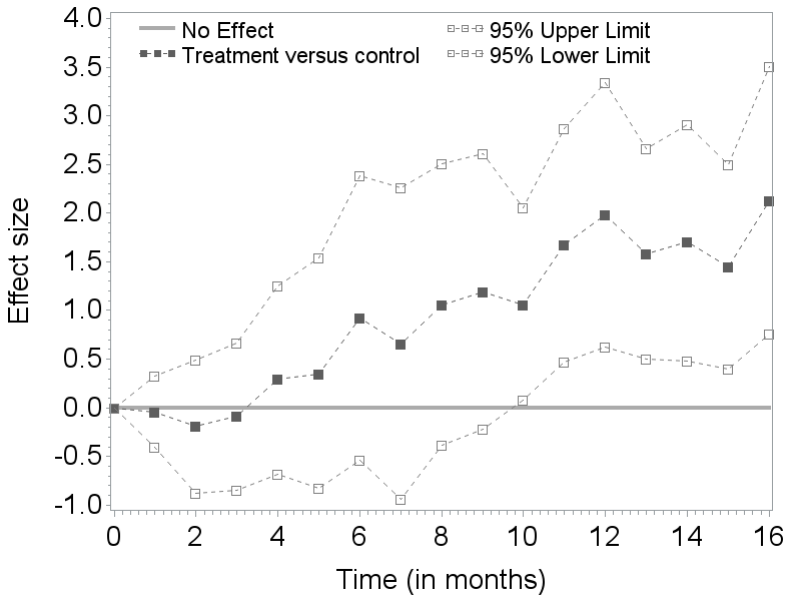
Lecture Notes

Analysis of Variance

[part of the course on longitudinal data analysis (2AMS10)]



Edwin van den Heuvel

Department of Mathematics and Computer Science

Eindhoven University of Technology

April 2021

Preface

Writing lecture notes on the ANalysis Of VAriance (ANOVA) in this day and age may be considered unnecessary. First of all, ANOVA is considered old-fashioned and it has grown into much more flexible statistical models for complex structured data that cannot be addressed by ANOVA models. This larger and more flexible class of models is referred to as the class of mixed models (linear, non-linear, and generalized linear mixed models). Secondly, there is a vast literature on ANOVA, including many books covering this topic (e.g., Scheffé, 1959; Graybill, 1961; Searle 1971; McCullagh and Searle, 2000; Searle et al., 2006). Thus writing about this topic, while highly-esteemed researchers who have strongly contributed to the development of ANOVA and who have written excellent books on the topic, may seem somewhat arrogant.

Nevertheless, I decided to write lecture notes on ANOVA, because I think there is no high-level overview of ANOVA that summarizes all the important elements and prepares for more complicated analysis with mixed effects models. Moreover, the excellent monographs on ANOVA are also somewhat elaborate and may therefore not be ideal to quickly grasp all the elements of ANOVA. This does not imply that I endorse students not to read these monographs, but I am just trying to summarize all the important elements of ANOVA in a condensed way with my lecture notes. Most of the material presented in this lecture notes can also be found in Searle et al. (2006) and McCullagh and Searle (2000). However, I have also included programming statements for the software program **SAS**, since this program is very strong in the analysis of mixed effects models (including the analyses with ANOVA models).

The current lecture notes on ANOVA are part of a new course on longitudinal data analysis at the Eindhoven University of Technology for master students in data science and artificial intelligence, where the students will learn about the breadth of mixed models in more detail. Thus my focus in the lecture notes is on random and mixed effects ANOVA models, and I mostly ignore the field of fixed effects ANOVA. I expect that students know the fixed effects ANOVA model (at least at the level of the one-way fixed effects ANOVA model) and that they have some mathematical maturity to read the lecture notes. Nevertheless, some background information on traditional statistical thinking will be provided in the first chapter. I am convinced that understanding random and mixed effects ANOVA models can help by a better understanding of the more complicated mixed effects models. ANOVA may help with the thinking in modeling data, which is an important element in all data analyses.

Edwin van den Heuvel

August 2021

Contents

1 Brief overview basic statistical theory

Within the field of statistics it is assumed that the *observed data* is being produced according to some form of (complex) probabilistic mechanism. This mechanism describes how certain variables are being dependent on each other and how they may vary across all the units of interest (e.g., sets of people, internet connections, or products). The set of units of interest is referred to as the *population* under study. The probability element in the probabilistic mechanism may come from different sources: (1) the way we collect the data (sampling units from the population), (2) the way we quantify results (measurement reliability of the variables of interest), and (3) possible other intrinsic sources of variation that causes differences between and within units. The ultimate goal of the field of statistics is to mathematically describe and recover this probabilistic mechanism for real data sets. When we would understand this probabilistic mechanism we would be able to make statements on the population under study for which this mechanism would apply. In many cases though, we do not need to know the exact probabilistic mechanism to be able to make certain statements. In many settings we would be satisfied with an approximate description of certain characteristics of the probabilistic mechanism underlying the data[1].

Making certain statements from data using mathematical models that addresses the uncertainty in data that is induced by the probabilistic mechanism is typically referred to as *statistical inference*. For instance, we may want to study the influence of insulin on the cognitive development of children having the Phelan-McDermid syndrome using a clinical trial (Zwanenburg et al., 2016). Here, the cognitive development of children is being compared within children by comparing the developmental growth under the placebo treatment with the growth under the insuline treatment (see Figure 1.1). In this particular study, the cognitive growth over time is being described by a piece-wise linear statistical model that would formulate the probability mechanism behind the observed data. The statement that would be formulated from such a data analysis is whether such treatment would benefit the cognitive development.

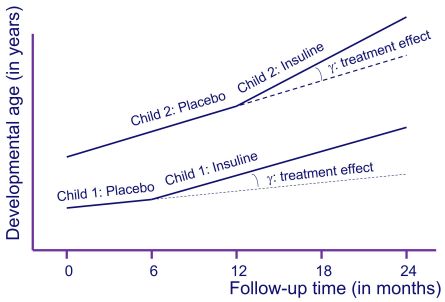


Figure 1.1: Piece-wise linear trajectories of cognitive development for Phelan-McDermid children who start on placebo and switch to insulin each at their own time. The treatment effect is then quantified by the change in slope with respect to cognitive development.

Thus part of the field of statistics is to estimate certain parameters that are

[1]In practice it is very complicated to recover the probabilistic mechanism, since the data only provides us with a partial description of the probabilistic mechanism. We never obtain all data or get a perfect insight in all sources of variation that causes the uncertainty. There are always pieces of information not present or available due to certains selection mechanisms (certain units are excluded, data is missing, and data can contain mistakes). For that reason we may be able to choose many and almost equally good mathematical models describing the data.

included in the selected mathematical model that is used to describe the probabilistic mechanism underneath the collected data. These parameters are typically denoted by Greek letters. For instance, we may use μ for the mean of the (sub)population, σ for the standard deviation, ρ for the correlation coefficient (between two variables), and β for a regression parameter (like the slope or change in slope in Figure 1.1). If we want to be generic, we typically use the notation θ for the parameter. Clearly, parameters may not just be one-dimensional and we are rather interested in a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_q)^T$. Vectors will be denoted in boldface. For instance, we may be interested in the average intercept, the slope, and the change in slope of the individual developmental trajectories for the study of insulin. More generically, parameters can also be much more abstract and they can be equal to *functions* f that may describe relations between variables or represent density or distribution functions. The principles will remain the same, but the statistical techniques may be different when we would consider these more abstract parameters. Here we will focus on parameters as elements of the set of all real numbers $\mathbb{R}$.

The collected data is now used to create estimators of the parameters of interest. Such estimators are typically denoted by the same Greek letter we use for parameters, but now with an additional hat on top of it, i.e., $\hat{\theta}$ is the estimator of the parameter θ (or $\hat{f}$ for the abstract parameter f). We may use subscripts or superscripts to add additional information to the estimator. For instance, we may use $\hat{\theta}_{\text{MLE}}$ to indicate that the estimator was obtained by the maximum likelihood estimation method. Additionally, we may use superscripts to indicate other aspects. For instance, we may use $\hat{\theta}^F_{\text{MME}}$ to indicate that the estimator was determined under a specific statistical assumption or for specific subgroups of data indexed by the letter F making use of the method of moment estimation approach. Furthermore, we will not use the notation $\hat{\theta}$ for estimators consistently, since in many settings estimators may be represented by capital roman letters. For instance, we may use S^2 or V as a notation for an estimator for a variance parameter σ^2. Again we could apply subscripts to refer to specific assumptions, subgroups of data, or estimation approaches.

Throughout the lecture notes, we will assume that every estimator $\hat{\theta}$ has a finite variance $\mathsf{VAR}(\hat{\theta}) < \infty$[2]. Often we may have a mathematical expression of this variance. For instance, the variance of $\hat{\mu}_{\text{WLS}} = \sum_{k=1}^m [X_k/\sigma_k^2] / \sum_{k=1}^m [1/\sigma_k^2]$ is equal to $\mathsf{VAR}(\hat{\mu}_{\text{WLS}}) = [\sum_{k=1}^m \sigma_k^{-2}]^{-1}$, when X_1, X_2, ..., X_m are independent and identically normally distributes (i.e., $X_k \sim_{\text{i.i.d.}} N(\mu, \sigma_k^2)$). In practice we often do not know the variance $\mathsf{VAR}(\hat{\theta})$, since this variance could depend on unknown model parameters (like the parameters σ_k). Therefore, the $\mathsf{VAR}(\hat{\theta})$ must be estimated too. We will assume that such an estimator exists. It should be noted that $[\mathsf{VAR}(\hat{\theta})]^{1/2}$ is called the *standard error* of the estimator $\hat{\theta}$. The standard error is often quantified to help us understand how precise we have been able to quantify the parameters of the probabilistic mechanism. This may help us determine our confidence in our understanding of the probabilistic mechanism.

[2] A finite variance of an estimator may not exist in all cases. For instance, the estimator $\sum_{k=1}^m X_k/(2m) = 0.5\bar{X}$ for the scale parameter β, when X_1, X_2, ..., X_m are i.i.d. inverse Gamma distributed with unknown scale parameter β and known shape parameter $\alpha = 3/2$ (i.e.,$X_i \sim_{\text{i.i.d.}} IG(\beta, 3/2)$), is unbiased ($\mathbb{E}[0.5\bar{X}] = \beta$), but has a variance that is infinite ($\mathsf{VAR}(0.5\bar{X}) = \infty$).

By estimating the parameters of the selected probabilitic mechanism for the observed data we can do inferences on the population of units. We would be able to make statements about the population. For instance, the average developmental age for Phelan-McDermid is equivalent to 30% of their biological age, indicating that Phelan-McDermid children develop 70% slower than "normal children". Or, the use of insulin would improve the cognitive development with, say, 20%. Or in terms of another example in the area of prediction, we may conclude that detecting breast cancer using an MRI is 88% accurate and 10% better than the detection of a mammography. Such statements are valuable, but does not provide evidence about how confident we are on these statements. Thus, in statistics it is common to report a 95% confidence interval on the quantitative values mentioned in the statement. Such an interval would provide a range of values that are also supported by the data. Thus, in statistics we would rather prefer the statement that detecting breast cancer using an MRI is 88% [78%, 98%] accurate and 10% [5%, 15%] better than the detection of a mammography.

The complete cycle of statistical inference that starts with a population under study, the way that the data has been collected from this population, the possible choices of probabilistic mechanisms that may be used to describe the observed data, and the approaches of estimating this mechanism using the observed data is visualized in Figure 1.2 (see also chapter 5 of Kaptein and Van den Heuvel, 2022). This whole process is meant to obtain a (better) understanding of the population under study using statistical thinking and methods.

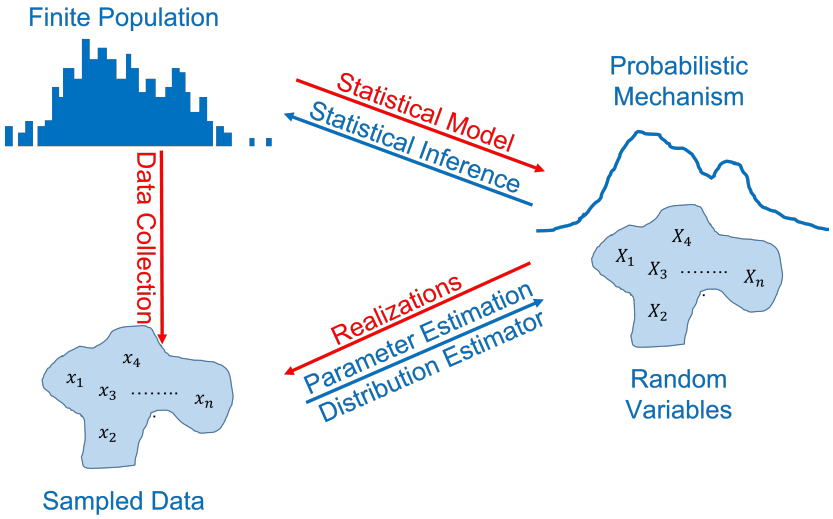


Figure 1.2: Visualization of statistical inference

Throughout the lecture notes we will make use of several statistical theories that are considered common knowledge. These theories are briefly described below (see also Kaptein and Van den Heuvel, 2022). Here we will describe some distributional theory that is based on independently and normally distributed random variables X_1, X_2, ..., X_m. Thus the probabilistic mechanism in Figure 1.2 is being described by a normal distribution. Throughout the lecture notes we will frequently make the assumption that the observed data is collected form

a normal distribution and we wish to make use of the following well-known distributional theories.

1.1 Some distributional theory

The data analysis methods that are used in statistics are heavily based on probability distributions. A *random variable* X is a variable that can take on different values (often in the set of real numbers $\mathbb{R}$). It is a description of all the values that may occur in the population, i.e., the range of values present in the population. The random variable can be continuous or discrete, depending on the application of interest. The random variable is seen as an abstraction of the data that was obtained for this variable. A *realization* of the random variable X is denoted as x and in real data we obtain seveal realizations (as visualized in Figure 1.2).

The *probability distribution* F of a random variable X describes the proportion of units from the population that have their value in the interval $(-\infty, x]$, i.e., $F(x) = P(X \leq x) \in [0, 1]$. This formulation holds for both continuous as discrete random variables. Thus the random variable X and F are uniquely connected. Knowing X means that we know F and knowing F means that we know X. The function $F(x)$ satistifies certain conditions: (1) $\lim_{x \downarrow -\infty} F(x) = 0$, (2) $\lim_{x \to \infty} F(x) = 1$, and (3) $F(x_1) \leq F(x_2)$ when $x_1 < x_2$. In probability theory many different distribution function (or densities) have been developed. These distributions may have arisen from practical settings as well as from theoretical derivations.

Many distribution functions have a derivative f, called the *density function*. The relation with the distribution function F is given by

$$F(x) = \int_{-\infty}^{x} f(u)du. \tag{1.1}$$

This relation is typically used for continuous random variables, since continuous random variables have continuous distribution functions F. For discrete distribution functions we have to replace the integral in (1.1) by a summation:

$$F(x) = \sum_{k=1}^{\infty} 1_{(-\infty, x]}(x_k) p_k,$$

with $\{x_1, x_2, ..., x_k, ...\}$ the set of values that X can attain, $p_k = P(X = x_k)$ the probability that X takes on value x_k, $\sum_{k=1}^{\infty} p_k = 1$, and with $1_A(x)$ the indicator function being equal to 1 when $x \in A$ and zero otherwise.

In statistics we are often interested in several characteristics of the random variable, like its mean value or its variance. The mean and variance of X (or interchangeably F) are given by

$$\begin{array}{lrcl} \text{Mean :} & \mathbb{E}(X) & = & \int_{-\infty}^{\infty} x f(x) dx, \\ \text{Variance :} & \mathsf{VAR}(X) & = & \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, \end{array} \tag{1.2}$$

with μ another notation for $\mathbb{E}(X)$. More general, the expectation of the random variable $\psi(X)$ is given by

$$\mathbb{E}(\psi(X)) = \int_{-\infty}^{\infty} \psi(x) f(x) dx.$$

For a discrete random variable X this generic formulation is given by

$$\mathbb{E}(\psi(X)) = \sum_{k=1}^{\infty} \psi(x_k) p_k.$$

With this general formulation we may be able to obtain any feature of the random variable we would like to calculate (e.g., other moments like skewness and kurtosis).

1.1.1 The weighted and arithmetic average

Assume that we have collected X_1, X_2, ..., X_m independent random variables having a normal distribution with mean μ_k and variance σ_k^2. The weighted average $\bar{X}_W = \sum_{i=1}^{n} w_k X_k$, with known weights w_k and with $\sum_{k=1}^{m} w_k = 1$, is also normally distributed with mean $\mu = \sum_{k=1}^{m} w_k \mu_k$ and variance $\sigma^2 = \sum_{k=1}^{m} w_k^2 \sigma_k^2$. This also implies that the normalized weighted average $(\bar{X}_W - \mu)/\sigma$ has a standard normal distribution function. In case we choose the weights equal to $w_k = 1/m$, the arithmetic average $\bar{X} = \sum_{k=1}^{n}(X_k/m)$ has a normal distribution function with average mean $\bar{\mu} = \sum_{k=1}^{m}(\mu_k/m)$ and variance $\bar{\sigma}^2/m$, where $\bar{\sigma}^2 = \sum_{k=1}^{m}(\sigma_k^2/m)$ is the average variance. If in addition, the means μ_k and variances σ_k^2 do not vary with unit k either (i.e., $\mu_k = \mu$ and $\sigma_k^2 = \sigma^2$), the arithmetic average $\bar{X}$ is normally distributed with mean μ and variance σ^2/m.

1.1.2 Sums of squared variables

If Z is a standard normally distributed random variable, then Z^2 is χ^2-distributed random variable with only one degrees of freedom. Thus also $(X_k - \mu_k)^2/\sigma_k^2$ is χ^2-distributed with one degrees of freedom. The sum of the squared standardized random variables $\sum_{k=1}^{m}(X_k - \mu_k)^2/\sigma_k^2$ is known to be χ^2-distributed with m degrees of freedom. The reason is that the sum of independently χ^2-distributed random variables is again χ^2-distributed. Indeed, if χ_1^2, χ_2^2, ..., χ_p^2 are independent χ^2-distributed variables with d_1, d_2, ..., d_p degrees of freedom, then $\sum_{i=1}^{p} \chi_i^2$ is χ^2-distributed with a degrees of freedom equal to $\sum_{i=1}^{p} d_i$. Furthermore, the sum $\sum_{k=1}^{m}(X_k - \hat{\mu}_{\mathrm{WLS}})^2/\sigma_k^2$ is χ^2-distributed with $m-1$ degrees of freedom when the means are all equal $\mu_k = \mu$ (Searle et al., 2006). Note that $(\hat{\mu}_{\mathrm{WLS}} - \mu)^2 \sum_{k=1}^{m}(\sigma_k^{-2})$ is χ^2-distributed with just one degrees of freedom when $\mu_k = \mu$, since $\hat{\mu}_{\mathrm{WLS}} = \sum_{k=1}^{m}[X_k/\sigma_k^2] / \sum_{i=1}^{m}[1/\sigma_k^2] \sim N(\mu, [\sum_{i=1}^{m}(\sigma_k^{-2})]^{-1})$. When additionally, all the variances are equal ($\sigma_k^2 = \sigma^2$), we have that $\sum_{k=1}^{n}[(X_k - \bar{X})^2/\sigma^2]$ is χ^2-distributed with $m-1$ degrees of freedom. We will make use of the χ^2-distribution when we calculate confidence intervals for variance components (see Section 2.9.3).

1.1.3 Student's t-distribution

Let Z be a standard normally distributed random variable, $Z \sim \mathcal{N}(0,1)$, S_m^2 be χ^2-distributed with m degrees of freedom, $S_m^2 \sim \chi_m^2$, and assume that Z and S_m^2 are independent. The distribution of the ratio $Z/(S_m/\sqrt{m})$ has a t-distribution with m degrees of freedom. Thus the ratio $(\bar{X} - \mu)/[\sum_{k=1}^{m}(X_k - \bar{X})^2/(m(m-1))]^{1/2}$ is t-distributed with $m-1$ degrees of freedom when the

means and variances do not vary with units ($\mu_k = \mu$ and $\sigma_k^2 = \sigma^2$). The random variables $\bar{X}$ and $\sum_{k=1}^{m}(X_k - \bar{X})^2$ are independent when $X_k \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$. We will see the use of the t-distribution when we calculate confidence intervals for normally distributed random variables.

1.1.4 Fisher's F-distribution

If χ_1^2 and χ_2^2 are two independently χ^2-distributed random variables with d_1 and d_2 degrees of freedom, then the ratio $F = [\chi_1^2/d_1]/[\chi_2^2/d_2]$ is F-distributed with d_1 and d_2 degrees of freedom. We will use the F-distribution for confidence intervals on ratio's of variance components (see Section 2.9.4)

1.2 Central limit theorem

In many settings we may study averages or sums of random variables, i.e., $Y_m = \sum_{k=1}^{m}(X_k/m)$, with $X_k \sim_{\text{i.i.d.}} F$, where F is an unknown distribution function with finite variance ($\mathsf{VAR}(X_k) < \infty$). When the sample size is large, the distribution function of Y_m is approximately normal irrespective of the shape of the distribution F. In mathematical terms this means that

$$\lim_{m\to\infty} P(\sqrt{m}[Y_m - \mathbb{E}(X_k)]/[\mathsf{VAR}(X_k)]^{1/2} \leq y) = \Phi(y),$$

with Φ the standard normal distribution function. This theorem is called the *central limit theorem*. There also exists a multivariate formulation of this central limit theorem.

It is not necessary that X_1, X_2, ..., X_m are identically distributed. If we assume that $\mathbb{E}(X_k) = \mu_k$ and $\mathsf{VAR}(X_k) = \sigma_k^2$, the distribution function of $\sum_{k=1}^{m}(X_k - \mu_k)/s_m$ converge to the standard normal distribution function, with $s_m^2 = \sum_{k=1}^{m}\sigma_k^2$ the sum of the variances. However, to proof this form of the central limit theorem it requires an additional restriction on the moments of X_k and finite variances is not enough (Billingsley, 1968). There are also other extensions of the central limit theorem where some dependence between the random variables X_1, X_2, ..., X_m are allowed, but this is outside the scope of these lecture notes.

1.3 Estimation methods

We will make use of the estimation techniques called the *method of moments* (MME) and *maximum likelihood estimation* (MLE). These two estimation methods will resort in parameter estimates of the statistical model that will be formulated for the collected data. A statistical model is a mathematical formulation of the probabilistic mechanism that is essentially nothing more than a (possibly complex) density function for the data, although we will use other forms of model descriptions. For these other model formulations we may apply the *ordinary least squares* (OLS) and *weighted least squares* (WLS) as estimation techniques. They are strongly related to the MME and MLE. Whatever estimation technique is applied, the parameter estimates will be functions of our collected data and are therefore referred to as the *parameters estimators*, since we will treat the data as a realization of a set of random variables that may

be seen the probabilistic mechanism or data generators (see Figure 1.2). The different estimation approaches may or may not result in the same parameter estimators, depending on the distributional assumptions of the data and the model choices.

Here we will assume that we want to estimate the vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_q)^T$ and that we have collected data $\boldsymbol{x}$ to do so. As just mentioned, this data $\boldsymbol{x}$ is considered a realization of the random variables $\boldsymbol{X}$. The random variables $\boldsymbol{X}$ can be simply a vector of single observations $\boldsymbol{X} = (X_1, X_2, ..., X_m)^T$ (as we have discussed in Section 1.1), or it could be represented in more complicated data structures (see Section 2.9). Whatever the data structure, we assume that the density of $\boldsymbol{X}$ is given by $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. In case the random variable $\boldsymbol{X}$ is a vector of independent random variables X_1, X_2, ..., X_m, the density $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ can be written as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{k=1}^{m} f_{\boldsymbol{\theta}}(x_k)$.

1.3.1 Method of moments

Here we will assume that we have observed the simpler data structure $\boldsymbol{X} = (X_1, X_2, ..., X_m)^T$ and that $X_k \sim_{i.i.d.} f_{\boldsymbol{\theta}}(x)$ (leading to$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{k=1}^{m} f_{\boldsymbol{\theta}}(x_k)$). For the ANOVA models we will see that more complicated data structures are allowed for moment estimation of the model parameters. To be able to estimate the elements in the vector $\boldsymbol{\theta}$ we will calculate the first q moments of X_k (assuming that they all exist). We would then obtain a set of equations:

$$\begin{array}{rcccl} \mu_1 & = & \mathbb{E}(X_k) & = & g_1(\boldsymbol{\theta}) \\ \mu_2 & = & \mathbb{E}(X_k - \mu_1)^2 & = & g_2(\boldsymbol{\theta}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_q & = & \mathbb{E}(X_k - \mu_1)^2 & = & g_q(\boldsymbol{\theta}) \end{array} \tag{1.3}$$

with $g_r : \mathbb{R}^q \longrightarrow \mathbb{R}$ a known function for the expectations. If we would now solve the equations in (1.3) such that each parameter θ_r is written as function of the moments $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_q)^T$, i.e., $\theta_r = h_r(\boldsymbol{\mu})$, we have created a new set of equations where the parameters are expressed in terms of the moments. The moments μ_r can be directly estimated by the sample moments $\hat{\mu}_r = \frac{1}{m}\sum_{k=1}^{m}\left(X_k - \bar{X}\right)^r$, with $\hat{\mu}_1 = \bar{X}$ the sample average. Then substituting the moment estimators $\hat{\mu}_r$ in the set of equations $\theta_r = h_r(\boldsymbol{\mu})$, we obtain the moment estimators $\hat{\boldsymbol{\theta}}_r = h_r(\hat{\boldsymbol{\mu}})$.

1.3.2 Maximum likelihood estimation

For the maximum likelihood we will assume the general setting with density $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. This density is called the likelihood function and maximizing this function over $\boldsymbol{\theta}$ results in the maximum likelihood estimator. Often it is easier to maximize the logarithm of the density $\ell(\theta) = \log(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$, in particular when some independence between observations are present. The maximum likelihood estimators are obtained by solving the likelihood equations:

$$\ell_k^{'}(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_k} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right) / f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = 0, \tag{1.4}$$

with $\ell_k^{'}(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\theta_k$ the derivative of $\ell(\boldsymbol{\theta})$ with respect to θ_k. This derivative is referred to as the *score function*. The solution $\hat{\boldsymbol{\theta}}$ of (1.4) does not always result in a closed form expression, which means that we have to resort to numerical approaches if we want to determine the MLE on data. Furthermore, there may also be certain boundary constraints present on the parameters leading to issues in solving the likelihood equations.

The standard errors of the maximum likelihood estimators are calculated via the variances and covariances of the large sample distribution (or asymptotics) of the maximum likelihood estimators. Under certain regularity conditions, it can be shown that $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is approximately multivariate normal $\mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$, with $I(\boldsymbol{\theta})$ the so-called Fisher information matrix. The Fisher information matrix is determined by the second derivative of the log likelihood function. The kth and lth element $I_{kl}(\boldsymbol{\theta})$ of the Fisher information matrix $I(\boldsymbol{\theta})$ is given by $I_{kl}(\boldsymbol{\theta}) = -\mathbb{E}[\ell_{kl}^{'}(\boldsymbol{\theta})]$, with $\ell_{kl}^{'}(\boldsymbol{\theta})$ the second derivative of the log likelihood function: $\ell_{kl}^{'}(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/(\partial\theta_k\partial\theta_l)$. The covariance of the maximum likelihood estimators $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\theta}}_l$ is now defined by $\mathsf{COV}(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\theta}}_l) = I_{kl}^{-1}(\boldsymbol{\theta})$, with $I_{kl}^{-1}(\boldsymbol{\theta})$ the kth and lth element of the inverse Fisher information matrix $I^{-1}(\boldsymbol{\theta})$.

1.4 Asymptotic confidence intervals

A confidence interval for a parameter θ represents a range of values that would still be a reasonable alternative estimate of the parameter with the used confidence level. For instance, let $\hat{\gamma} = 0.2$ be the estimate for the treatment effect γ of the insulin treatment on cognitive development from the insulin trial let $[-0.1, 0.5]$ be the calculated 95% confidence interval. The trial has provided us evidence that the most likely value for the treatment effec is 0.2, but we are also confident that the true effect size would fall within -0.1 and 0.5 with 95% confidence. Thus, a confidence interval gives a range of values for the parameter that would be reasonable alternative estimates of the parameter that is supported by the data with the stated confidence level. The *confidence level* is typically indicated by $100\%(1-\alpha)$, with α the significance level (see Section 1.7). It is common to choose $\alpha = 0.05$, to obtain 95% confidence intervals.

Asymptotic confidence intervals for a specific parameter θ are typically based on the asymptotic distribution of the estimator $\hat{\theta}$. Since many estimators have an asymptotic normal distribution (when the estimator is properly normalized), the confidence interval for θ is based on the normal distribution as if we would be creating a confidence interval for the mean of the normal distribution. The normal distribution is then used as an approximation of the distribution of the estimator to construct a confidence interval. Thus the basic idea is that the probability distribution for the normalized estimator $(\hat{\theta}-\theta)/[\mathsf{VAR}(\hat{\theta})]^{1/2}$ can be approximated by the standard normal distribution function. This means that

$$P((\hat{\theta} - \theta)/[\mathsf{VAR}(\hat{\theta})]^{1/2} \leq x) \approx \Phi(x),$$

with Φ the standard normal distribution function. In case we have collected enough data, we know that the approximation is quite close due to the asymptotic theory.

To obtain an approximate $100\%(1-\alpha)$ confidence interval for θ, we could now choose $\hat{\theta} \pm z_{1-\alpha/2}[\mathsf{VAR}(\hat{\theta})]^{1/2}$, with $z_q = \Phi^{-1}(q)$ the q^{th} quantile of a standard normal distribution. As mentioned earlier, in practice we often do not

know the variance $\mathsf{VAR}(\hat{\theta})$, since this variance could depend on the unknown model parameters, and should therefore be estimated. In many applications an estimator $V^{1/2}(\hat{\theta})$ of the standard error $[\mathsf{VAR}(\hat{\theta})]^{1/2}$ would exist. In case the sample sizes are large, the estimator $V^{1/2}(\hat{\theta})$ may be close to the unknown standard error $[\mathsf{VAR}(\hat{\theta})]^{1/2}$ and we could just substitute $V^{1/2}(\hat{\theta})$ for $[\mathsf{VAR}(\hat{\theta})]^{1/2}$ in the confidence interval $\hat{\theta} \pm z_{1-\alpha/2}[\mathsf{VAR}(\hat{\theta})]^{1/2}$. However, we believe that it would be better to alter the quantile $z_{1-\alpha/2}$ by a quantile of the t-distribution when a degrees of freedom for the estimated standard error $V^{1/2}(\hat{\theta})$ is obtained. Not addressing the uncertainty of the estimator $V^{1/2}(\hat{\theta})$ for the standard error $[\mathsf{VAR}(\hat{\theta})]^{1/2}$ would lead to asymptotic confidence intervals that are too narrow. The $100\%(1-\alpha)$ asymptotic confidence interval for θ can now be calculated as

$$\hat{\theta} \pm t^{-1}_{df_V}(1-\alpha/2)V^{1/2}(\hat{\theta}), \tag{1.5}$$

with df_V the degrees of freedom for the estimated standard error $V^{1/2}(\hat{\theta})$ and $t^{-1}_{df}(q)$ the q^{th} quantile of the t-distribution with df degrees of freedom. Calculation of the degrees of freedom df_V is often conducted with generic approaches using chi-square distributions formulated by Satterthwaite (1946), see Section 2.4.4.

1.5 The delta method

Let $g(\boldsymbol{X})$ be a random variable for which we would like to calculate its variance, with $g : \mathbb{R}^K \rightarrow \mathbb{R}$ a K-dimensional function and $\boldsymbol{X} = (X_1, X_2, ..., X_K)^T$ a K-dimensional random vector. We assume that the random variable X_k has mean $\mu_k = \mathbb{E}(X_k)$ and variance $\sigma^2_k = \mathbb{E}(X_k - \mu_k)^2$, for $k = 1, 2, ..., K$. Furthermore, we denote the covariances by $\sigma_{kl} = \mathbb{E}(X_k - \mu_k)(X_l - \mu_l)$, with $\sigma_{kk} = \sigma^2_k$. Then we approximate the random variable $g(\boldsymbol{X})$ by its first derivative:

$$g(\boldsymbol{X}) \approx g(\boldsymbol{\mu}) + \sum_{k=1}^{K} g'_k(\boldsymbol{\mu})(X_k - \mu_k), \tag{1.6}$$

with $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_K)^T$ and g'_k the derivative of the function g with respect to the k^{th} dimension (i.e., $\partial g(\boldsymbol{x})/\partial x_k = g'_k(\boldsymbol{x})$). Now applying the approximation in (1.6), we obtain that

$$\mathbb{E}(g(\boldsymbol{X}) - g(\boldsymbol{\mu}))^2 \approx \sum_{k=1}^{K}\sum_{l=1}^{K} g'_k(\boldsymbol{\mu})g'_l(\boldsymbol{\mu})\sigma_{kl}. \tag{1.7}$$

To get an estimate of this term, we need to substitute an estimate for the parameters $\boldsymbol{\mu}$ and σ_{kl}. Such estimates are often available in the settings that we will discuss (as we already indicated).

It is important to realize that $\mathbb{E}(g(\boldsymbol{X}) - g(\boldsymbol{\mu}))^2$ is most likely not equal to the variance of $g(\boldsymbol{X})$, because it is likely that $\mathbb{E}(g(\boldsymbol{X})) \neq \mathbb{E}(g(\boldsymbol{\mu}))$, in particular when we are dealing with non-linear functions g. The use of $g(\boldsymbol{\mu})$ is warranted due to asymptotic reasoning. The random vector $\boldsymbol{X}$ is often a vector of estimators based on underlying data (see Section 1.4). If the underlying data would then become very large, the random vector $\boldsymbol{X}$ would converge to $\boldsymbol{\mu}$ and $g(\boldsymbol{X})$ would converge to $g(\boldsymbol{\mu})$. Thus asymptotically, $\mathbb{E}(g(\boldsymbol{X}) - g(\boldsymbol{\mu}))^2$ would represent the variance $\mathsf{VAR}(g(\boldsymbol{X}))$.

1.6 Correlation analysis

In the field of statistics, it is assumed that the relation between certain variables, or in more generic terms the dependency between variables, is fully captured by the joint distribution function of these variables. If we consider just two variables, represented by X and Y, obtained at the same units, we may investigate the joint distribution function $H(x, y)$ to understand the dependency. This distribution function would describe the probabilistic mechanism for X and Y for the population of units under study (see Figure 1.2). If the two variables are unrelated or *independent*, the joint distribution function is the product of the marginal distribution functions of X and Y, i.e., $H(x, y) \equiv P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \equiv F(x)G(y)$.

There have been many approaches towards capturing or quantifying the dependency of variables. One way is making use of statistical models, like the mixed effects models that will be discuss in these lecture notes in more detail, but also non-model based approaches have been proposed which are often considered more basic statistical theory. For instance, many *measures of association* have been developed over more than 100 years for different types of data and they are still actively used for data analysis. These measures have been developed to quantify the strength of the relationship between X and Y (Kaptein and Van den Heuvel, 2022). When both X and Y are continuous common measures of association are Pearson's correlation coefficient, Kendall's τ, and Spearman's ρ. When X and Y are both categorical variables common measures of association are Pearson's χ^2 statistics, Pearson's ϕ coefficient, Cramér's V, Goodman and Kruskal's γ, and Cohen's (weighted) κ statistic. These measures do also apply when X and Y are both binary, but for that situation there exist many more measures of association as well, like the risk difference, the relative risk, the odds ratio, many (dis)similarity measures, and Yule's Q.

Here we will discuss Pearson's correlation coefficient between two variables X and Y, since this measure of association connects very nicely with the correlations we will discuss in terms of mixed effects models. Based on the collected data (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), Pearson suggested to calculate the correlation by

$$\hat{\rho}_P = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})/(S_X S_Y), \tag{1.8}$$

with $\bar{X} = \sum_{i=1}^{n}(X_i/n)$ and $\bar{Y} = \sum_{i=1}^{n}(Y_i/n)$ the averages of the variables X and Y and with $S_X^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ and $S_Y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ the variances of these variables. It can be shown that this statistic is an element of the interval $[-1, 1]$. It can be viewed as an estimator of the parameter ρ_P, which is defined by

$$\rho_P = \mathsf{COV}(X, Y)/\sqrt{\mathsf{VAR}(X)\mathsf{VAR}(Y)}. \tag{1.9}$$

It should be noted that Pearson's estimator in (1.8) for the parameter in (1.9), is not unbiased (i.e., $\mathbb{E}(\hat{\rho}_P) \neq \rho_P$). This means that on average the estimator is not on target. However, asymptotically the estimator is *consistent*, which means that the estimator converges to the parameter when the sample size increases to infinity, i.e., $\lim_{n\to\infty} \mathbb{E}(\hat{\rho}_P) = \rho_P$. Furthermore, the asymptotic distribution of Pearson's correlation coefficient is normal, i.e., $\sqrt{n}(\hat{\rho}_P - \rho_P) \longrightarrow_{n\to\infty} Z$, with

Z being normally distributed with zero mean and some variance. This variance depends on the variances and covariances of the variables X^2, Y^2, and XY, but is approximately equal to $(1-\rho_P^2)^2$ (see Bonett and Wright, 2000).

If we assume that the joint distribution of (X, Y) is bivariate normal[3], with mean parameters (μ_X, μ_Y), variance parameters (σ_X^2, σ_Y^2), and correlation coefficient ρ, the correlation coefficient defined in (1.9) is equal to the correlation parameter ρ from the bivariate normal distribution. The asymptotic variance of Pearson's estimator in (1.8) is then exactly equal to $(1-\rho^2)^2$. Moreover, the finite distribution function of Pearson's estimator $\hat{\rho}_P$ is determined by the t-distribution but only when the correlation coefficient $\rho = 0$. In this case $\hat{\rho}_P\sqrt{n-2}/\sqrt{1-\hat{\rho}_P}$ has a t-distribution with $n-2$ degrees of freedom. Sir Ronald Fisher was able to demonstrate (without making the assumption that $\rho = 0$) that the distribution of $z_{\hat{\rho}_P} = 0.5[\log(1+\hat{\rho}_P) - \log(1-\hat{\rho}_P)]$ is approximately normal with mean $0.5[\log(1+\rho) - \log(1-\rho)]$ and variance $1/(n-3)$. The transformation function $0.5[\log(1+\rho) - \log(1-\rho)]$ is referred to as Fisher's z transformation. Confidence intervals on the correlation coefficient ρ are therefore often calculated through Fisher's z transformation using Pearson's estimator. A 95% confidence interval for $0.5[\log(1+\rho) - \log(1-\rho)]$ is then given by

$$[z_{\hat{\rho}_P} - 1.96/\sqrt{n-3}, z_{\hat{\rho}_P} - 1.96/\sqrt{n-3}]. \tag{1.10}$$

To obtain a 95% confidence interval on ρ, the confidence interval in (1.10) can be transformed back by using the inverse function $[\exp\{2z\} - 1]/[\exp\{2z\} + 1]$.

1.7 Hypothesis testing

In many settings we would like to know whether the probabilistic mechanism underneath the data is different for certain settings or subgroups of units. For example, in the insulin trial we may want to know whether the treatment effect (the difference in slope for placebo and insulin treatment) is different from zero or alternatively we may want to know whether girls benefit more than boys. This formulation refers to the probabilistic mechanism underneath the data and not to the data self. The data is used to make a decision on the probabilistic mechanism. In statistics such decision problems are formulated into a *null hypothesis* and an *alternative hypothesis* and the goal of hypothesis testing is to reject the null hypothesis in favor of the alternative hypothesis.

In case of demonstrating a treatment effect of insulin, the null hypothesis may be formulated as $H_0 : \gamma = 0$, with γ the parameter that describes the change in slopes for cognitive development under the placebo and insulin treatment. An alternative formulation could be $H_0 : \beta_P = \beta_I$, with β_P the slope for cognitive growth under placebo and β_I the slope for cognitive growth under insulin. The parameter γ represents the difference in slopes, i.e., $\gamma = \beta_I - \beta_P$. The alternative hypothesis can be formulated as $H_1 : \gamma \neq 0$ or $H_1 : \beta_P \neq \beta_I$. This is refered to as *two-sided* hypothesis testing, since the alternative hypothesis does not favor

[3]The bivariate normal distribution is defined by its density function given by

$$\phi_2(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{z_X^2 - 2\rho z_X z_Y + z_Y^2}{2(1-\rho^2)}\right\},$$

with $z_X = (x - \mu_X)/\sigma_X$ and $z_Y = (x - \mu_Y)/\sigma_Y$.

any direction. In case we wish to demonstrate that insulin is beneficial (on average) we would formulate a *one-sided* alternative hypothesis as $H_1 : \gamma > 0$ or $H_1 : \beta_P < \beta_I$.

In this formulation of the problem, we have to make a binary decision using the available data. The data is either convincing us that the alternative hypothesis is correct or the data does not show enough evidence to reject the null hypothesis. This is similar to our justice system where the evidence should lead to a conviction without a reasonable doubt. Our decision based on the data may not be correct. The reason is that in statistics we believe that the data does not fully describe the probabilistic mechanism perfectly but only approximately. We could make two mistakes: we incorrectly accept the alternative hypothesis or we incorrectly do not reject the null hypothesis. Again, this is similar to our justice system were we may convict an innocent individual or acquit a criminal..

The *type 1 error* refers to the probability of incorrectly rejecting the null hypothesis in favor of the alternative and the *type 2 error* refers to probability of incorrectly not rejecting the null hypothesis. The *power* is the probability of correctly rejecting the null hypothesis. Thus the power is one minues the type 2 error. In practice, we can bound (from above) the type 1 error independent of the amount of available data. This is called the *significance level* α and it is typically set equal to $\alpha = 0.05$. The type 2 error depend on several things and is difficult to bound (from above) in practice. For instance, it depends on the amount of collected data and on the real setting of the alternative hypothesis. Indeed, in the insulin trial we will less likely make a type 2 error when we study more children and/or more temporal observations. Additionally, the larger the value for γ in the insulin trial the less likely we will make a type 2 error in our decision. However, it also depends on the significance level. If we choose a large significance level, we are less likely to make a type 2 error (but of course at the expense of a potentially larger type 1 error). Finally, the type 2 error also depends on the measurement precision. Measuring cognition is relatively noisy and having more precise measurements would provide us with better determination of the cognitive growth.

Estimation of the parameters in the statistical model (e.g., the piece-wise linear growth model in the insulin trial) in combination with the construction of (asymptotic) confidence intervals on these parameters may help make us a decision between the null and alternative hypothesis. As indicated in Section 1.4, the confidence interval would indicate a range of values for the parameter that would be supported by the data with the stated confidence level used for the interval. Thus, this confidence interval can be used to evaluate a null hypothesis that is formulated on the parameter. For instance, if we assume that the calculated 95% confidence interval for the treatment effect γ is $[-0.1, 0.5]$, there will not be enough evidence for the alternative hypothesis $H_1 : \gamma \neq 0$ when a significance level of $\alpha = 0.05$ is used. The trial has indicated that a reasonable alternative estimate for γ would be equal to $\hat{\gamma} = 0$, since the value 0 falls inside the 95% confidence interval $[-0.1, 0.5]$. The null hypothesis $\gamma = 0$ is thus a likely solution that is supported by the data when we wish to use a maximum type 1 error of $\alpha = 0.05$. Instead, if the 95% confidence interval would have been equal to $[0.1, 0.3]$, an estimate of $\hat{\gamma} = 0$ would not be very likely, since we expect the treatment effect γ to be away from zero and inside the interval $[0.1, 0.3]$ with 95% certainty. Thus, the null hypothesis $H_0 : \gamma = 0$ is considered unlikely and

will therefore be rejected in favor of the alternative hypothesis $H_1 : \gamma > 0$.

In many application areas, there is a wish to quantify how likely the null hypothesis is supported by the data. This is called the *p-value*. It is defined by the probability of observing the obtained estimate and more extreme values when the null hypothesis is correct. For the alternative hypothesis $H_1 : \gamma > 0$ the asymptotic p-value is defined by $p \approx 1 - \Phi(\hat{\gamma}/\sqrt{\mathsf{VAR}(\hat{\gamma})})$, with Φ the standard normal distribution function and $\mathsf{VAR}(\hat{\gamma})$ the variance of the estimator $\hat{\gamma}$. For the two-sided alternative hypothesis $H_1 : \gamma \neq 0$, the p-value would be calculated by $p \approx 2(1 - \Phi(|\hat{\gamma}|/\sqrt{\mathsf{VAR}(\hat{\gamma})}))$,with $|\cdot|$ the absolute value function. Since the variance $\mathsf{VAR}(\hat{\gamma})$ must be estimated in practice as well, we often use the t-distribution with an appropriate degrees of freedom (see Section 2.4.4) instead of the normal distribution when we replace $\mathsf{VAR}(\hat{\gamma})$ by its estimator (similar to the discussion on confidence intervals in Section 1.4).

1.8 Introduction to SAS

There exists many different statistical software packages that could be used for data analysis. Many of them are menu-based (e.g., SPSS, Minitab, Stata, JMP), where the user can load in the data in the package and use a menu of tools to select from to analyze the data. From the menu the user can select one tool of interest at the time. There also exists packages that do not make use of a menu from which you can choose your analysis (**SAS**, R, Python), but you have to write your codes yourself and then run the codes on the data. These software packages still have pre-defined macros or procedures that one can use in the codes. Thus, they are in principle not that much different from menu-based software packages, they merely differ in the way they are used.

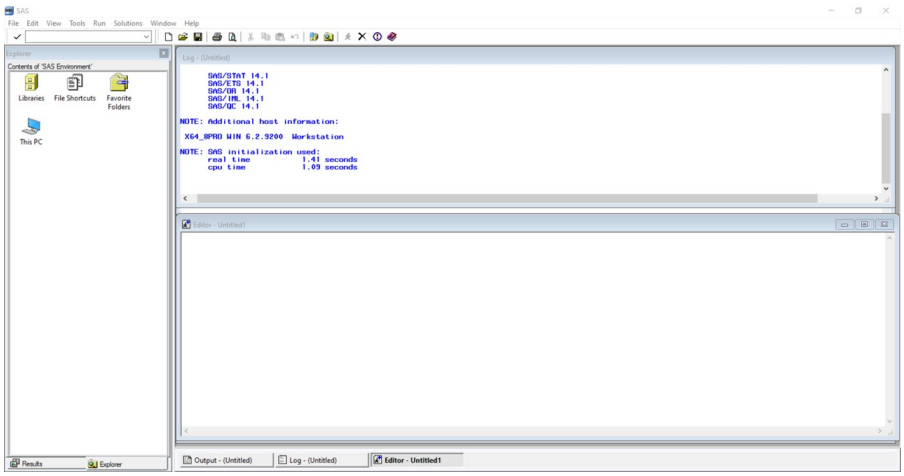


Figure 1.3: Visualization of **SAS**'s layout.

Within the field of data science the software packages Python and R are very common, while in field of statistics the packages **SAS** and R are very common. Here we have decided to use **SAS**, because **SAS** is very strong when it comes to the use of mixed models on real data. It can implement mixed effects models that are currently not possible in R and Python. Furthermore, **SAS** has an

extensive set of documentation on the details of the available procedures, the underlying statistical methodology underneath the procedures and SAS has several manuals with statistical content. Another advantage of the software package SAS is the validity of the procedures that have been implemented. SAS spends a lot of time of providing reliable procedures, while R and Python are generic repositories of procedures that can be attributed to by anyone who would like to make a contribution. The validity of the implemented procedures is often unknown or sparsely studied. Nooraee et al., (2014), demonstrated that packages in R do not always provide the correct results and this is a real concern if the results of a data analysis is used for inference. A disadvantage of SAS is that it works quite different from R and Python and thus not as easy to learn when you are used to the other two software packages.

In Figure 1.3 we provide a screen shot of SAS's layout when you would open the package. It shows a SAS environment, a log-screen, a editor in which the user can write the code, and an output screen that is not opened. The environment is connected to the computer on which the software has been installed. The log screen will provide information on the codes that are being processed (e.g., codes that have been submitted to the data, warnings and errors of the procedures). The output file provides the results of the procedures (e.g., parameter estimates, model fit information, likelihood information).

1.8.1 Importing data into SAS

The data can be entered into the SAS software package in different ways. One way is to type in the data manually. This would only be convenient when we would deal with relatively small data sets. In most other cases we would read in the data using the procedure PROC IMPORT (link documentation). The data must then be stored in some format on the computer (or somewhere else that SAS can access). SAS can handle many different data storage formats (e.g., CSV, TXT, Excel, DBF, or SPSS files). If the data is already in the SAS data format and it is stored on your computer somewhere, you can use the LIBNAME statement of SAS. To demonstrate these different options we will read in the data listed in Table 1.1.

Table 1.1: Simple data set for illustration purposes of SAS programming codes.

School	Child	Sex	SES	IQ
1	1	Girl	10	10.5
1	2	Girl		14.0
1	3	Boy	23	15.0
1	4	Boy	10	14.5
1	5	Girl	15	9.5
2	1	Girl	15	11.5
2	2	Boy	15	10.5
2	3	Boy	20	8.0
2	4	Girl	15	7.5
2	5	Boy	10	7.5

To manually read in the data, we would use the followin codes:

```
DATA schooldata;
  INPUT School Child Sex $ SES IQ;
  DATALINES;
  1  1  Girl  10  10.5
```

```
  1  2  Girl  .    14.0
  1  3  Boy   23   15.0
.....
  2  5  Boy   10   7.5
  ;
RUN;
```

In the first line of these codes a data set will be created that has the name schooldata. The second line introduces the names of the variables (as they have been given in Table 1.1). The variable Sex contains text labels. To tell **SAS** that we wish to use text als labels we put a dollar sign after the variable. The third line indicates that after this statement numbers or text will be given for each variable. Then each line represents a row in Table 1.1. To let **SAS** know which number belongs to which variable, a space is used between the values. In case of missing data for a variable, the punctuation "." symbol is used. After all records have been listed we close with ";" that is used for each line in the **SAS** codes. Then finally we use the text **RUN**; to end the data step procedure. To execute the programming codes, you select the full codes and press the "running man" on top of Figure 1.3.

To read in the data from some file (like Excel), we could use the following import statement. Here we have assumed that the data was stored in an Excel file on a memory disk with name E in the folder "E:\Onderwijs\2AMS10-LDA". The name of the Excel file was SAS_Tutorial.

```
PROC IMPORT
  OUT = WORK.Schooldata
  DATAFILE = "E:\Onderwijs\2AMS10-LDA\SAS_Tutorial.xlsx"
  DBMS = XLSX
  REPLACE;
RUN;
```

The OUT statement puts the data into the data set schooldata. The word WORK infront of the name of the data set (and connected with a punctuation) indicates that the data is being stored in the work file in the SAS system. When the SAS system is closed, all the data sets in the work file will vanish. The work file can be found in the file cabinet "Libraries" in the SAS environment in Figure 1.3. The DATAFILE is just a link to the Excel data set, where the full path is specified. The DBMS (Data Base Management System) is a statement that informs **SAS** about the type of data format that is being read in. Then finally, the REPLACE statement would replace any schooldata already present in the work file. Again, the **RUN** statement finishes the import procedure. To execute the programming codes, you select the full codes with your mouse and press the "running man" on top of Figure 1.3.

When the data was already stored on the computer in the **SAS** data format (sas7bdat) under the name Data_SAS_Tutorial, the following statements would have been easier.

```
LIBNAME SAS "E:\Onderwijs\2AMS10-LDA";
DATA WORK.SCHOOLDATA;
  SET SAS.DATA_SAS_TUTORIAL;
```

```
RUN;
```

The libname statement connects a name (here the word SAS, but any name would be possible) to a folder on the computer (indicated by the path). Now all data sets in this folder on the computer can be found in the file cabinet "Libraries". If you would click on "Libraries" with your mouse you would find the name SAS. If you would then click on this folder SAS in "Libraries" with your mouse you would see the data set `DATA_SAS_TUTORIAL` and you can open it from there. The next three lines of codes shifts the **SAS** data set `DATA_SAS_TUTORIAL` to the **SAS** data set `SCHOOLDATA` in the work file. One nice feature of the **SAS** system is that **SAS** is not sensitive to capitals. In the codes we used only capitals `DATA_SAS_TUTORIAL`, while the original name may be written quite differently, e.g., Data_SAS_Tutorial. This insensitivity to lower and capital letter would also apply to the names of the variables in the data set.

1.8.2 Calculating summary statistics with SAS

Now that we know how to import data into **SAS**, we need to learn how we can use **SAS**'s procedures to conduct data analysis. There exist many procedures and most of them work in the same way. To illustrate we will use the procedure **MEANS** (link documentation), which provides summary statistics from the imported data set. The following code could be used on the data in Table 1.1.

```
PROC MEANS DATA = SCHOOLDATA options;
  CLASS SCHOOL;
  VAR SES IQ;
RUN;
```

To use a procedure we start with **PROC** and then the name of the procedure (here **MEANS**). For most procedures it would be useful to indicate on which data this procedure should be conducted. The name of the data is `SCHOOLDATA` and it is located in the work file of the cabinet "Libraries". It would be more precise to write `WORK.SCHOOLDATA`, but if you do not specify the location in the file cabinet "Libraries", it will automatically be the work file. If the data is not yet read into SAS, you could have specified another location `SAS.SCHOOLDATA`, but then you should have run the libname statement to connect **SAS** to your data stored elsewhere. To calculate the summary statistic of interest, **SAS** has many calculation options you may find in the **SAS** manual and which can be mentioned at the place options. The default set of summary statistics (if we do not make use of the options) are the sample size, the average, the standard deviation, the minimum, and the maximum. The CLASS statement can be used to help calculate the summary statistics per strata (in this case summary statistics for the two levels of the variable `SCHOOL`). The VAR statement lists the variables of which the summaries should be calculated. This can only contain numerical variables. The output of the programming statement of **PROC MEANS** used on the data in Table 1.1 leads to the following self-explanatory output.

This output data can also be stored into a new **SAS** data set. This could be beneficial when additional analysis must be performed on the summary statistics or when the results must be stored digitally for later access. The programming

statements that would store the averaged, the standard deviations, and the sample sizes of the variables SES and IQ are

```
PROC SORT DATA = SCHOOLDATA;
  BY SCHOOL;
RUN;
PROC MEANS DATA = SCHOOLDATA NOPRINT;
  VAR SES IQ;
  BY SCHOOL;
  OUTPUT OUT = SUMMARY
      MEAN = M_SES M_IQ STD = S_SES S_IQ N = N_SES N_IQ;
RUN;
```

Table 1.2: SAS output of PROC MEANS on the data of Table 1.1

School	N Obs	Variable	Label	N	Mean	Std Dev	Min	Max
1	5	SES	SES	4	14.50	6.137	10.0	23.0
		IQ	IQ	5	12.70	2.515	9.5	15.0
2	5	SES	SES	5	15.00	3.536	10.0	20.0
		IQ	IQ	5	9.00	1.871	7.5	11.5

Before we summarize the data per stratum, we have to make sure we sort the data per stratum first. The procedure SORT (link documentation) sorts the data in order of the levels of the variable SCHOOL. With the BY statement, the sorting is conducted for each (combinations of) level(s) of the variable(s) involved in chronological order. The NOPRINT option in the MEANS procedure is used to make sure we do not get any output in the output file. This is useful when the data is summarized by many strata. With the BY statement the summarizing in conducted for each (combinations of) level(s) of the variables involved. The OUTPUT statement creates a new data set and the name of this data set was chosen equal to SUMMARY. Each type of summary statistic that needs to be calculated has to be mentioned in the same OUTPUT statement. The names of the variables in the summary data set are given the names M_SES and M_IQ for the means of SES and IQ, S_SES and S_IQ for the standard deviations, and N_SES and N_IQ for the sample sizes.

Variables like SEX which have been recorded with text information can not be summarized with the procedure MEANS, but they can be summarized using the procedure FREQ (link documentation). This procedure counts for each level of the variable listed in the procedure the number of records (or rows) it appears in the data set. To summarize the frequency of the levels of categorical variables the following statement can be executed. It has the same structure as the procedure MEANS, except the VAR statement is being replaced by TABLES statement.

```
PROC FREQ DATA = SCHOOLDATA;
  TABLES SCHOOL SEX;
RUN;
```

The output of this procedure consists of a table with the number of counts for each level of each variable separately. Here we have only provided the output of

Table 1.3: **SAS** output of **PROC FREQ** on the data of Table 1.1

Sex				
Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Boy	6	60.00	6	60.00
Girl	4	40.00	10	100.00

the variable `SEX` in Table 1.3. It represents counts, percentages and cumulative information. The order in accumulating the levels of the variable is chronological. For text information **SAS** uses the alphabetical order and for numeric information **SAS** uses the order of size (from small to large).

Procedure **FREQ** can also be operated with a BY statement to have frequency counts of a variable per stratum. However, it has also the option to make contingency tables for variables. The following programming codes will provide a 2×2 contingency table for the variables `SEX` and `SCHOOL`. The order of the variables determine which one will end up in the column and which one in the row. The first variable mentioned will be in the row of the contingency table.

```
PROC FREQ DATA = SCHOOLDATA options;
  TABLES SCHOOL*SEX;
RUN;
```

The output of this procedure is given in Table 1.4. The procedure also provides percentages of the counts of the total sample size, the row sample sizes, and the column sample sizes (which are not listed here). In this way the procedure **FREQ** is used to create summary statistics, but the procedure is also capable of calculating statistics from the contingency table (like Pearson's χ^2 statistics, Pearson's ϕ coefficient, Cramér's V, Goodman and Kruskal's γ, Cohen's (weighted) κ statistic, risk difference, relative risk, and odds ratio's). These methods can be implemented in the option statement. The procedure **FREQ** is also capable of calculating such measures of association from aggregated data. The data set would then exists of three variables, where one variable consists of the count of each cell and the other two variables list the combinations of levels of the variables that makes up the cells in the contingency table. Procedure **FREQ** then requires a WEIGHT statement (i.e., WEIGHT `COUNT`) that uses the count variable to indicate that each combination of levels originally occured many more times in the data.

Table 1.4: **SAS** output for a 2×2 contingency table

Table of school by sex			
School	Sex		
	Boy	Girl	Total
1	3	2	5
2	3	2	5
Total	6	4	10

To obtain the correlation coefficients between two numerical or continuous variables (like SES and IQ), we could use the procedure **CORR** (link documentation). The procedure can calculate Pearson's correlation coefficient, Sprearman's ρ correlation, and Kendall's τ correlation. The following programming statements provides Pearson's correlation coefficient.

```
PROC CORR DATA = SCHOOLDATA PEARSON;
  VAR SES IQ;
```

RUN;

The output is given in Table 1.5. The procedure also provides summary statistics for the variables involved, but here we just report the correlation information we are interested in. Furthermore, the procedure also provides the number of observations on which the correlation estimate is based and it automatically test the null hypothesis $H_0 : \rho = 0$, using the t-dsitribution. The p-value for this null hypothesis based on the data in Table was determined at $p = 0.6375$. The information can be stored using a so-called ODS (Output Delivery System) statement. Adding the statement "ODS OUTPUT PEARSONCORR = CORR_SES_IQ;" to the programming statements, results in a SAS data set CORR_SES_IQ in the work file with the information you want.

Table 1.5: Pearson's correlation between SES and IQ.

Pearson's Correlation		
	SES	IQl
SES	1.00000	0.18298
IQ	0.18298	1.00000

The procedure CORR also has the option to use the Fisher-z transformation with its corresponding 95% confidence interval we discussed in (1.10). Unfortunately, SAS uses by default a bias correction for the Fisher-z transformed correlation coefficient. The reason for this correction is that the expected value of $z_{\hat{\rho}_P} = 0.5[\log(1+\hat{\rho}_P) - \log(1-\hat{\rho}_P)]$, with $\hat{\rho}_P$ Pearson's correlation coefficient, is not exactly equal to $0.5[\log(1+\rho) - \log(1-\rho)]$. We do not recommend to use this bias correction, so we need to eliminate this calculation from the analysis. If we also wish to store the results into a SAS data set, the final programming statements are

```
ODS OUTPUT FISHERPEARSONCORR = CORR_SES_IQ;
PROC CORR DATA = SCHOOLDATA PEARSON FISHER(BIASADJ = NO);
  VAR SES IQ;
RUN;
```

The final correlation coefficient (in its original scale) would now be reported by 0.183 $[-0.548, 0.755]$. The p-value for testing the null hypothesis $H_0 : \rho = 0$ using the Fisher-z transformation is now dtermined at $p = 0.650$. Thus, the null hypothesis of no correlation is not rejected and we can not demonstrate (with the data in Table 1.1) that there is a correlation between the variables SES and IQ. Clearly, the data set is very small to have any power to reject the null hypothesis.

1.8.3 Other SAS programming information

In Section 1.8.2 we demonstrated that the output of a SAS procedure can be stored in a SAS data set. All outcome data of the SAS procedures is stored by default internally, but SAS uses separate files and has invented its own names. To know the names of the output files under which SAS stores the output data we can use the ODS TRACE statements around any procedure. To illustrate we use the procedure MEANS.

```
ODS TRACE ON;
PROC MEANS DATA = SCHOOLDATA;
  VAR SES IQ;
```

```
RUN;
ODS TRACE OFF;
```

The log file provides a list of all output files that the programming statements are producing. The more options we may include in a procedure the more output files could be listed. Thus, we only see the the names of the output files we create with the programming statements, and not all output files that the procedure is potentially capable of making when we would use all options. Since we only request a single output file with the current programming statement, we only see the name `SUMMARY` in the log file[4]. To store this file we can add an ODS statement again, like we did for the procedure CORR. However, we have already seen in Section 1.8.2 that the procedure **MEANS** has a separate output statement. Not all procedures have separate output statement though.

Similar to the packages R and Python, we can also manipulate data in a **SAS** data set, but R and Python are much more flexible in their calculations than **SAS**. These mathematical manipulations are conducted row-wise, since **SAS** stores the data in columns, which makes it more difficult to conduct manipulations for the data that is stored in one column. To illustrate the manipulations, let's consider the data set `SUMMARY` that we created with the procedure **MEANS** in Section 1.8.2. In this data set we had collected the average and standard deviation of the two numerical variables per school. The calculation of a 95% confidence interval on the average IQ and SES per school is rather straightforward. The following data step will provide use these intervals.

```
DATA SUMMARY;
  SET SUMMARY;
  LCL_SES = M_SES - TINV(0.975,N_SES-1)*S_SES/SQRT(N_SES);
  UCL_SES = M_SES + TINV(0.975,N_SES-1)*S_SES/SQRT(N_SES);
  LCL_IQ = M_IQ - TINV(0.975,N_IQ-1)*S_IQ/SQRT(N_IQ);
  UCL_IQ = M_IQ + TINV(0.975,N_IQ-1)*S_IQ/SQRT(N_IQ);
RUN;
PROC PRINT DATA = SUMMARY;
  VAR SCHOOL LCL_SES UCL_SES LCL_IQ UCL_IQ;
RUN;
```

The programming codes are provided in what we call a data step. We put the results of our calculations in a data set called `SUMMARY` and we use the available data set `SUMMARY` with the statement SET `SUMMARY`. We could have changed the name of the data set in the first line to make a new data set, but here we chose to overwrite the data set `SUMMARY`. The calculations are now conducted per row, which means we would obtain confidence intervals for each school separately.

Table 1.6: 95% Confidence intervals on the mean SES and IQ per school.

School	SES		IQ	
	LCL	UCL	LCL	UCL
1	4.73	24.3	9.58	15.8
2	10.6	19.4	6.68	11.3

[4]The name `SUMMARY` for the output file of the procedure **MEANS** coincides with the name for the data set we used in the OUTPUT statement of procedure **MEANS** in Section 1.8.2. This is a coincidence, since we could have used a completely different name in the OUTPUT statement, like `DESCRIPTIVES`.

To calculate the confidence intervals we used two functions that are available in SAS: the function TINV that calculate the quantiles of the t-distribution and the function SQRT that calculates the square root. The function TINV requires at least two inputs, the quantile value (here 0.975) and the number of degrees of freedom. The mathematical operations `-`, `+`, `*`, and `/` represent subtraction, summation, multiplication, and diving, respectively. The procedure **PRINT** prints the results of the variables `LCL_SES`, `UCL_SES`, `LCL_IQ`, and `UCL_IQ` in the **SAS** output file per school. Table 1.6 shows the results of the calculations and the print statement.

To calculate the difference in averages between the two schools manually is more elaborate with **SAS**. One way is to make two separate summary data sets: one for each school and then merge the results of the two data sets into one data set so that the averages are in one row to be able to manipulate. Making the two data sets with the average IQ can be done as follows:

```
DATA SUMMARY1;
  SET SUMMARY;
  WHERE SCHOOL = 1;
  M_IQ1 = M_IQ;
  S_IQ1 = S_IQ;
  N_IQ1 = N_IQ;
  KEEP M_IQ1 S_IQ1 N_IQ1;
RUN;
DATA SUMMARY2;
  SET SUMMARY;
  WHERE SCHOOL = 2;
  M_IQ2 = M_IQ;
  S_IQ2 = S_IQ;
  N_IQ2 = N_IQ;
  KEEP M_IQ2 S_IQ2 N_IQ2;
RUN;
```

The data steps have now created two separate data sets with the IQ descriptives of school 1 in data set `SUMMARY1` and the descriptives of school 2 in data set `SUMMARY2`. The WHERE statement selects the rows of the data set that is read in (here data set `SUMMARY`) for which the statement is true. Thus the statement "`WHERE SCHOOL = 1;`" selects all rows in data set `SUMMARY` where the variable `SCHOOL` is equal to the value one. With the KEEP statement we can select the variables we wish to keep in the data set. Merging the two data set `SUMMARY1` and `SUMMARY2` is now conducted by the programming statements:

```
DATA MEAN_DIFF;
  MERGE SUMMARY1 SUMMARY2;
RUN;
```

If we merge data sets the names of the variables should be different, unless we would merge a data set on one or more variables. The merging is then conducted for each (combination of) level(s) of the merging variable. The merged data set `MEAN_DIFF` has now six colums with names `M_IQ1`, `S_IQ1`, `N_IQ1`, `M_IQ2`, `S_IQ2`, and `N_IQ2` and with just one row where the descriptive value is stored.

1.8.4 SAS procedures for repeated outcomes

SAS has many different procedures for all kinds of statistical analyses. It goes too far to list them all here. The procedures that are relevant to longitudinal data are the procedures MIXED, HPMIXED, GENMOD, GLIMMIX, and NLMIXED. The HPMIXED is a high-performance version of the procedure MIXED. This means that HPMIXED can better handle large data sets than procedures MIXED, but HPMIXED does not have all the options of procedure MIXED. The procedure MIXED and HPMIXED assume that the repeated outcome variable is normally distributed (which we will assume throughout the lecture notes). The procedure GENMOD and GLIMMIX were essentially developed for modeling repeated outcomes that do not follow a normal distribution. The procedure GENMOD typically was developed for marginal models, while GLIMMIX was developed for subject-specific models, To understand the difference between marginal and subject-specific models we refer to Section 2.10. When the data is normally distributed marginal and subject-specific models can both be handled with procedure MIXED (and HPMIXED). The procedure NLMIXED makes it possible to program your own models. You could specify your own likelihood function, but it also has incorporated certain options to help built your likelihood with the inherent components. We will mostly discuss the procedures MIXED and a little bit NLMIXED and HPMIXED.

1.9 Case studies

Throughout the lecture notes we will make use of several data sets. These data sets are either publically available or we have created the data sets ourselves. When we have created the data sets ourselves, they are still based on real data sets we have worked with in the past. This means that we use real data, but that we have made adjustements to the data to make sure we do not jeopardize the privacy of real individuals.

1.9.1 A case study on children's cognitive capacity

To help understand ANOVA models and its use on real data, we will make use of a case study for which the data was made publicly available. The data has been described in the literature (Brandsma and Knuver, 1989; Snijders and Bosker, 1999). We used the file `mlbook2_data_preparations.zip` from website http://www.stats.ox.ac.uk/~snijders/mlbook.htm#data.

This study contains 4106 pupils or children in grade seven and eight. They are approximately eleven years old. The children were collected from 216 elementary schools in the Netherlands. They were evaluated twice on a Dutch language and arithmetic test. The first time in grade 7 and the second time in grade 8, approximately one year later. They also conducted a verbal and performal IQ test. We constructed from the data file the following variables and we will refer to this data set as `schooldata`. The data is available in a SAS data set.

Variables at school and/or class level:

- ▷ `SCHOOL`: Indicator variable for school.
- ▷ `CLASS`: Indicator variable for class.

- ▷ `COMBI`: Indicator variable for students being taught in a multi-grade class or not ($0 = \text{no}; 1 = \text{yes}$). A multi-grade class combines grade 7 and 8.
- ▷ `SIZE`: Class size, which can be different from the observed number of children since not all children in a class may have participated in the tests.
- ▷ `SSES`: Social economic status at school level.

Variables at child level:

- ▷ `CHILD`: Indicator variable for individual.
- ▷ `GIRL`: Binary indicator for the child being a girl ($0 = \text{no}$; $1 = \text{yes}$).
- ▷ `CSES`: Social economic status of the child's family.
- ▷ `MINORITY`: Binary indicator for being born outside an industrialized country ($0 = \text{no}$; $1 = \text{yes}$).
- ▷ `SITTERS`: The number of times the child repeated a grade.
- ▷ `IQV`: Verbal IQ test score.
- ▷ `IQP`: Performal IQ test score.
- ▷ `PRE_LANG`: Language test score at grade 7
- ▷ `POST_LANG`: Language test at grade 8.
- ▷ `PRE_ARITH`: Arithmetic test score at grade 7.
- ▷ `POST_ARITH`: Arithmetic test score at grade 8.

The main research goal of this study by the original researchers, was to investigate how school related organizational factors affect children's progress in proficiency of language and arithmetic (Brandsma and Knuver, 1989) and how much the language test score depends on the children's intelligence and his/her social economic status (Snijders and Bosker, 1999). We will use this data set to illustrate ANOVA models with all its features but will also use it to demonstrate the limitations of ANOVA models.

1.9.2 A case study on anthropometric growth of children

1.9.3 A phase II clinical trial study on pain treatment

2 Analysis of variance models

ANalysis Of VAriance (ANOVA) is a statistical analysis technique that was developed in the first half of the 20th century by Sir Ronald Fisher. Before the 20th century, this technique was applied only for specific examples without describing its generic mathematical characteristics (see Section 2.2). Originally, ANOVA was used to quantify and test only differences between means, the so-called fixed effects ANOVA (see Section 2.1). However, over more than 50 years of research, ANOVA has undergone tremendous developments to be able to deal with all kinds of correlated data sets that are observed under different conditions (see Searle *et al.*, 2006). It grew into what is called the random and mixed effects ANOVA. Instead of just focusing on differences in means, it has become a statistical modeling approach where correlations between observations could be properly described.

Due to these developments, ANOVA has found its way into many different application areas. To name just a few:

Agriculture: In agriculture, it is important to maximize (crop) yield with minimal use of pesticides. Over the years, all kinds of experiments have been performed to support such endeavors. These experiments are not straightforward because testing a condition is confounded by the quality of the soil. Thus rather complicated experimental designs with repeated plots were invented to deal with these settings (block designs, Latin squares, split-plot designs). Also, the analysis of the observed data became more complicated and they helped to develop ANOVA models.

Measurement reliability: Collecting data often involves some kind of measurement instrument. For instance, blood pressure is obtained with a sphygmomanometer. The cognitive ability of a person is obtained through well-defined psychological tests. The strength of plastic is obtained with an extensometer, that quantifies how much the plastic can be stretched before it ruptures. The activity of a medical drug is measured with biological cells or with animals in a so-called bioassay. To quantify the accuracy and precision of a measurement instrument or system, ANOVA has played an important role. ANOVA can help separate the different sources of variability.

Clinical Trials: New treatments must be tested before it can be marketed and commercial medicinal products are being monitored for effectiveness and side effects. This is typically done in clinical trials. Like in agriculture, trial designs can be complicated with different sources of variation (e.g., cluster randomized trials, stepped wedge designs, cross-over designs). The collected trial data is then often analyzed with ANOVA models to be able to obtain an unbiased and precise estimator of the treatment effect.

Quality improvement: The quality of products and services is often determined by measuring their quality characteristics. These quality characteristics should often stay within certain limits to guarantee a certain performance. The quality of products and services is often increased when the variability between products and services can be reduced. To make quality improvements, ANOVA is often used to quantify the variability between products and services and to help minimize variability across products of the same type.

Heritability: Understanding the genetic contribution in the occurrence of diseases is an old and important field of study. ANOVA can help quantify this

contribution when a specific phenotype is observed in several family members. This field has also led to the introduction of intraclass correlation coefficients, discussed in these lecture notes.

Nowadays, ANOVA can be viewed as a subset of modeling techniques in the much larger class of mixed effects models (see Section 2.10). The mixed effect models can handle more complicated correlation structures and handle more complicated associations between different variables than ANOVA models can. In longitudinal data sets, this could be relevant. Nevertheless, a good understanding of ANOVA is very valuable. What Sir Ronald Fisher said about ANOVA when he wrote a letter to George Snedecor in 1934 (see Searle *et al.*, 2006) is still relevant today

> *"The analysis of variance is (not a mathematical theorem but) a simple method of arranging arithmetical facts so as to isolate and display the essential features of a body of data with the utmost simplicity"*

2.1 Terminology

The data structure for an ANOVA is organized in a specific way, also called the *data layout* or *data design*. First of all, there exists a continuous or numerical outcome, like blood pressure, crop yield, tensile strength, bioactivity, IQ score, etc. The numerical variable is the variable of interest for which we would like to understand how it is influencing or how it can be influenced by other variables. In ANOVA, these other variables are typically categorical. They are called *factors* and have only a fixed set of *levels*, like sex, type of fertilizer, type of plastic, dosage, or education. However, factors can also be individuals, families, schools, cities, hospitals, etc. These factors together form a specific layout of the data on the numerical variable. Table 2.1 shows an example of a specific layout where boys and girls were repeatedly measured on their cognitive performance in grade 7 and grade 8 (see Section 1.9.1). In each cell of the layout we would collect a certain number of units with a value of the numerical variable on each unit (in this case a cognitive test score on children). If the number of units in each cell is constant, the data is called *balanced*. Otherwise, the data is called *unbalanced*.

Table 2.1: Example data layout

Sex	Grade	Schools		
		1	$\cdots$	206
Girls	7			
	8			
Boys	7			
	8			

The influence of factors on the numerical outcome is typically referred to as the *effects* of a factor. Girls may have a higher cognitive performance in grade 7 than boys (or the other way around). The effect is typically determined by the mean difference in cognitive performance. The same is true for the factor school. Each school may have its own average level of cognitive performance and the differences between the schools are called the school effects. There are as many effects as there are levels. We can treat the effects either as *fixed effects* or as *random effects*. For fixed effects, the levels of a factor enter the ANOVA model as a mean parameter, while for random effects the levels of a factor enter the ANOVA model as a random variable. Fixed or random is often a choice by the researchers, but there are a few rules that are used among all researchers:

- ▷ **Treat as fixed effects:** When all possible levels of the factors are included in the study or when the study is only interested in quantifying the mean differences (the so-called *contrasts*) between the levels of the factor that was included.
- ▷ **Treat as random effects:** When the levels of a factor are a random selection of all possible levels of this factor or when the interest is in quantifying the variability (the so-called *variance components*) in the outcome for the levels of the factor.

Thus the factor sex is typically treated as a factor that contributes to the statistical model with only fixed effects, while the factor school can be treated as either fixed or random. If a study was focusing on understanding differences between a few schools, to improve the education within some of the schools, then the factor school would be treated as a fixed effects factor. We would be interested in the mean differences between schools on several numerical outcomes. However, if the factor schools is a way of collecting the information on children in a country or region, then schools can be viewed as a random sample from all possible schools in that region that could have participated in the study. In this setting, schools would typically be treated as a random effects factor.

If the ANOVA model contains only fixed effects, then the ANOVA model is referred to as *fixed effects ANOVA*. If it contains only random effects, the ANOVA model is called a *random effects ANOVA* model. In the case that fixed and random effects are included, the ANOVA model is called a *mixed effects ANOVA* model. In mixed effects and random effects ANOVA models, the random parameters are not directly estimated, but the variability in these coefficients is estimated. Thus the statistical analysis would estimate the *variance* (and possibly *covariance*) *components* that are associated with the sets of random effects. Then based on the fitted model, the individual random effects can be retrieved by using a technique called *Empirical Bayes*.

Factors can be viewed as *crossed* or *nested* factors. Two factors A and B are considered crossed when the levels of factor A are identical across the levels of factor B (or the other way around). In Table 2.1 the factors school and grade are crossed. Grade 7 and grade 8 mean the same thing within each school. When factors are crossed, the influence or *effect* of one factor on the outcome can depend on the levels of the other factor. These types of effects are referred to as *interaction effects*. To study interaction effects we use the notation AB to indicate the cross product. For instance, the difference in cognitive performances of children in grade 7 and grade 8 can depend on the school the children are going to. Factor A is nested within factor B when the levels of factor A are different for the levels of factor B. For instance, shools in Table 2.1 could be organized into schools that either have single classes for grade 7 and grade 8 and schools that combine seven and eight grade classes. Thus there may exist another factor called multi-grade classes (yes/no). In this example, the factor school is nested within the factor multi-grade classes, because schools with multi-grade classes are typically different schools than schools with single classes for each grade.

The goal of ANOVA is to partition the variability in the numerical outcome by the different crossed and nested factors and their potential interaction effects. The ANOVA model would describe all the fixed and random effects of the factors and their interactions. These elements of the ANOVA model are referred to

as the *terms* in the statistical model. Additionally, there may be variability unexplained by these factors and their interactions. The unexplained variability is referred to as the *residual* or *error term* of the ANOVA model. Based on the ANOVA model the following results can be obtained:

- ▷ **Estimate contrasts:** Quantify or estimate the effect sizes of (the interactions of) fixed effects factors.
- ▷ **Hypothesis testing:** Test the contribution of a factor on the outcome as a whole or test if differences between two levels are significant.
- ▷ **Estimate variance components:** Quantify the variability in the outcome that is associated with the (interactions of) factors.
- ▷ **Correlation analysis:** Quantify how much each factor contributes to the overall variability in the outcome.

2.2 A brief history of ANOVA models

As mentioned in the beginning of this chapter, ANOVA was developed by Sir Ronald Aylmer Fisher (1890-1962). The simplest ANOVA model is the *balanced one-way fixed effects model*. It is defined by

$$y_{ij} = \mu + \alpha_i + e_{ij}, \tag{2.1}$$

with y_{ij} the numerical outcome for unit $j \in \{1, 2, ..., n_0\}$ in group $i \in \{1, 2, ..., m\}$, μ is the overall mean, α_i a fixed effects parameter for group i with constraint $\alpha_1 + \alpha_2 + \cdots + \alpha_m = 0$, e_{ij} an unobserved random error term having a normal distribution with mean zero and variance σ_E^2, i.e., $e_{ij} \sim N(0, \sigma_E^2)$, and with all the error terms being independently distributed. The constraint on the fixed effects parameters α_i is needed for model identifiability. Without the constraint, the model parameters are not all estimable. There is one parameter too many.

Applications of the fixed effects ANOVA for Fisher was in agriculture where he studied the yield of plants (unit j) for different varieties (group i). He was interested in the fixed effects α_i to identify which variety could lead to the largets yield. In this setting, the ANOVA method was a natural extension of the t-test where only two groups ($m = 2$) are being compared.

Earlier accounts of the fixed effects one-way ANOVA model go back to Adrien-Marie Legendre (1752-1833) and Carl Friederich Gauss (1777-1855) in the 19th century, who both studied problems from astronomy (Searle *et al.*, 2006). They invented and developed the *ordinary least squares* estimation technique (Sheynin, 1993) that is still relevant today.

Another account of the one-way ANOVA model, but where the fixed parameters α_i were treated as random terms a_i, was by George Biddell Airy (1801-1892). The *balanced one-way random effects model* is of the same form as in (2.1), but where α_i would be replaced by the random variable a_i.[5] Instead of making

[5] We will use roman letters for random effects and random variables and we will use greek letters for fixed effects and parameters. In case we estimate a fixed effects parameter, we may use a hat on top of the parameter to indicate that it is an estimator and therefore also random.

constraints on these random effects, the random effects a_i are considered independent and identically normally distributed, $a_i \sim N(0, \sigma_G^2)$, and independent of the error terms e_{ij}. Airy also studied astronomic phenomena by studying the same atmospheric phenomena at different nights and used different numbers of observations per night ($j \in \{1, 2, ..., n_i\}$). The nights were considered some random draw of many more nights that could have been studied as well. Thus the unbalanced one-way random effects ANOVA model can be written as

$$y_{ij} = \mu + a_i + e_{ij}, \tag{2.2}$$

with $j \in \{1, 2, ..., n_i\}$ and $i \in \{1, 2, ..., m\}$, μ is the overall mean, $a_i \sim N(0, \sigma_G^2)$ independent random effects, $e_{ij} \sim N(0, \sigma_E^2)$ independent error terms, and with a_i and e_{ij} independent.

Interestingly, Airy estimated the *residual variance component* σ_E^2 by squaring the average standard deviation over the groups, i.e., $\hat{\sigma}_E^2 = [\sum_{i=1}^{m} S_i/m]^2$, with $S_i = [\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/(n_i - 1)]^{1/2}$ the standard deviation for group i, $\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i$ the average for group i, and n_i the sample size for group i. Thus in the early days, the notion of making use of the *degrees of freedom* $df_i = n_i - 1$ within groups for the calculation of sample variances was well understood. However, nowadays we would estimate the residual variance by $\hat{\sigma}_E^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2/(n - m)$, with $n = \sum_{i=1}^{m} n_i$ the total sample size and $n - m$ the associated degrees of freedom.

In the context of genetics, Fisher developed the general principles for estimating the parameters of the balanced one-way random effects ANOVA ($n_i = n_0, \forall i$) in the early part of the 20th century. He used this model to quantify the correlation between pairs of units in the same group. He calculated two sums of squares, similar to what was used for the one-way fixed effects analysis:

$$\begin{array}{rcrcl} \text{Between groups} & : & SS_B & = & \sum_{i=1}^{m}[n_0(\bar{y}_{i.} - \bar{y}_{..})^2], \\ \text{Within groups} & : & SS_W & = & \sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij} - \bar{y}_{i.})^2, \end{array} \tag{2.3}$$

with $\bar{y}_{..} = \sum_{i=1}^{m}\sum_{j=1}^{n_0}[y_{ij}/(mn_0)]$ the *overall* or *grand average*. Fisher calculated the *expected sums of squares* and found that $\mathbb{E}[SS_W] = m(n_0 - 1)\sigma_E^2$ and $\mathbb{E}[SS_B] = (m - 1)[n_0\sigma_G^2 + \sigma_E^2]$. By solving these equalities and replacing the expected values by the corresponding sums of squares, Fisher determined the *ANOVA estimators* or, in other words, the *method of moment estimators* for the variance components σ_E^2 and σ_G^2:

$$\begin{array}{rcl} \hat{\sigma}_E^2 & = & MS_W, \\ \hat{\sigma}_G^2 & = & [MS_B - MS_W]/n_0, \end{array} \tag{2.4}$$

with $MS_W = SS_W/[m(n_0 - 1)]$ and $MS_B = SS_B/(m - 1)$ the within and between group mean squares, respectively. The mean parameter μ would be estimated by the overall average: $\hat{\mu} = \bar{y}_{..}$, an estimator that was already used for the balanced one-way random effects model many decades earlier by William Chauvenet (1820-1870). He even established the variance of this estimator to be equal to $\mathsf{VAR}(\bar{y}_{..}) = [\sigma_G^2 + \sigma_E^2/n_0]/m$ in 1863.

Fisher used these variance component estimators to estimate the intraclass correlation coefficient:

$$\mathsf{ICC} = \sigma_G^2/[\sigma_G^2 + \sigma_E^2]. \tag{2.5}$$

This coefficient quantifies how much the units from the same group are alike. It was a way of quantifying the heritability in certain phenotypes. An intraclass correlation coefficient equal to one would indicate that all units within one group would be identical, while an intraclass coefficient of zero suggests that units within one group are as alike as units between groups. The natural estimator of the intraclass correlation coefficient is to substitute the variance component estimators: $\hat{\mathsf{ICC}} = \hat{\sigma}_G^2/[\hat{\sigma}_G^2 + \hat{\sigma}_E^2]$.

The intraclass coefficient has found many applications, but especially in measurement reliability since the ICC can quantify the agreement among repeated measures on a single unit. It is straightforward to demonstrate that the intraclass correlation coefficient is equal to the correlation of two outcome values within the same group: $\mathsf{ICC} = \mathsf{CORR}(y_{ir}, y_{is})$, with $r \neq s$.

The construction of confidence intervals on the ICC has been a topic of research for many decades (Wald, 1940; Donner, 1986; Demetrashvili *et al.*, 2016). A general principle for the construction of confidence intervals on ICC's from any mixed effects ANOVA model will be discussed in Section 2.9.4. For the balanced one-way random effects model, there exists an *exact confidence interval* on the ICC, but this can not be guaranteed for unbalanced data. An exact confidence interval means that the probability that the interval contains the true parameter is exactly equal to the imposed confidence level. The exact $100\%(1-\alpha)$ confidence interval is given by

$$\left[\frac{F/F_U - 1}{F/F_U + n_0 - 1}, \frac{F/F_L - 1}{F/F_L + n_0 - 1}\right], \tag{2.6}$$

with $F = MS_B/MS_W$ the ratio of the mean squares (often used for hypothesis testing), F_U the $\alpha/2$ upper quantile of the F-distribution with $m-1$ and $m(n_0-1)$ degrees of freedom, and F_L the $\alpha/2$ lower quantile of the F-distribution with $m-1$ and $m(n_0-1)$ degrees of freedom. In Section 2.4.8 we will explain how such an exact confidence interval can be created for specific ANOVA models, but in Section 2.9.4 we will provide a more generic approach that is applicable to any ANOVA model.

Although Fisher developed the principles for parameter estimation in the random and mixed effects ANOVA models, he never studied more complicated random and mixed effects ANOVA model than the one-way layout. It was Lenoard Henry Caleb Tippett (1902-1985) who extended the balanced one-way random effects model to the *balanced two-way crossed random effects model* around the 1930's:

$$y_{ijk} = \mu + a_i + b_j + e_{ijk}, \tag{2.7}$$

with y_{ijk} the kth value of the outcome for the levels $i \in \{1, 2, ..., I\}$ and $j \in \{1, 2, ..., J\}$ of the factors A and B, μ the overall mean, $a_i \sim N(0, \sigma_A^2)$, $b_j \sim N(0, \sigma_B^2)$, $e_{ij} \sim N(0, \sigma_E^2)$, and all random terms being mutually independent. To be a balanced design, we require K observations ($k \in \{1, 2, ..., K\}$) for each combination of factor levels. Unfortunately, Tippett did not study a possible random interaction term $(ab)_{ij} \sim N(0, \sigma_{AB}^2)$ in model (2.7), while interaction effects where very common in fixed effects analyses (Searle *et al.*, 2006). With the calculation rules for expected mean squares explained in Section 2.4, which were developed many years later, the estimators for the variance components in

the balanced two-way random effects model can be easily established at

$$\begin{array}{rcl} \hat{\sigma}_A^2 & = & [MS_A - MS_E]/I, \\ \hat{\sigma}_B^2 & = & [MS_B - MS_E]/J, \\ \hat{\sigma}_E^2 & = & MS_E, \end{array} \tag{2.8}$$

with MS_A, MS_B, and MS_E the mean squares for factor A, factor B, and the residual of model (2.7). These mean squares are defined by

$$\begin{array}{l} MS_A = \frac{1}{I-1}\sum_{i=1}^{I} JK(\bar{y}_{i..} - \bar{y}_{...})^2, \\ MS_B = \frac{1}{J-1}\sum_{j=1}^{J} IK(\bar{y}_{.j.} - \bar{y}_{...})^2, \\ MS_E = \frac{1}{IJK-I-J+1}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2, \end{array} \tag{2.9}$$

with $\bar{y}_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K}[y_{ijk}/(JK)]$, $\bar{y}_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K}[y_{ijk}/(IK)]$, and $\bar{y}_{...} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}[y_{ijk}/(IJK)]$ the averages for level i of factor A, level j of factor B, and the overall average, respectively.[6]

If we would add the random interaction effect $(ab)_{ij} \sim N(0, \sigma_{AB}^2)$ to model (2.7), the variance component estimators for σ_A^2 and σ_B^2 in (2.8) would change to $\hat{\sigma}_A^2 = [MS_A - MS_{AB}]/I$ and $\hat{\sigma}_B^2 = [MS_B - MS_{AB}]/I$, respectively, with the mean squares MS_A and MS_B defined in (2.9) unchanged, MS_{AB} the mean squares for the interaction effects,

$$MS_{AB} = \frac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J} K(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2,$$

and $\bar{y}_{ij.} = \sum_{k=1}^{K}[y_{ijk}/K]$ the average for level i and j of factors A and B. The variance component σ_{AB}^2 is then estimated by $\hat{\sigma}_{AB}^2 = [MS_{AB} - MS_E]/I$, where the mean squares MS_E for the residual has changed t0

$$MS_E = \frac{1}{IJ(K-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2.$$

The reason for the changed variance component estimators and the change in MS_E is that the residual mean square MS_E in (2.9) is also including information on the variance component for the interaction effects if we would include this in model (2.7). In the current model (2.7), where these interaction effects are considered non-existing, there is no need to calculate the mean squares MS_{AB} for model (2.7). Therefore, if some interaction effect would exist, it will end up in the residual mean squares, and its expectation will be part of the expectation of the mean squares MS_A and MS_B. Inclusion of interaction effects in crossed ANOVA models was only investigated in the 1940s (Searle et al., 2006).

In this period, a landmark paper by Franklin E. Satterthwaite was published (Satterthwaite, 1946). This paper discussed the distribution of linear combinations of independently and chi-square distributed variables (i.e., $\sum_{k=1}^{p}\eta_k X_k$,

[6] In two-way ANOVA models and models with more than two factors it is common to use the letters of the factors: in this case A for factor A, and B for factor B. If the factors would represent a certain variable, like days and class, the letters D and C would be used. For the residuals the letter R or E is used. Clearly, the R refers to residual, while the letter E refers to the error term. In one-way ANOVA models it is more common to use W for within groups and B for between groups.

with $X_k \sim \chi^2_{d_k}$). It was well-known that this distribution is chi-square when the coefficients in the linear combination were all the same (i.e., $\eta_k = \eta_0$). However, for other linear combinations, the distribution deviates from the chi-square distribution, but it is approximately chi-square distributed, with a degrees of freedom that must be calculated from the coefficients η_k and the degrees of freedom d_k (Satterthwaite, 1946). The importance of this paper is that it can be applied to linear combinations of mean squares since mean squares have a direct relation to the central or non-central chi-square distribution in balanced designs (see Section 2.4.4). Satterthwaite degrees of freedom are used for hypothesis testing to determine the statistical significance of factors and for construction of confidence intervals. For hypothesis testing purposes, it has been implemented in many software packages (including **SAS**).

In the same period (around the 1930s) as the two-way crossed random effects model (2.7), the *balanced two-way crossed mixed effects* model without interaction was discussed in literature as well (Searle et al., 2006). In the mixed effects model, one of the factors is treated as fixed, say factor A, while the other factors is treated as random. This means that the effects enter the statistical model in the same way as we have done in the one-way fixed effects ANOVA in (2.1), including the constraints we addressed for model (2.1). The mean squares defined in (2.9) remain unchanged, but the variance component σ^2_A would not exist and does not have to be estimated. Instead, the fixed effects α_i are estimated by $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$.

With the introduction of the two-way crossed ANOVA models, a serious deficiency of the ANOVA method became visible. The variance component estimators can become negative, leading to a negative estimate for a parameter that is non-negative (Ganguli, 1941; Crump 1946). It is common practice to truncate such estimates to zero, to avoid a negative estimate, but truncation as a general principle would make the moment estimators biased (see Section 2.4.5). The issue of a negative estimate can happen to any variance component in any mixed effects ANOVA model (including the one-way ANOVA model), except for the estimation of the residual variance component. This estimator will never be non-negative. The feature of negative estimates for the variance components with the methods of moments is often used as an argument to avoid the method of moments as an estimation approach, but this is not entirely fair (see the discussion in Section 2.7).

The early work on random and mixed effects analysis were for balanced designs. For such designs, estimation is straightforward and the estimators have very nice properties under normality, as we will see in Section 2.4. In unbalanced designs though, estimation of the variance components is not unique anymore and many ways of estimation exists (see Section 2.5). The early work on unbalanced designs by Airy for the one-way random effects model in the 1860s is a special case where uniqueness of estimation remains, as in some other special ANOVA models, but in most *higher-order ANOVA models* issues with estimation remains (Section 2.5). The unbalanced one-way random effects model in (2.2) was extensively discussed by William Gemmell Cochran (1909-1980) in 1939, although he was not specifically interested in estimating the variance components (see Searle *et al.*, 2006). The variance component estimators for this unbalanced

model are given by

$$\begin{array}{rcl} \hat{\sigma}^2_E & = & MS_W, \\ \hat{\sigma}^2_G & = & (m-1)[MS_B - MS_W]/[n - \sum_{i=1}^{m}[n_i^2/n^2]], \end{array} \tag{2.10}$$

with the mean squares given by

$$\begin{array}{l} MS_W = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2, \\ MS_B = \frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2. \end{array}$$

It can be shown that the estimators in (2.10) reduces to the estimators in (2.4) when the design becomes balanced ($n_i = n_0$, $\forall i$). Cochran also showed that the constant $(n - \sum_{i=1}^{m}[n_i^2/n])/(m-1)$ is always smaller than the average sample size $\bar{n}_{.} = n/m$.

In the 1950s and 1960s, higher-order (e.g., three-ways or higher) and *nested ANOVA models* are being discussed for balanced and unbalanced designs. In this period different estimation approaches are being proposed for unbalanced designs. The method of moment approach for estimation of variance components was further developed by Charles Roy Henderson (1911-1989) in 1953. Despite the non-uniqueness of the method of moments for unbalanced data, he described three ways of partitioning the sums of squares for unbalanced designs, referred to as methods I, II, and III. His third method is the most general method of moments approach that can be applied to any ANOVA model (Henderson, 1953). These methods are discussed in Section 2.5.2 and the third method is implemented in most software programs.

Due to the non-uniqueness of the method of moments and its numerical complexity for unbalanced data in these pre-computer days, alternative approaches were sought. In a landmark paper by Herman Otto Hartley (1912-1980) and J.N.K. Rao (1937), the *maximum likelihood* (ML) approach as a generic estimation method for estimating the fixed and random effects parameters from any type of ANOVA model was proposed in 1967. This generic formulation required a formulation of any ANOVA model in terms of matrices. We will discuss the ML method in Section 2.6. A disadvantage of the ML estimation is that it may typically underestimate the variance components since it does not incorporate the appropriate number of degrees of freedom (See Section 2.6). This is a well-know issue, since the ML estimator for the variance of a normal distribution based on a simple random sample of m observations (i.e., $y_i \sim N(\mu, \sigma^2)$) leads to $\hat{\sigma}^2 = \sum_{i=1}^{m}[(y_i - \bar{y}_{.})^2/m]$, instead of using $m-1$ as denominator. Although the bias of ML estimators is more complex than in this example (Section 2.6), it led to the development of the *restricted* or *residual maximum likelihood* (REML) estimation. The restricted likelihood maximizes a likelihood function of a linear combination of the observations, such that the fixed effects are eliminated from the likelihood (see Section 2.6). The REML estimator for σ^2 in our simple example would be equal to $\hat{\sigma}^2 = \sum_{i=1}^{m}[(y_i - \bar{y}_{.})^2/(m-1)]$, including the degrees of freedom of $m-1$ instead of the sample size m.

An advantage of the ML and REML estimation methods, often used as an argument to choose for ML or REML (McCullagh and Searle, 2000; Searle et al., 2006), is their constraint to non-negative estimates of the variance components, that is not automatically implemented in the method of moments. But this constraint leads to bias (see Section 2.6), similar to the truncation of the moment estimators to zero.

The general formulation of ANOVA models in terms of matrices generalized the ANOVA models to *linear mixed effects models*, of which ANOVA models forms a small subset. Thus ML and REML estimation initiated research in the area of linear mixed models and their applications. It made it possible to better analyze longitudinal data, where more complicated correlation structures over time are needed that ANOVA models cannot provide. Furthermore, it also made it possible to extend the linear mixed models to *non-linear mixed effects models* and *generalized linear mixed effects models*. In Section 2.10 we will briefly describe these models and explain why it forms a much larger class. For more details on the development of mixed models, we refer to Fitzmaurice et al. (2008).

2.3 Two-way nested mixed effects ANOVA model

To understand all aspects of ANOVA models in a brief and condensed way, it may be best to consider a single ANOVA model that is not too simple and not too complex either. It should be a model that contains most of the important features and relevant complexities, without making things too complicated. It should also be a model that can be extended easily to illustrate features that are not contained in the model. We decided to use the unbalanced mixed effects two-way nested ANOVA model. It contains fixed and random effects, it contains nested factors, it is unbalanced and it connects nicely with the case study described in Section 1.9.1. Unfortunately, the model does not contain crossed factors, but these factors can be added easily to be able to explain expected mean squares and interaction effects. In Section 2.9 we will discuss a few alternative ANOVA models.

We will study the verbal IQ variable as the numerical outcome, but we could have taken any of the other numerical variables on the children or alternatively have calculated a difference in the pre and post scores. Since ANOVA models typically assumes normality, it could be useful to study transformations of numerical outcomes to make the variables more normally distributed. For instance, the logarithm of the ratio of the post and pre test scores may be better that the difference in post and pre scores if the transformation would make the numerical outcome more normal.

It is important to recognize that the collection of the data of the case study was conducted in two steps or two stages. This sampling approach is called *cluster sampling* (Levi and Lemeshow, 2013; Kaptein and Van den Heuvel, 2021). In the first stage of cluster sampling a random sample of schools was collected and in the second stage a single random class within each sampled school was collected (when multiple classes would have been available). Then each child in the class was asked to participate.

Table 2.2: Numbers of children with an verbal IQ score

	Mean	Min	Max
Single	22.6	7	36
Multi	12.7	5	23
Overall	19.0	5	36

Note that this form of sampling may introduce correlation between the IQ scores of the children in one class. Children in one class may be more alike when compared to children from another school or class, similar to the one-way random effects ANOVA model. Thus this aspect should be addressed when we would

like to analyze the IQ data. Since the schools can be classified in two types of schools, those schools that educate children in multi-grade classes and those schools that educate children in single-grade classes. Thus schools can be seen as a nested factor within the factor multi-grade classes.

Finally, the data is unbalanced in two ways. The number of children may vary with class and the number of schools may vary with the levels of the multi-grade factor. The average number of children per class with a verbal IQ score and the minimum and maximum are provided in Table 2.2. There were 131 schools with single grade classes and 75 schools with multi-grade classes.

The variance component model for the IQ-score as dependent variable can now be formulated as follows:

$$y_{ijk} = \mu_i + a_{j(i)} + e_{ijk}, \tag{2.11}$$

with y_{ijk} the IQ-score of the k^{th} child ($k = 1, 2, ..., K_{ij}$) in the j^{th} class ($j = 1, 2, ..., J_i$) of the i^{th} class type ($i = 1, 2, ..., I$), μ_i the mean IQ-score of children in class type i, $a_{j(i)}$ a random effect for class j within class type i, and e_{ijk} a random effect for child k within class j for class type i. Thus there are two sets of random effects for model (2.11): one set for the variability between classes ($a_{j(i)}$) and one set for the variability between children (e_{ijk}). However, the random effects e_{ijk} are often not seen as random effects, since they often play a different role in the analysis. They typically represent *unexplained* variability in the outcome variable, while the other random effects typically try to *explain* the variability. The random effects e_{ijk} are referred to as the *error terms* or *residuals*. They also appear in a linear regression model, but the term $a_{j(i)}$ would not be part of a linear regression analysis. The fixed effect parameters in this model would be given by μ_1 and μ_2 (since $I = 2$) or equivalently given by $\mu = \mu_1$ and $\alpha = \mu_2 - \mu_1$ (or $\mu = \mu_2$ and $\alpha = \mu_1 - \mu_2$) in an alternative parametrization of the fixed effects. Frequently, only the difference $\mu_2 - \mu_1$ (or $\mu_1 - \mu_2$) in means between the two types of classes is referred to as the fixed effect of type of classes, but we will also refer to the intercept μ as a fixed effect parameter.

Since model (2.11) contains only two factors (i.e., class type and class), it is called a *two-way* ANOVA model. Thus, there also exists n-way ANOVA models, with n the number of categorical factors involved. Since class is nested within class type, model (2.11) is a two-way *nested* ANOVA model. The term nested is only added if all factors are nested within each other. Nested refers to the hierarchical structure of levels of factors. The first class of the eight grade classes is a different class than the first class of the multi-grade classes, although they may both be recorded as class one. Opposite to the nested factors are *crossed* factors, as discussed in Section 2.1. For instance, the factor sex, could enter model (2.11) and could be crossed with the factors class type and with the nested factor class within class type (see Exercise 4). Male and female in one class is and means the same as male and female in another class. If all factors would be crossed (e.g., class is not nested within class type) the model would be referred to as a three-way *crossed* ANOVA model. Note that these cross products in a model are referred to as *interaction terms*.

In mixed models it is common to assume that the random effects are normally distributed

$$a_{j(i)} \sim N(0, \sigma^2_{C(T)}) \text{ and } e_{ijk} \sim N(0, \sigma^2_R),$$

although distributional extensions also exist (e.g., Asar *et al.*, 2020). As we just mentioned, for variance component models the random effects and the residuals are all considered mutually independent, but for mixed effects models, random effects can be correlated with each other and the residuals can be correlated with each other as well, see Section 3. Dependency between random effects and residuals are typically uncommon in literature, although more research in this direction is currently appearing (e.g., Regis *et al.*, 2019; Geraci and Farcomeni, 2020).

The two-way nested ANOVA model can be viewed as a combination of two one-way random effects ANOVA models where the variance components have been taken equal in both one-way ANOVA models. For the schools with only single grade classes we would have model (2.2) with mean μ_1 and variance components $\sigma^2_G(S)$ and $\sigma^2_E(S)$ and for the schools with multi-grade classes we would have model (2.2) with mean μ_2 and variance components $\sigma^2_G(M)$ and $\sigma^2_E(M)$. Since we do not expect that variability between children would be really different for the two class types it may be reasonable to assume that $\sigma^2_E(S) = \sigma^2_E(M) = \sigma^2_R$. Furthermore, differences between classes may vary in the same way for the two types of classes: $\sigma^2_G(S) = \sigma^2_G(M) = \sigma^2_{C(T)}$. If we do not want to make these assumptions we could run two one-way ANOVA models on the data of schools with single classes and on the data of schools with multi-grade classes. However, if we wish to make just one these assumptions, the procedure MIXED of **SAS** could run the two-way nested ANOVA model with the extension that $\sigma^2_E(S)$ and $\sigma^2_E(M)$ or $\sigma^2_G(S)$ and $\sigma^2_G(M)$ are different.

2.3.1 Intraclass correlation coefficient

Even if all random effects are assumed independent, they can still induce or introduce certain dependencies in the outcomes. Indeed, model (2.11) leads to correlation of IQ-scores for children within classes, but there will be no correlation between children from different classes due to the assumption of independence of the random effects (and residuals). To calculate the correlation between IQ-scores within classes, the covariance is calculated first:

$$\mathsf{COV}(y_{ijk}, y_{ijl}) = \mathbb{E}[(y_{ijk} - \mu_i)(y_{ijl} - \mu_i)] = \mathbb{E}[a^2_{j(i)} + e_{ijk}e_{ijl}] = \sigma^2_{C(T)},$$

when $k \neq l$. Note that the expected value of y_{ijk} is equal to $\mathbb{E}[y_{ijk}] = \mu_i$, that $\mathbb{E}[e_{ijk}e_{ijl}] = 0$ due to the independence of the residuals and their zero means (i.e., $\mathbb{E}[e_{ijk}] = 0$), and that the second moment of $a_{j(i)}$ is equal to $\mathbb{E}[a^2_{j(i)}] = \mathsf{VAR}(a^2_{j(i)}) = \sigma^2_{C(T)}$, since the random effects have zero mean ($\mathbb{E}[a_{j(i)}] = 0$). Furthermore, the total variability in the IQ-score is $\mathsf{VAR}(y_{ijk}) = \sigma^2_{C(T)} + \sigma^2_R$, which makes the correlation between IQ-scores of children in the same class equal to

$$\mathsf{ICC} \equiv \mathsf{CORR}(y_{ijk}, y_{ijl}) = \frac{\sigma^2_{C(T)}}{\sigma^2_{C(T)} + \sigma^2_R}. \tag{2.12}$$

This correlation is called the *intraclass correlation coefficient* (ICC). It varies within the unit interval $\mathsf{ICC} \in [0, 1]$. The larger the difference between IQ-scores across classes, the more alike the IQ-scores of children within classes.

Finally, it is straightforward to see that the covariance between children from different classes is zero: $\mathsf{COV}(y_{ijk}, y_{rst}) = \mathbb{E}[a_{j(i)}a_{s(r)} + e_{ijk}e_{rst}] = 0$, when $i \neq r$ and/or $j \neq s$, due to the independence of random effects and residuals.

2.3.2 Estimation techniques

ANOVA models contain fixed effects parameters and variance components. For instance, model (2.11) contains two fixed effects parameters μ_1 and μ_2 and two variance components $\sigma^2_{C(T)}$ and σ^2_R. In the past these parameters were estimated by *generalized least squares* (GLS) for the fixed effects parameters and *mean squares* for the variance components. These estimation techniques were typically referred to as the *method of moments* or *ANOVA methods*. This estimation approach does not make distributional assumptions, besides the existence of means and variances. Normality is assumed only when confidence intervals are being constructed or when hypothesis testing on fixed or random effects are being applied. Moment-based estimation techniques have been further developed for non-normally distributed outcomes under the name of *generalized estimating equations* (GEE), see Section 5.

Alternative estimation approaches are *maximum likelihood* (ML) and *restricted maximum likelihood* (REML) estimation. They have been developed for variance component models many years ago, but software programs were not fully available to execute these estimation techniques before the 1990s (Littell, 2002). The likelihood-based estimation approaches do immediately implement the assumption of normality for all random elements. ML estimation can be applied to any mixed model, while REML is mostly developed for linear mixed models.

It is important to realize that the moment-based and likelihood-based estimation techniques both have their advantages and disadvantages and that there does not exist one estimation technique that is superior to the other techniques in all situations. We will discuss the method of moments for ANOVA models in Sections 2.4 and 2.5. The likelihood-based estimation techniques are discussed in Section 2.6. We will give advise on the use of moment-based and likelihood-based estimation for ANOVA models in Section 2.7.

2.4 Method of moments: Balanced data

Estimation of ANOVA and variance component models with the method of moments is easiest when the data is *balanced*. This means that the number of observations in each combination of levels of the categorical variables are the same. In model (2.11) this would mean that each class has the same number of children ($K_{ij} = K$, $\forall i, j$) and that each class type has the same number of classes ($J_i = J$, $\forall i$). For unbalanced data the method of moments is typically not unique, which is a serious disadvantage. Before we explain this issue of non-uniqueness, let us first assume that the data is balanced.

2.4.1 Sums of squares

The *total variability* in the outcome variable for the balanced two-way ANOVA model (2.11) is defined by

$$SS_{\text{Total}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{...})^2,$$

with $\bar{y}_{...}$ the overall average $\bar{y}_{...} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}[y_{ijk}/(IJK)]$. It calculates the total squared distance from each observation to the overall average. It is referred to as the *total sums of squares*. Since model (2.11) has two factors (class type T and class within class type $C(T)$) and a residual R, the total sums of squares can be uniquely partitioned into three sums of squares: SS_T, $SS_{C(T)}$, and SS_R.[7] There exist generic rules to help determine how the different sums of squares are calculated (Searle *et al.*, 2006, Chapter 4), i.e., knowing which squared distances to calculate for each factor. These rules make use of the factor structure of the data and the *degrees of freedom* of each sum of squares. The degrees of freedom indicate how many independent observations are used to quantify the variability in the outcome variable. To illustrate these generic rules we will explain them for model (2.11).

For the IQ-scores we have the following factor structure: class type T, class within class type $C(T)$, and the residual within class type and class $R(TC)$. Note that it is uncommon to write $R(C(T))$ and that in most cases it is well known that the residuals are completely nested within all other factors and therefore a notation of R is frequently applied. Each factor is associated with a degrees of freedom that can be determined in the following way. Factors outside the brackets (e.g., C for factor $C(T)$) count for their number of levels minus and factors within brackets (e.g., T for factor $C(T)$) count for their levels. Since factor T is associated with index i in model (2.11), factor T is replaced by $I-1$ (their number of levels minus one). Thus the degrees of freedom for class type T is $df_T = I - 1$. Factor class within class type is associated with index j. Therefore, $C(T)$ is replaced by $(J-1)I$, where $J-1$ replaces C (number of levels minus one) and I replaces T (the number of levels). Thus the number of degrees of freedom for class within class type is $df_{C(T)} = I(J-1)$. Finally, the residual is associated with index k, which implies that $R(TC)$ is replaced by $(K-1)IJ$. Thus the degrees of freedom for the residual is $df_R = IJ(K-1)$. If we add all degrees of freedom, we get the total degrees of freedom $df_{\text{Total}} = I-1+IJ-I+IJK-IJ = IJK-1$, which is equal to all observations minus one (for the overall mean). These degrees of freedom belongs to SS_{Total}.

The number of degrees of freedom for each factor now indicates which averages of the IQ-scores are being used in the sums of squares. For instance, if we have IJ in the degrees of freedom we use the average $\bar{y}_{ij.} = \sum_{k=1}^{K} y_{ijk}/K$ for class j in class type i. Thus the letters that are missing (here index k) are being averaged out. The degrees of freedom for $C(T)$ is $IJ-I$, so we need to study the squared difference between $(\bar{y}_{ij.} - \bar{y}_{i..})^2$, with $\bar{y}_{i..} = \sum_{j=1}^{J}[\bar{y}_{ij.}/J]$ the average value for class type i. For the value 1 we use the overall average $\bar{y}_{...}$ since no

[7] Note that we use the letters of the variables they represent: T for the factor type of class, C for the factor class, and R for the residual. Since class is nested within type of class we denote $C(T)$.

letters are involved and all indices must be averaged out. The sum of squares for a factor is then calculated by summing over all indices (even if the squared differences does not contain all indices) to calculate the total squared distance for this factor. Thus the three sums of squares for model (2.11) with balanced data are now defined by

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{i..} - \bar{y}_{...})^2 = \sum_{i=1}^{I} JK(\bar{y}_{i..} - \bar{y}_{...})^2,$$
$$SS_{C(T)} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{ij.} - \bar{y}_{i..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} K(\bar{y}_{ij.} - \bar{y}_{i..})^2,$$
$$SS_R = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{ijk.} - \bar{y}_{ij.})^2.$$

Note that the rule for establishing sums of squares also applies to the total sum of squares. The degrees of freedom is $IJK - 1$, thus we need to sum up all squared distances $(y_{ijk} - \bar{y}_{...})^2$. This leads to SS_{Total}.

2.4.2 Mean squares and expected mean squares

As we just mentioned, the sums of squares calculate total variabilities or squared distances, but to estimate the variance components of an ANOVA model, we need to normalize the sums of squares by their degrees of freedom. These normalized sums of squares are called *mean squares*. Thus the mean squares for model (2.11) are $MS_T = SS_T/[I-1]$, $MS_{C(T)} = SS_{C(T)}/[I(J-1)]$, and $MS_R = SS_R/[IJ(K-1)]$. It is uncommon to calculate a mean square for the total variability.

Since the mean squares are descriptive statistics, we can determine the *expected mean squares* if we assume that the data is generated according to model (2.11). These expectations will tell us what the mean squares are trying to estimate. Specific rules have been developed to quickly determine the expected mean squares for balanced ANOVA models. The first step is to assume that all factors provide random effects. Thus we assume for a moment that we have no fixed effects and only at the end of the calculations we will discuss their role in the expected mean squares. Thus for model (1) we now assume that the fixed effect μ_i for class type i is a random variable having a mean μ and a variance equal to σ_T^2.

For the second step we have to create a table with the factors in the rows and the variance components in the columns (see Table 2.3) and determine which cells are zero and which are not. A cell in the table is non-zero when the letters of the factors in the corresponding row are present in the letters of the variance component in the corresponding column. Otherwise the cell becomes zero. For example, the letter T for the factor class type in row T is present in each of the variance components, since R stands for $R(TC)$. Thus we need a positive number in each cell for this row. Why these numbers are JK, K, and 1, respectively, is discussed in a minute. The letters C and T for factor or row $C(T)$ are present in the variance components $\sigma^2_{C(T)}$ and σ^2_R but they are not <u>both</u> present in σ^2_T. Thus the cell for row $C(T)$ and column σ^2_T becomes zero, but the other cells in this row will be non-zero. Since the letters $R(TC)$ for the residual are all present in the residual variance component only, all cells in this row will be equal to zero, except for the last column.

Then the third step is to determine the numbers for the non-zero cells. The number to put in the cell are the number of levels for the factors that are missing in the variance component. For the residual variance all indices are present, thus the number becomes always equal to one. For the variance component $\sigma^2_{C(T)}$ we miss index k, since the letter R is missing. Thus the number in the non-zero cells of column $\sigma^2_{C(T)}$ becomes the number of levels for this factor, which is K. For variance component σ^2_T the indices j and k are missing since the factors C and R are missing. Then the product of all the levels of these factors are put in the non-zero cells, leading to JK. Table 2.3 now provides the following expected mean squares for a balanced model (2.11) with all factors being random:

Table 2.3: Expected mean squares for the balanced two-way nested ANOVA model.

Factor	σ^2_T	$\sigma^2_{C(T)}$	σ^2_R
T	JK	K	1
$C(T)$	0	K	1
$R(TC)$	0	0	1

$$\begin{aligned} \mathbb{E}(MS_T) &= JK\sigma^2_T + K\sigma^2_{C(T)} + \sigma^2_R \\ \mathbb{E}(MS_{C(T)}) &= K\sigma^2_{C(T)} + \sigma^2_R \\ \mathbb{E}(MS_R) &= \sigma^2_R \end{aligned} \tag{2.13}$$

The fourth and final step is to bring back the fixed effects. When there are also fixed effects in the model, the assumed variance components do not exist. So, this means that we can not include them in the expected mean squares. Thus, if the factor $C(T)$ would be modeled as fixed effects, σ^2_{CT} does not exist and can therefore not play a part in the expected mean squares for factor T. The expected mean square for factor T would then become $\mathbb{E}(MS_T) = JK\sigma^2_T + \sigma^2_R$. All mean squares are cleaned up in this way. For model (2.11) the only fixed effects are from factor T. Thus the variance component σ^2_T does not exists, but this variance component is never mentioned in any of the expected mean squares, except in the mean square for the fixed effects factor T. So, the expected means squares $\mathbb{E}(MS_{C(T)})$ and $\mathbb{E}(MS_R)$ in (2.13) remain unchanged, since they do not contain σ^2_T. According to this fixed effects rule, $\mathbb{E}(MS_T)$ should become $\mathbb{E}(MS_T) = K\sigma^2_{C(T)} + \sigma^2_R$, since σ^2_T does not exist, but the fixed effects μ_1 and μ_2 for factor T do need to end up in this expected mean squares somewhere. Indeed, they do and they replace the term $JK\sigma^2_T$ by a quadratic function of the fixed effects, i.e., by $Q(\mu_1, \mu_2)$. The expected mean square $\mathbb{E}(MS_T)$ therefore becomes $\mathbb{E}(MS_T) = Q(\mu_1, \mu_2) + K\sigma^2_{C(T)} + \sigma^2_R$. The term $Q(\mu_1, \mu_2)$ for model (2.11) can be explicitly determined and is equal to $Q(\mu_1, \mu_2) = 0.5JK(\mu_1 - \mu_2)^2$. If class within class type $C(T)$ would have been a fixed effects factor, the expected mean squares for factor $C(T)$ would become $\mathbb{E}(MS_{C(T)}) = Q(a_{1(1)}, a_{2(1)}, ..., a_{J(I)}) + \sigma^2_R$ and $\mathbb{E}(MS_T)$ would become $\mathbb{E}(MS_T) = Q(\mu_1, \mu_2) + \sigma^2_R$.

2.4.3 Hypothesis testing for fixed and random effects

Under the assumption of a normally distributed outcome variable, there are two important distributional characteristics for the mean squares in balanced ANOVA models (Searle *et al.*, 2006). When data is unbalanced, these characteristics would not hold anymore. The mean squares are all

1. Independently distributed.

2. χ^2-distributed when properly scaled.

The mean squares for factors that are modeled as random effects have a central χ^2-distribution, while the mean squares for fixed effects factors have a non-central χ^2-distribution. If H represents a random effects factor, we have that $df_H MS_H/\mathbb{E}(MS_H)$ is χ^2-distributed with df_H degrees of freedom. Thus for the balanced ANOVA model (2.4) and a normally distributed IQ-score, we have that $df_{C(T)}MS_{C(T)}/[K\sigma^2_{C(T)} + \sigma^2_R]$ is $\chi^2_{df_{C(T)}}$-distributed and $df_R MS_R/\sigma^2_R$ is $\chi^2_{df_R}$-distributed. If H represents a fixed effects factor and $\boldsymbol{\xi}_H$ is the vector of fixed effects parameters, then $df_H MS_H/[\mathbb{E}(MS_H) - Q(\boldsymbol{\xi}_H)]$ has a non-central χ^2-distribution with df_H degrees of freedom and non-centrality parameter $df_H Q(\boldsymbol{\xi}_H)/\mathbb{E}(MS_H) - Q(\boldsymbol{\xi}_H)]$. Thus $df_T MS_T/[K\sigma^2_{C(T)} + \sigma^2_R]$ has a non-central $\chi^2_{df_T}$-distribution with non-centrality parameter given by $0.5JK(\mu_1 - \mu_2)^2/[K\sigma^2_{C(T)} + \sigma^2_R]$. Note that for the IQ-scores in model (2.11), the number of levels for class type is $I = 2$.

Now that we have determined the distribution of the mean squares, we could test whether factors contribute to the variability in the outcome variable. For model (2.11) there are two hypotheses that can be tested: the contribution of factor class type T and of factor class within class type $C(T)$. The number of hypotheses depends on how many factors or terms are in the model, excluding the residual. The null hypotheses for class type and class within class type are defined by $H_0 : \mu_1 = \mu_2$ and $H_0 : \sigma^2_{C(T)} = 0$, respectively. The test statistics are derived from the mean squares. For instance, if class type does not contribute to the variability in the IQ-score, the expected mean square of MS_T is equal to the expected mean square of $MS_{C(T)}$. Indeed, under $H_0 : \mu_1 = \mu_2$, the quadratic term $Q(\mu_1, \mu_2) = 0.5JK(\mu_1 - \mu_2)^2$ becomes zero and the distribution of the statistic $df_T MS_T/[K\sigma^2_{C(T)} + \sigma^2_R]$ becomes a central χ^2-distribution with df_T degrees of freedom. Combining this result with the independence of mean squares, the test statistic $F_T = MS_T/MS_{C(T)}$ has an F-distribution with df_T and $df_{C(T)}$ degrees of freedom, respectively, when the null hypothesis $H_0 : \mu_1 = \mu_2$ is true and the outcome variable is normally distributed. Thus, the test statistic F_T can now be compared with the critical value of the corresponding F-distribution. If the F_T is larger than this critical value, the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected. Something similar occurs for $MS_{C(T)}$. Under the null hypothesis $H_0 : \sigma^2_{C(T)} = 0$, the expected mean square of $MS_{C(T)}$ is equal to the expected mean square of MS_R. Thus under the null hypothesis $H_0 : \sigma^2_{C(T)} = 0$, the statistic $F_{C(T)} = MS_{C(T)}/MS_R$ has an F-distribution with $df_{C(T)}$ and df_R degrees of freedom, respectively. Thus when $F_{C(T)}$ is larger than the 95% quantile of the F-distribution with $df_{C(T)}$ and df_R degrees of freedom, the null hypothesis $H_0 : \sigma^2_{C(T)} = 0$ is rejected.

2.4.4 Satterthwaite approach

In model (2.11), the expected mean squares for the different factors reduces to an expected mean square of another factor, when the factor of interest does not contribute to the variability in the outcome. Indeed, $\mathbb{E}(MS_T)$ reduces to $\mathbb{E}(MS_{C(T)})$ and $\mathbb{E}(MS_{C(T)})$ reduces to $\mathbb{E}(MS_R)$ under their respective null hypotheses. Unfortunately, this does not occur in all ANOVA models (see Exercise 5). If another random effects factor H would enter model (2.11), which would

also interact with class type (and possibly with class within class type), we obtain an expected mean square $\mathbb{E}(MS_H)$ for factor H that does not reduce to any of the other expected mean squares in this extended model. Thus an exact F-test can not be constructed for testing the contribution of factor H to the variability of the outcome variable.

The approach that is typically followed for such factors H is to construct an approximate F-test, where the mean square MS_H is compared to a linear combination of other mean squares in the model (see Searle *et al.* (2006) and also Section 2.5). The expectation of this linear combination is the expected mean square of MS_H when factor H does not contribute to the variability of the outcome variable. The linear combination of mean squares is then approximated with a χ^2-distribution (Satterthwaite, 1946), where the degrees of freedom are being estimated from the data.

This Satterthwaite approach can be formulated as follows. Let MS_1, MS_2, ..., and MS_n be a set of independent mean squares (for random effects factors) with degrees of freedom given by d_1, d_2,....., and d_n, respectively. This implies that $d_r MS_r/\mathbb{E}(MS_r)$ is a χ^2-distributed random variable. Now assume that we are interested in the linear combination $MS_L = \sum_{r=1}^{n} \eta_r MS_r$, with η_1, η_2, ..., and η_n known constants. Then the mean and variance of MS_L is given by

$$\begin{array}{rcl} \mathbb{E}(MS_L) & = & \sum_{r=1}^{n} \eta_r \mathbb{E}(MS_r), \\ \mathsf{VAR}(MS_L) & = & \sum_{r=1}^{n} 2\eta_r^2 [\mathbb{E}(MS_r)]^2/d_r, \end{array}$$

since the mean and variance of a χ^2-distribution with df degrees of freedom are equal to df and $2df$, respectively. If we now assume that the normalized linear combination of mean squares $df_L MS_L/\mathbb{E}(MS_L)$, is (approximately) a χ^2-distributed random variable with df_L degrees of freedom, then the variance $\mathsf{VAR}(df_L MS_L/\mathbb{E}(MS_L))$ would be equal to $2df_L$, but it would also be equal to $df_L^2 \mathsf{VAR}(MS_L)/[\mathbb{E}(MS_L)]^2$. Solving the degrees of freedom for this equality leads to

$$df_L = \frac{\left[\sum_{r=1}^{n} \eta_k \mathbb{E}(MS_r)\right]^2}{\sum_{r=1}^{n} d_r^{-1} \eta_r^2 [\mathbb{E}(MS_r)]^2}. \tag{2.14}$$

Unfortunately, this solution contains the expected mean squares $\mathbb{E}(MS_L)$, which would typically be unknown in practice. But when we would replace $\mathbb{E}(MS_L)$ by MS_L in (2.14), we obtain Satterthwaite degrees of freedom.

2.4.5 Variance component estimators

Solving the equations in (2.13) for the variance components, lead to variance components that are expressed in linear combinations of the expected mean squares. Substituting the mean squares for the expected mean squares, results in the moment estimators for the variance components. The estimator for σ_R^2 is just the residual mean square: $\hat{\sigma}_R^2 = MS_R$, while the estimator for $\sigma_{C(T)}^2$ is equal to $\hat{\sigma}_{C(T)}^2 = [MS_{C(T)} - MS_R]/K$. The factor T is modeled as fixed effects factor, thus σ_T^2 does not exist and is therefore not estimated in this way, but if factor T was a random effects factor, σ_T^2 would have been estimated by $\hat{\sigma}_T^2 = [MS_T - MS_{C(T)}]/JK$.

Since the variance components estimators are linear combinations of the mean squares, it is straightforward to calculate its variance when the outcome variable

is normally distributed. Under the normality assumption we can make use of the moments of the χ^2-distributions and their independence. Thus the variance of variance component estimators $\hat{\sigma}^2_R$ and $\hat{\sigma}^2_{C(T)}$ are now given by

$$\begin{aligned} \mathsf{VAR}(\hat{\sigma}^2_R) &= 2df_R^{-1}\sigma^4_R, \\ \mathsf{VAR}(\hat{\sigma}^2_{C(T)}) &= 2K^{-2}\left(df_{C(T)}^{-1}[\sigma^2_R + K\sigma^2_{C(T)}]^2 + df_R^{-1}\sigma^4_R\right). \end{aligned} \tag{2.15}$$

These variances can be estimated by substituting the variance components estimators for the variance components. The estimator for the variance of $\hat{\sigma}^2_R$ is then equal to $V(\hat{\sigma}^2_R) = 2MS^2_R/df_R$ and the estimator for the variance of $\hat{\sigma}^2_{C(T)}$ is then given by $V(\hat{\sigma}^2_{C(T)}) = 2K^{-2}[MS^2_{C(T)}/df_{C(T)} + MS^2_R/df_R]$. By taking the square root, we obtain standard errors of the variance component estimators.

This principle of constructing and estimating variances of variance component estimators can be applied to any (balanced) ANOVA model, but it is important to realize that these variance estimators are not unbiased estimators of the true variances. Indeed, the expectations of the estimator $V(\hat{\sigma}^2_R)$ and $V(\hat{\sigma}^2_{C(T)})$ are given by

$$\begin{aligned} \mathbb{E}[V(\hat{\sigma}^2_R)] &= \tfrac{2\sigma^4_R}{df_R}\left[1 + \tfrac{2}{df^2_R}\right], \\ \mathbb{E}[V(\hat{\sigma}^2_{C(T)})] &= \tfrac{2}{K^2}\left(\tfrac{[\sigma^2_R + K\sigma^2_{C(T)}]^2}{df_{C(T)}}\left[1 + \tfrac{2}{df^2_{C(T)}}\right] + \tfrac{\sigma^4_R}{df_R}\left[1 + \tfrac{2}{df^2_R}\right]\right). \end{aligned}$$

Thus the estimators $V(\hat{\sigma}^2_R)$ and $V(\hat{\sigma}^2_{C(T)})$ overestimate the variances in (2.15) of the variance component estimators.

A disadvantage of the moment estimators for estimation of variance components, is the possibility of obtaining a negative estimate. If we consider estimator $\hat{\sigma}^2_{C(T)}$, the residual mean square MS_R can be larger then the mean square $MS_{C(T)}$ for factor $C(T)$. Of course, this could happen when $\sigma^2_{C(T)} = 0$, but it may also occur for other reasons when $\sigma^2_{C(T)} > 0$. If the number of classes within class type is small, say less than or equal to 5, it is more likely to have a negative estimate then when the degrees of freedom $df_{C(T)}$ would be large. Alternatively, if the variance component $\sigma^2_{C(T)}$ is much smaller than the residual variance σ^2_R, a large probability for a negative estimate may still occur even if we have a larger sample size. The probability of having a negative estimator $\hat{\sigma}^2_{C(T)}$ is given by

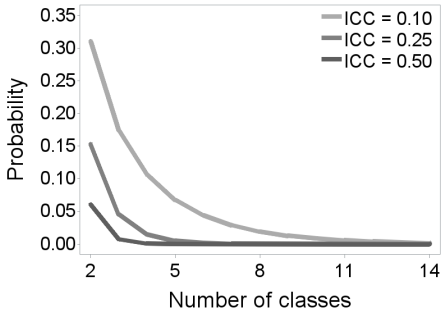


Figure 2.1: Probability of a negative estimator for $\sigma^2_{C(T)}$ as function of J and for three different ICC values (and with $K = 15$ children per class).

$$\mathbb{P}(\hat{\sigma}^2_{C(T)} \leq 0) = \mathbb{P}(MS_{C(T)} \leq MS_R) = F_{df_{C(T)},df_R}(\sigma^2_R/[K\sigma^2_{C(T)} + \sigma^2_R]), \tag{2.16}$$

with F_{d_1,d_2} the F-distribution function with d_1 and d_2 degrees of freedom, respectively. Note that we have made use of the central χ^2-distribution of the

mean squares in balanced ANOVA models for factors that are modeled as random effects if the outcome is normally distributed and the fact that the ratio of independent χ^2-distributed random variables are related to an F-distribution. The probability in (2.16) shows that it is less likely to have a negative estimate when the degrees of freedom $df_{C(T)}$ is large and $K\sigma^2_{C(T)}$ is large with respect to σ^2_R (see also Figure 2.1 for several settings).

In most practical situations, it is recommended to truncate negative variance component estimates to zero, since variance components can never be negative. Theoretically, this would make the estimators biased due to the positive probability in (2.16) of obtaining a negative estimate. Indeed, the truncated estimator overestimates the variance component:

$$\mathbb{E}[\max\{0, \hat{\sigma}^2_{C(T)}\}] = \mathbb{E}[\hat{\sigma}^2_{C(T)} 1_{(MS_R,\infty)}(MS_{C(T)})] > \mathbb{E}[\hat{\sigma}^2_{C(T)}] = \sigma^2_{C(T)}.$$

However, this issue of bias does not outweigh a nonsensical negative estimate.

2.4.6 Fixed effects estimators

Now that we have determined the variance component estimators for balanced ANOVA models, we need to determine the estimators for the fixed effects. One easy way of doing this, is to apply the *ordinary least squares* (OLS) method. This method makes use of the squared distance of the observations with respect to their expectations. For model (2.11), this means calculating the squared distances

$$\mathsf{OLS} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \mu_i)^2.$$

The solutions for the fixed effects that minimizes this squared distance OLS are the ordinary least squares estimators. For μ_i this would lead to $\hat{\mu}_i = \bar{y}_{i..}$. Indeed, taking the derivative of OLS with respect to μ_i and then equating this derivative to zero, leads to $\hat{\mu}_i = \bar{y}_{i..}$.

The OLS estimators for the fixed effects in ANOVA models are typically represented by certain averages. However, in case we would have selected an alternative parametrization, having one mean μ and one parameter that expresses the difference α_i between the two class types (with either $\alpha_1 = 0$ or otherwise $\alpha_2 = 0$), we would have obtained the mean $\hat{\mu} = \bar{y}_{2..}$ and $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{2..}$ if $\alpha_2 = 0$. This solution is often implemented by software packages, like **SAS**, to be able to deal with both balanced and unbalanced data. For balanced data, the ideal parametrization would be one mean parameter μ that represents the overall mean and an effect α_i for class type, but now with the restriction $\alpha_1 + \alpha_2 = 0$, to make sure that we have as many parameters as we have levels for the fixed effects. This symmetry would make sense in balanced data, since we have as many observations for one class type as we have for the other class type, but much less in unbalanced data. The imbalance in data would then request some strange weighted average of the α's to be equal to zero. For this ideal parametrization, the OLS estimators would now become $\hat{\mu} = \bar{y}_{...}$, $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{...}$, and $\hat{\alpha}_2 = \bar{y}_{2..} - \bar{y}_{...}$. Note that $\hat{\alpha}_1 + \hat{\alpha}_2 = 0$ is satisfied in this solution for balanced data. This parametrization fits best with the balanced ANOVA, because it now contains the estimate $\bar{y}_{...}$ that is also used in the total sums of squares.

It should be noted that the OLS approach does not take into account the data and correlation structure that is due to the random effects. In other words, the OLS approach, that was invented for linear regression analysis with independent and identically distributed residuals, would be the most appropriate approach only when the model is given by $y_{ijk} = \mu_i + \tilde{e}_{ijk}$, with $\tilde{e}_{ijk}$ i.i.d. normally distributed residuals $\tilde{e}_{ijk} \sim N(0, \sigma^2)$. Clearly, model (2.11) can be rewritten as $y_{ijk} = \mu_i + \tilde{e}_{ijk}$, but now the residuals $\tilde{e}_{ijk} = a_{j(i)} + e_{ijk}$ are not independent nor identically distributed anymore. This questions whether we should somehow use the data correlation structure in the calculation of the squared distances. An approach that accomplishes this is called the *generalized least squares* (GLS) method, which will be discussed when we discuss estimation for unbalanced data. The GLS estimation method is programmed in procedure MIXED of **SAS**. For balanced ANOVA models, the GLS will be identical to the OLS (Searle *et al.*, 2006), but when data is unbalanced differences in estimation may occur.

Even though the OLS estimators are constructed without using the data correlation structure, when the distribution of the OLS estimator is investigated, the random effects are being implemented. The variance of the OLS estimator $\hat{\mu}_i = \bar{y}_{i..}$ is given by

$$\mathbb{E}(\hat{\mu}_i - \mu_i)^2 = \mathbb{E}(\bar{a}_{.(i)} + \bar{e}_{i..})^2 = [K\sigma^2_{C(T)} + \sigma^2_R]/[JK],$$

with $\bar{a}_{.(i)} = \sum_{j=1}^{J} a_{j(i)}/J$ and $\bar{e}_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} e_{ijk}/(JK)$ the average class effect within class type i and the average residuals within class type i, respectively. Since the term $K\sigma^2_{C(T)}+\sigma^2_R$ is equal to the expected mean square of $MS_{C(T)}$, the variance of the OLS estimator $\hat{\mu}_i$ can be estimated by $MS_{C(T)}/[JK]$. The corresponding degrees of freedom is therefore equal to $df_{C(T)}$. In case we would have chosen the μ_2 and $\alpha = \mu_1 - \mu_2$, the OLS estimator for α is equal to $\hat{\alpha} = \bar{y}_{i..} - \bar{y}_{2..}$. This estimator has a mean α and a variance $2[K\sigma^2_{C(T)} + \sigma^2_R]/[JK]$.

2.4.7 Predictions of random effects

Although the effects of class within type of class are considered random, we may still want to calculate some average of the IQ-score of the children in each class. For instance, we may want to know which class has the highest and which class has the lowest average IQ-score. It may be tempting to just use the OLS estimator $\bar{y}_{ij.}$, but that would imply that we have treated $\mu_i + a_{j(i)}$ as a fixed effects parameter, while we chose it to be random. Since we do not view it as a parameter, we should in principle not talk about estimation anymore, but rather talk about *prediction* (although estimation is still often used in this context). The classes are seen as a random sample of many classes, and thus act as a random variable. This means that we are trying to determine the *realization* of the *unobserved* random variable $\mu_i + a_{j(i)}$ and this will be typically different from $\bar{y}_{ij.}$.

To obtain a prediction of a random variable, the best possible way forward is to consider the expected value of the random variable, conditionally on the information that is available for this random variable. Thus instead of using $\bar{y}_{ij.}$, we may want to use the conditional mean $\mathbb{E}(\mu_i + a_{j(i)}|\bar{y}_{ij.})$ as prediction. To determine an expression for this predictor, we will investigate the joint distribution of $\mu_i + a_{j(i)}$ and $\bar{y}_{ij.}$. Based on the normality assumptions in model (2.11), we

obtain the bivariate normal distribution

$$\begin{pmatrix} \mu_i + a_{j(i)} \\ \bar{y}_{ij.} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2_{C(T)} & \sigma^2_{C(T)} \\ \sigma^2_{C(T)} & \sigma^2_{C(T)} + \sigma^2_R/K \end{pmatrix} \right).$$

Making use of the rules for conditional distributional[8], we can determine the conditional mean $\mathbb{E}(\mu_i + a_{j(i)}|\bar{y}_{ij.})$ by

$$\mathbb{E}(\mu_i + a_{j(i)}|\bar{y}_{ij.}) = \mu_i + \sigma^2_{C(T)}(\bar{y}_{ij.} - \mu_i)/[\sigma^2_{C(T)} + \sigma^2_R/K_{ij}]. \tag{2.17}$$

Unfortunately, this expression depends on the model parameters, but they can be substituted by its estimators. Thus the average prediction for the IQ-score in class j of class type i is determined by

$$\hat{y}_{ij.} = \hat{\mu}_i + \hat{\sigma}^2_{C(T)}(\bar{y}_{ij.} - \hat{\mu}_i)/[\hat{\sigma}^2_{C(T)} + \hat{\sigma}^2_R/K],$$

This prediction is often referred to as the *Empirical Bayes* estimation or as the *best linear unbiased prediction* (BLUPs). This latter term is used, because the expression on the right hand side in (2.17) is a linear combination of the observations and its expectation is the expectation of $\mu_i + a_{j(i)}$, i.e., $\mathbb{E}[\mathbb{E}(\mu_i + a_{j(i)}|\bar{y}_{ij.})] = \mathbb{E}(\mu_i + a_{j(i)})$. Note that the prediction $\hat{y}_{ij.}$ can be rewritten into $\hat{y}_{ij.} = \bar{y}_{i..} + [MS_B - MS_W](\bar{y}_{ij.} - \bar{y}_{i..})/MS_B$ for balanced data.

From the expression in (2.17), it is obvious that the prediction $\hat{y}_{ij.}$ will only be close to the average $\bar{y}_{ij.}$ when σ^2_R/K is close to zero. Thus, when the residual variance is small or when the number of children in a class is large. Furthermore, this procedure typically shrinks the averages $\bar{y}_{ij.}$ towards the fixed effect estimator $\hat{\mu}_i$. Large averages ($\bar{y}_{ij.} > \hat{\mu}_i$) are predicted smaller with the prediction $\hat{y}_{ij.}$ and thus closer to $\hat{\mu}_i$ (i.e., $\hat{\mu}_i < \hat{y}_{ij.} < \bar{y}_{ij.}$), while low averages ($\bar{y}_{ij.} < \hat{\mu}_i$) are predicted larger with the prediction $\hat{y}_{ij.}$ and thus closer to $\hat{\mu}_i$ (i.e., $\bar{y}_{ij.} < \hat{y}_{ij.} < \hat{\mu}_i$).

In balanced designs, the ranking of average IQ-scores based on $\bar{y}_{ij.}$ will not be changed when we would use $\hat{y}_{ij.}$, the predictions will just be closer to $\hat{\mu}_i$. However, in unbalanced designs, the ranking can change due to the different sample sizes K_{ij}. The conditional mean in (2.17) for unbalanced designs is still equal to (2.17), but with K replaced by K_{ij}. Thus a small class with a large average IQ-score may shrink strongly since K_{ij} is just small, while a large average IQ-score in a large class will not shrink as much. Thus if the highest average IQ-score is observed in a small class, it will be likely that the prediction is not the largests among all predictions.

2.4.8 Confidence intervals

Under assumption of normally distributed random effects, the OLS estimators have a normal distribution and they are independent of the mean squares of the model (Searle *et al.*, 2006). These two characteristics, together with the χ^2-distribution of the mean squares, can be used to create an appropriate confidence

[8] If $(Z_1, Z_2)^T$ is bivariate normally distributed with mean $\mathbb{E}Z_i = \mu_i$, variance $\mathsf{VAR}(Z_i) = \sigma^2_i$, and covariance $\mathsf{COV}(Z_1, Z_2) = \sigma_{12}$, the conditional mean of Z_1 given $Z_2 = z$ is given by $E(Z_1|Z_2 = z) = \mu_1 + \sigma_{11}(z - \mu_2)/\sigma^2_2$.

interval for the fixed effects parameters. The confidence interval is of the well-known form for normally distributed data. To illustrate this, we will consider the OLS estimator $\bar{y}_{i..}$. The $100\%(1-\alpha)$ confidence interval is given by

$$\bar{y}_{i..} \pm t^{-1}_{df_{C(T)}}(1-\alpha/2)[MS_{C(T)}/(JK)]^{1/2}, \tag{2.18}$$

with $t_d^{-1}(q)$ the q^{th} quantile of the t-distribution with d degrees of freedom. The form in (2.18) is used for all fixed effects estimators, where the standard error is estimated with a (linear combinations of) mean square(s). In case a linear combination of mean squares must be used for the standard error, the appropriate degrees of freedom in (2.18) is calculated with Satterthwaite approach in (2.14). Note that the confidence interval in (2.18), with the Satterthwaite approach, is implemented in procedure MIXED of **SAS**.

The form of confidence interval in (2.18) is used for variance component estimators in procedure MIXED of **SAS**, see **SAS** code box 2.5.2, except for the residual variance component.[9] The fixed effect estimator in (2.18) is then replaced by the variance component estimator, the degrees of freedom is then set at ∞, and the standard error of the variance component estimator is calculated as discussed earlier. This approach is convenient, since it has been implemented in software procedures, but it has been demonstrated that this approach for variance component estimators only work when the number of levels for this factor is large. For the factor class within class type, this approach may be reasonable, but it is often better to use a logarithmic transformation on the estimator. The standard error of the estimator in the log scale can be obtained with the delta method described in Section 1.7. Then in the log scale we may use the form in (2.18) and then invert this confidence interval back to the original scale. For the residual variance component estimator, the χ^2-distribution is used, i.e., $[df_R MS_R/\chi^{-2}_{df_R}(1-\alpha/2);\ df_R MS_R/\chi^{-2}_{df_R}(\alpha/2)]$, with $\chi_d^{-2}(q)$ the qth quantile of the χ^2-distribution with d degrees of freedom (see **SAS** code box 2.5.2).

In case a variance component is truncated to zero, there can be no standard error calculated. Thus in the original and logarithmic transformed scale, confidence interval (2.18) will not work anymore for these variance component estimators. Then the probability in (2.16) may be used to determine an upper confidence limit. Here we will search for the maximum value of the variance component that leads to a probability of α in (2.16), with α the significance level (e.g., $\alpha = 0.05$). This upper confidence limit represents the maximum value of the variance component for which it is unlikely to get a negative estimate. For the variance component $\sigma^2_{C(T)}$ for class within class type, the upper confidence limit would be equal to

$$UCL = K^{-1}[1 - F^{-1}_{df_{C(T)},df_R}(\alpha)]\sigma^2_R,$$

if the estimate is $\hat{\sigma}^2_{C(T)} = 0$. The value $F^{-1}_{d_1,d_2}(q)$ represents the q^{th} quantile of the F-distribution with d_1 and d_2 degrees of freedom. To be able to use this

[9] The confidence interval in (2.18) is only used in the procedure MIXED of **SAS** when the moment-based estimators are used, since the variance component estimators can become negative. When the likelihood-based estimators are used, the variance component estimator can not be less than zero and the procedure MIXED of **SAS** uses a confidence interval that is based on the chi-square distribution: $[df\hat{\sigma}^2/\chi^{-2}_{df}(1-\alpha/2);\ df\hat{\sigma}^2/\chi^{-2}_{df}(\alpha/2)]$, with $\hat{\sigma}^2$ the variance component estimator. The degrees of freedom are based on Satterthwaite approach with $df = 2[\hat{\sigma}^2/SE(\hat{\sigma}^2)]^2$.

upper confidence limit in practice, we could estimate the variance component σ^2_R with it variance component estimator MS_R.

Intraclass correlation coefficient: In some cases we would like to construct confidence intervals on the intraclass correlation coefficient. For the balanced two-way nested design an exact confidence interval exists on the intraclass correlation coefficient in (2.12). This interval is equal to (2.6). To show that (2.6) is the exact confidence interval, notice that the ratio $\sigma^2_R MS_{C(T)}/[(n_0\sigma^2_{C(T)} + \sigma^2_R)MS_R]$ follows an F-distribution with $m-1$ and $m(n_0-1)$ degrees of freedom[10]. The estimator for the ICC in (2.12) is typically obtained by substituting the variance component estimators, leading to

$$\hat{\mathsf{ICC}} = \frac{\hat{\sigma}^2_{C(T)}}{\hat{\sigma}^2_{C(T)} + \hat{\sigma}^2_R} = \frac{MS_{C(T)} - MS_R}{MS_{C(T)} + (n_0-1)MS_R} = \frac{F-1}{F+n_0-1}, \tag{2.19}$$

with $F = MS_{C(T)}/MS_R$ the ratio of the two mean squares for the random terms in the ANOVA model (2.11). Furthermore, notice that the ratio $\sigma^2_R/[n_0\sigma^2_{C(T)} + \sigma^2_R]$, which is the multiplication factor that makes F an F-distributed random variable, can be written as $[\mathsf{ICC}^{-1} - 1]/[\mathsf{ICC}^{-1} + n_0 - 1]$. This has the same form as the estimator $\hat{\mathsf{ICC}}$ in (2.19). Thus a natural upper and lower bound on the ICC is to divide the statistic F in (2.19) with the $\alpha/2$th lower and upper quantile of the F-distribution as we did in (2.6). It then follows that

$$P\left(\mathsf{ICC} \leq \frac{F/F_L - 1}{F/F_L + n_0 - 1}\right) = P\left(\frac{\mathsf{ICC}^{-1} - 1}{\mathsf{ICC}^{-1} + n_0 - 1} F \geq F_L\right) = 1 - \frac{\alpha}{2}$$

and $P\left(\mathsf{ICC} \geq [F/F_U - 1]/[F/F_U + n_0 - 1]\right) = \alpha/2$. Thus the confidence interval in (2.6) has a coverage probability exactly equal to $1-\alpha$.

For unbalanced data an exact confidence interval does not exist anymore, since the distribution of $MS_{C(T)}$, when properly scaled, is not exactly chi-square anymore. Many different approaches have been developed (Donner, 1986) all having their advantages and disadvantages. In Section 2.9.4we will present a very generic approach that can be applied to any balanced or unbalanced ANOVA model (Demetrashvili et al., 2016).

Sums of variance components: The total variability in the IQ-scores in children is defined by $\sigma^2_{\text{Total}} = \sigma^2_{C(T)} + \sigma^2_R$. It represents the population variance in IQ-scores among all children in seven grade classes in the Netherlands. Reporting this total variability with a 95% confidence interval could be an important aspect of the study. Here we will make use of Satterthwaite approach to construct confidence interval (Donner 1986; Nijhuis and Van den Heuvel, 2007; Van den Heuvel, 2010).

The estimator for the total variability is equal to $\hat{\sigma}^2_{\text{Total}} = \hat{\sigma}^2_{C(T)} + \hat{\sigma}^2_R$. For balanced data this can be rewritten into $\hat{\sigma}^2_{\text{Total}} = [MS_{C(T)} + (n_0-1)MS_R]/n_0$. The distribution of this estimator is unknown, but it can be approximated by a chi-square distribution, where the degrees of freedom can be obtained with Satterthwaite approach (2.14). The degrees of freedom for estimator $\hat{\sigma}^2_{\text{Total}}$ is

[10] If X_1 and X_2 are independent chi-square distributed random variables having d_1 and d_2 degrees of freedom, then the ratio $[X_1/d_1]/[X_2/d_2]$ follows an F-distribution with d_1 and d_2 degrees of freedom.

then determined by

$$df_{\text{Total}} = \frac{[MS_{C(T)} + (n_0 - 1)MS_R]^2}{MS_{C(T)}^2/(m-1) + (n_0-1)MS_R^2/m]}.$$

A $100\%(1-\alpha)$ confidence interval is then obtained with the chi-square distribution, as we discussed for the residual variance σ_R^2. Thus the $100\%(1-\alpha)$ confidence interval on σ_{Total}^2 is now given by

$$[\hat{\sigma}_{\text{Total}}^2/\chi_{df_{\text{Total}}}^{-2}(1-\alpha/2), \hat{\sigma}_{\text{Total}}^2/\chi_{df_{\text{Total}}}^{-2}(\alpha/2)]$$

with $\chi_d^{-2}(q)$ the qth quantile of the chi-square distribution with d degrees of freedom.

The confidence interval is not an exact confidence interval, but an approximate confidence interval. This means that the probability that σ_{Total}^2 is inside the confidence interval is only approximately equal to $1-\alpha$. Many studies have investigated this procedure and it gives good coverage results in many situations (Nijhuis and Van den Heuvel, 2007). The approach can be extended to any sums of variance components from unbalanced data (Van den Heuvel, 2010) as we will demonstrate in Section 2.9.3.

2.5 Method of moments: Unbalanced data

To understand the GLS estimation technique for the fixed effects of unbalanced ANOVA models, it is easier to change to a matrix notation. As we already have seen earlier, model (2.11) was rewritten into the form of $y_{ijk} = \mu_i + \tilde{e}_{ijk}$. In matrix notation, this model can be more generically written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{e}}$, with $\boldsymbol{y}$ the vector of all observations, $\boldsymbol{X}$ the design matrix for the fixed effects of class type, $\boldsymbol{\beta}$ the vector of fixed effects parameters, and $\tilde{\boldsymbol{e}}$ the vector of residuals having a multivariate normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{V}$. In linear regression, $\boldsymbol{V}$ would be equal to $\sigma_E^2\boldsymbol{I}$, with $\boldsymbol{I}$ the identity matrix having a value 1 at the diagonal and a value 0 at the off-diagonals, but for ANOVA models with random effects, like our model (2.11), the random effects induce a correlation structure for $\tilde{\boldsymbol{e}}$ that results into the matrix $\boldsymbol{V}$ that is different from $\boldsymbol{I}$.

For model (2.11), the design matrix $\boldsymbol{X}$ and variance-covariance matrix $\boldsymbol{V}$ are defined by

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{1}_{I_1} & \boldsymbol{0}_{I_1} \\ \boldsymbol{0}_{I_2} & \boldsymbol{1}_{I_2} \end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{0}_{K_{1.}\times K_{2.}} \\ \boldsymbol{0}_{K_{2.}\times K_{1.}} & \boldsymbol{V}_2 \end{pmatrix},$$

with $\mathbf{n}_p$ a vector of length p having the value n at each enrty (e.g., $\boldsymbol{1}_5 = (1,1,1,1,1)^T$), $\boldsymbol{0}_{r\times s}$ a $r \times s$ matrix completely filled with zeros, $K_{i.}$ the number of children in class type i (i.e., $K_{i.} = K_{i1} + K_{i2} + \cdots + K_{iJ_i}$), and $\boldsymbol{V}_i$ the variance-covariance matrix for the residuals in class type i given by

$$\boldsymbol{V}_i = \begin{pmatrix} \boldsymbol{V}_{K_{i1}\times K_{i1}} & \boldsymbol{0}_{K_{i1}\times K_{i2}} & \cdots & \boldsymbol{0}_{K_{i1}\times K_{iJ_i}} \\ \boldsymbol{0}_{K_{i2}\times K_{i1}} & \boldsymbol{V}_{K_{i2}\times K_{i2}} & \cdots & \boldsymbol{0}_{K_{i2}\times K_{iJ_i}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{K_{iJ_i}\times K_{i1}} & \boldsymbol{0}_{K_{iJ_i}\times K_{i2}} & \cdots & \boldsymbol{V}_{K_{iJ_i}\times K_{iJ_i}} \end{pmatrix},$$

with $\boldsymbol{V}_{K_{ij}\times K_{ij}}$ the $K_{ij}\times K_{ij}$ variance-covariance matrix for the children in class j of class type i given by

$$\boldsymbol{V}_{K_{ij}\times K_{ij}} = [\sigma^2_{C(T)} + \sigma^2_E] \begin{pmatrix} 1 & \rho_W & \cdots & \rho_W \\ \rho_W & 1 & \cdots & \rho_W \\ \vdots & \vdots & \ddots & \vdots \\ \rho_W & \rho_W & \cdots & 1 \end{pmatrix}$$

and ρ_W the intraclass correlation coefficient defined in (2.12).

2.5.1 Fixed effects estimators

In terms of these matrices, the OLS estimators $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ are defined by $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \text{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, where the residual correlation structure $\boldsymbol{V}$ is obviously not involved. The GLS estimators $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ are defined by $\hat{\boldsymbol{\beta}}_{\text{GLS}} = \text{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, which now shows a dependency on the variance-covariance matrix $\boldsymbol{V}$. Thus the only difference between the OLS and GLS estimator is the involvement of the covariance structure of the residuals in minimizing the squared distance of the observation to their expectation. In case the variance-covariance matrix $\boldsymbol{V}$ is a diagonal with different elements on the diagonal, the GLS is referred to as the weighted least squares (WLS).

An explicit expression for the GLS estimators in terms of these matrices can be determined. It can be demonstrated that the solution is equal to

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}, \tag{2.20}$$

assuming that the involved matrices are invertible. It is not directly obvious that for balanced data we have $\hat{\boldsymbol{\beta}}_{\text{GLS}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, but the balance in data imposes nice properties on the structure of $\boldsymbol{V}$, which makes the two solutions identical (Searle *et al.*, 2006). This does not mean that $\boldsymbol{V}^{-1}$ reduces to the identity matrix, but rather that its influence on the solution vanishes. However, in the unbalanced case the OLS and GLS estimators may be different, implying that the estimator for the fixed effects parameter μ_i in model (2.11) would not be equal to the average $\bar{y}_{i..}$ at class type i anymore.

It should be mentioned that both estimators $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ are unbiased estimators for the fixed effects (Searle *et al.*, 2006), i.e., $\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{GLS}}] = \boldsymbol{X}\boldsymbol{\beta}$. The variances of the two estimators are given by

$$\begin{array}{rcl} \mathsf{VAR}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) & = & (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \\ \mathsf{VAR}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) & = & (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}. \end{array} \tag{2.21}$$

Although they are difficult to compare, it has been demonstrated that the variance of the GLS estimators is less than or equal to the variance of the OLS estimators. This is the reason why the GLS is preferred over the OLS estimators. However, an important complicating factor is that $\boldsymbol{V}$ is typically unknown in practice, which implies that $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ can not be determined without estimating the variance components first. A consequence of the estimation of $\boldsymbol{V}$ is that the GLS estimator may not be unbiased anymore when the variance-covariance matrix $\boldsymbol{V}$ is replaced by an estimator, i.e., we may have

$$\mathbb{E}[(\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}] \neq \boldsymbol{\beta},$$

when $\hat{\boldsymbol{V}}$ is an estimated variance-covariance matrix for $\boldsymbol{V}$. When sample sizes for all variance components are large, the bias may be minimal, since $\boldsymbol{V}$ is estimated precisely, but when sample sizes are relatively small, estimation of $\boldsymbol{V}$ may cause some difficulties. Clearly, this also affects the calculation of the variance of the GLS estimators. This may require adjustments in the calculation of the variance of $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ when $\boldsymbol{V}$ is estimated and may therefore loose some of its preference. Note that the OLS estimators $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ are not concerned with these issues of estimating $\boldsymbol{V}$ and is therefore always unbiased. This is sometimes (correctly) used as arguments for choosing OLS over GLS estimators.

2.5.2 Variance component estimators

One approach for the estimation of the variance components in $\boldsymbol{V}$ for unbalanced data is based on Henderson's method III for *reduction* in sums of squares (Henderson, 1953). However, the reduction in sums of squares can be calculated in different ways and are therefore not unique. For balanced data, Henderson's method III reduces to the sums of squares that we discussed in detail above. Here we will explain two different ways of calculating reductions in sums of squares, which are referred to as the TYPE 1 and TYPE 3 method.

The TYPE 1 estimation method starts with a model without any factors involved (except the intercept) and then sequentially add one factor at the time. Whether factors are random or fixed is essentially irrelevant. Calculating the reduction in the residual sums of squares of the previous model by adding another factor is then the sums of squares used for this factor. Using the number of degrees of freedom for this factor, we can calculate the mean squares and then determine the expected mean squares. The order in which you can add the factors in the model sequentially is at the discretion of the user, as long as the user maintains a statistically hierarchical model. Thus a nested factor comes after the factor that contains the nested factor, e.g., class type comes before class within class type. Given the order of the factors, the TYPE 1 sums of squares are fixed.

The TYPE 3 estimation method calculates the reduction in the residual sums of squares for one factor when all other factors are already included in the model. Thus if we would have three factors A, B, and C, the TYPE 3 method calculates the contribution of factor C when factor A and B are already included in the model, but the contribution of factor A is calculated when the model contains the factor B and C. Thus the effect of one factor is calculated after the effects of all the other factors are corrected for. Due to the complete nested structure of model (2.11), TYPE 1 and TYPE 3 estimation will not be different for this model, but for many other models this will not be the true (see Exercise 4).

The TYPE 1 and TYPE 3 estimation approaches are implemented in procedure MIXED of **SAS** software. In the **SAS** code box 2.5.2, the programming codes together with an explanation of the statements are provided for model (2.11). The output of procedure MIXED of **SAS** is quite extensive, since it contains the parameter estimates, the sums of squares, the mean squares, the degrees of freedom, the expected mean squares, and the (approximate) F-tests for testing the contribution of the factors to the variability of the outcome. Here we will discuss only the most relevant aspects.

SAS code box: ANOVA Unbalanced Data

```
PROC MIXED DATA = schooldata METHOD = TYPE1 CL COVTEST;
  CLASS COMBI CLASS;
  MODEL IQV = COMBI / SOLUTION CL DDFM = SAT OUTP = PRED;
  RANDOM CLASS(COMBI);
RUN;
```

Any procedure in **SAS** starts with **PROC** followed by the name of the procedure (here **MIXED**). Since each procedure needs to know what data should be used, there is a "DATA = " statement. Our data is stored in `schooldata`. The statement "METHOD = TYPE1" informs the procedure to use the TYPE1 estimation technique. Here we can also use of TYPE3 course. The option "CL" will provide 95% confidence intervals on the variance components, but with the option "COVTEST" we also produce the standard errors of the variance component estimators. The statement "CLASS `COMBI CLASS`" tells the procedure that the two variables `COMBI` and `CLASS` in the data set must be treated as categorical. The statement "MODEL `IQV = COMBI`" is a regression analysis statement, where the IQ-score is the dependent variable and `COMBI` is the independent variable. Everything after this equality sign is treated as fixed effects. The options "SOLUTION CL" will provide the fixed effects estimates together with their 95% confidence intervals. The option "DDFM = SAT" calculates the Satterthwaite degrees of freedom for the estimated variance in (2.21). The option "OUTP = PRED" stores the predictions $\hat{y}_{ij}$ for each child (without considering the residuals e_{ij}). Finally, the statement "RANDOM `CLASS(COMBI)`" informs the procedure to consider the variable `CLASS` as nested within `COMBI` and to treat this nested variable as a random effect.

Example 1. Moment-based estimation of model (2.11)

The procedure **MIXED** parameterized the two fixed effects parameters μ_1 and μ_2 of model (2.11) as the mean of the second class type μ_2 and a difference between class types $\alpha = \mu_1 - \mu_2$. The output table "**Solution for Fixed Effects**" in the **SAS** output, provides the estimates for the fixed effects together with their 95% confidence interval: $\hat{\mu}_2 = 11.64\,[11.43;\, 11.85]$ and $\hat{\alpha} = 0.218\,[-0.040;\, 0.475]$. The standard errors of the fixed effects are calculated according to GLS formula in (2.21) where the variance-covariance matrix $\boldsymbol{V}$ is estimated with the TYPE 1 variance component estimators. They were estimated at 0.1075 and 0.1306, respectively. Sattertwaite degrees of freedom for these standard errors were determined at 265 and 236. Note that these deviate from 204, the degrees of freedom for the mean square of class within class type. In a balanced setting, the standard error would be fully determined by $MS_{C(T)}$ (see the fixed effects estimation in Section 2.4 and (2.18)).

The output table "**Covariance Parameter Estimates**" provides the estimates for the variance components $\sigma^2_{C(T)}$ and σ^2_R with their 95% confidence intervals: $\hat{\sigma}^2_{C(T)} = 0.5420\,[0.3967;\, 0.6873]$ and $\hat{\sigma}^2_R = 3.8056\,[3.6381;\, 3.9850]$. These estimates were calculated from the "**Analysis of Variance**" table (which contains the expected mean squares) shown in Table 2.4.

For a balanced ANOVA the standard error of the residual variance component can be calculated by $[2MS^2_R/df_R]^{1/2}$ and would be equal to 0.0884. The standard error of the variance component estimator $\hat{\sigma}^2_{C(T)}$ can be calculated by

Table 2.4: ANOVA table for model (2.11) using TYPE1 estimation.

Factor	*df*	*SS*	*MS*	$\mathbb{E}(MS)$
T	1	12.870352	12.87	$\sigma^2_R + 16.629\sigma^2_{C(T)} + Q(\mu_2, \alpha)$
$C(T)$	204	2874.695337	14.09	$\sigma^2_R + 18.978\sigma^2_{C(T)}$
R	3704	14096	3.8056	σ^2_R

$2[MS^2_{C(T)}/df_{C(T)} + MS^2_R/df_R]/(18.978)^2$ and would be equal to 0.0737. **SAS** reports the standard errors of 0.0886 and 0.0741. They deviate somewhat from our calculations, because the mean squares are not independent anymore. **SAS** takes this into account. Note that the standard error of the residual mean square is not used for calculating the confidence interval, since it makes use of the χ^2-distribution with $df_R = 3704$ degrees of freedom. The interval is calculated as $[3.6381; 3.9850] = [3704 \times 3.8056/3874.58; 3704 \times 3.8056/3537.21]$. For the confidence interval on $\sigma^2_{C(T)}$, the form in (2.18) is used, leading to $[0.3967; 0.6873] = 0.5420 \pm 1.96 \times 0.0741$.

The ANOVA table shows that $\mathbb{E}(MS_T)$ does not result to $\mathbb{E}(MS_{C(T)})$ when factor class type T does not contribute to the variability in the outcome ($H_0 : \alpha = 0$), like we have seen for balanced data in (2.13). To test null hypothesis $H_0 : \alpha = 0$, a linear combination of $MS_{C(T)}$ and MS_R is used in the denominator of the F-test. This error term with its Satterthwaite degrees of freedom, is provided in the output of procedure **MIXED** (see Table 2.5):

Table 2.5: Hypothesis testing for model (2.11) using TYPE1 estimation.

Factor	Error Term	Error *df*	F-value	P-value
T	$0.8762MS_{C(T)} + 0.1238MS_R$	219.85	1.00	0.3174
$C(T)$	MS_R	3704	3.70	$<.0001$

The output shows that the null hypothesis $H_0 : \alpha = 0$ (or $H_0 : \mu_1 = \mu_2$) is not rejected at the significance level of 0.05, since the p-value of the approximate F-test is determined at $p = 0.317$. On the other hand, the null hypothesis on the variance component for class within class type $H_0 : \sigma^2_{C(T)} = 0$ is rejected at significance level 0.05. The p-value is determined at $p < 0.001$. Note that this F-test is not an exact F-test either, even though the expected mean square $MS_{C(T)}$ reduces to MS_R when the null hypothesis is true. The reason is that the two features of a χ^2-distribution and independence for mean squares does not hold in the unbalanced setting.

Investigating the predictions for the IQ per class, we have listed the five highest and five lowest IQ scores for the classes in Table 2.6. These predictions represent the BLUPs for each class given in (2.17). We also calculated the class averages for these ten most extreme predictions to compare the differences with just the arithmetic averages.

The predictions are less extreme than the averages, as expected. We also see that the ranking in predictions is different from the rankings in the averages. The average IQ for the class with the second smallest prediction is larger than the average IQ for the class with the third smallest prediction. In this case the lowest rankings 2 and 3 are interchanged for the averages, by ranking the

Table 2.6: Most extreme predictions of IQ per class

Rankings		1	2	3	4	5
Highest IQ	Prediction ($\hat{y}_{ij}$)	13.59	13.18	13.17	13.15	13.14
	Average ($\bar{y}_{ij.}$)	14.34	13.90	13.93	13.68	13.55
Lowest IQ	Prediction ($\hat{y}_{ij}$)	9.07	9.50	9.63	10.15	10.32
	Average ($\bar{y}_{ij.}$)	7.06	8.58	8.46	9.53	9.00

predictions. Note that theoretically there may exist a class with an average IQ that is smaller than 7.06 and that is ranked sixth or higher with the predictions. It is recommended to use and report the predictions for the class IQ's instead of using the averages (Efron and Morris, 1977).

2.6 Maximum likelihood approaches

For the method of moments we have seen that estimation of the fixed effects parameters and the estimation of the variance components are separated. The variance components are estimated first and are then used in the generalized least squares estimators for the fixed effects. By assuming specific distributions for the random effects, we could construct a likelihood function and then estimate all parameters simultaneously. As we already mentioned, the most common distributional assumption on the random effects is normality. We will discuss two likelihood-based approaches: maximum likelihood (ML) estimation and restricted maximum likelihood (REML) estimation. In Section 2.7 we discuss the advantages and disadvantages of the moment-based and likelihood-based approaches.

2.6.1 Maximum likelihood estimation

To illustrate the ML estimation technique, we will focus on model (2.11) again, with all random effects being normally distributed. The likelihood function for model (2.11) is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\sigma}|\boldsymbol{y}) = \prod_{i=1}^{I}\prod_{j=1}^{J_i}\int_{\mathbb{R}}\prod_{k=1}^{K_{ij}}\left[\frac{1}{\sigma_R}\phi\left(\frac{y_{ijk}-\mu_i-z}{\sigma_R}\right)\right]\frac{1}{\sigma_{C(T)}}\phi\left(\frac{z}{\sigma_{C(T)}}\right)dz, \qquad (2.22)$$

with $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ the fixed effects parameters, $\boldsymbol{\sigma} = (\sigma_{C(T)}, \sigma_R)^T$ the variance components, and $\boldsymbol{y} = (y_{111}, y_{112}, ..., y_{2I_2K_{2I_2}})^T$ the complete vector of observations. To construct this likelihood function, we have made use of the conditional distribution of y_{ijk} given $c_{j(i)} = z$. This conditional distribution of child k is a normal distribution with mean $\mu_i + z$ and variance σ_R^2. The K_{ij} children in class j of class type i, are then independently and identically distributed with this normal distribution, conditionally on $c_{j(i)} = z$. That is why we see a product of the normal density for the K_{ij} children The random effect $c_{j(i)}$, which is also normally distributed, must then be integrated out. In this way, the likelihood function contains an integral over the repeated observation of all children in the class, but due to the assumption of normality, this integral has a closed-form expression. This closed-form represents the likelihood of an arbitrary class. Since

all classes are independent, this class likelihood should be applied to each class as a product of likelihoods.

The integrand in the likelihood function (2.22) is an exponential function with a quadratic function in z:

$$\left[2\pi\sigma_R^{K_{ij}}\sigma_{C(T)}\right]^{-1}\exp\left\{-\frac{1}{2}\left[\sigma_R^{-2}\sum_{k=1}^{K_{ij}}(y_{ijk}-\mu_i-z)^2+\sigma_{C(T)}^{-2}z^2\right]\right\}.$$

Rewriting the quadratic form within the brackets and then integrating the variable z out, results into another exponential function. Then substituting the new exponential function in the likelihood function (2.22), and then taking the logarithm, we obtain the following *log likelihood function* for model (2.11):

$$\begin{aligned}\ell(\boldsymbol{\mu},\boldsymbol{\sigma}|\boldsymbol{y}) = -\frac{1}{2}\Big[&J\log(2\pi)+\textstyle\sum_{i=1}^{I}\sum_{J=1}^{J_i}\log\left(K_{ij}\sigma_{C(T)}^2+\sigma_R^2\right)\\ &+(K-J)\log(\sigma_R^2)+\sigma_R^{-2}\textstyle\sum_{i=1}^{I}\sum_{J=1}^{J_i}\sum_{k=1}^{K_{ij}}(y_{ijk}-\mu_i)^2\\ &-\textstyle\sum_{i=1}^{I}\sum_{J=1}^{J_i}\left(\frac{\sigma_{C(T)}^2K_{ij}^2}{\sigma_R^2[K_{ij}\sigma_{C(T)}^2+\sigma_R^2]}(\bar{y}_{ij.}-\mu_i)^2\right)\Big],\end{aligned}$$

with $J=J_1+J_2$ the total number of classes and $K=\sum_{i=1}^{I}\sum_{j=1}^{J_i}K_{ij}$ the total number of children.

It should be noted that maximizing the likelihood function is the same as maximizing the log likelihood function, but the log likelihood function is mathematically easier to handle. Furthermore, maximization of the (log) likelihood must be conducted under restriction that all variance components are non-negative (i.e., $\sigma_{C(T)}^2\geq 0$). The parameters μ_1, μ_2, $\sigma_{C(T)}^2$, and σ_R^2 that maximize the log likelihood function, can be found by taking the derivatives of the log likelihood function with respect to μ_1, μ_2, $\sigma_{C(T)}^2$, and σ_R^2 and then equating them to zero. This set of equations is called the set of *likelihood equations* and the solutions $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_{C(T)}^2$, and $\hat{\sigma}_R^2$ of this set of equations are referred to as the *ML estimators*. However, the solution for this set of equations can be outside the parameter space, i.e., resulting into a negative variance component as we have also seen for the mean squares approach. In case the solution $\hat{\sigma}_{C(T)}^2$ becomes negative, the ML estimator sets the solution to zero. This truncated solution is then implemented in the log likelihood function and a new set of likelihood equations is being determined where now only the derivatives with respect to μ_1, μ_2, and σ_R^2 is considered. Thus a potential negative variance component estimator effectively removes the corresponding factor from the model. Note that it can be demonstrated that the solution $\hat{\sigma}_R^2$ is always positive and therefore will never be removed from the set of likelihood equations.

For unbalanced data the ML estimators can not be written in closed-form expressions. This is mostly due to the estimation of the variance components. To illustrate this point we have calculated the likelihood equations for the parameters μ_1, μ_2, $\sigma_{C(T)}^2$, and σ_R^2 in model (2.11). These equations are given

by

$$\begin{aligned}
&\hat{\mu}_i = \frac{\sum_{J=1}^{J_i}\left[\sigma_R^2[K_{ij}\sigma_{C(T)}^2+\sigma_R^2]^{-1}\bar{y}_{ij.}\right]}{\sum_{J=1}^{J_i}\sigma_R^2[K_{ij}\sigma_{C(T)}^2+\sigma_R^2]^{-1}},\\
&\sum_{i=1}^{I}\sum_{J=1}^{J_i}\left[\frac{K_{ij}}{K_{ij}\sigma_{C(T)}^2+\sigma_R^2}\right] = \sum_{i=1}^{I}\sum_{J=1}^{J_i}\left[\frac{K_{ij}^2(\bar{y}_{ij.}-\mu_i)^2}{[K_{ij}\sigma_{C(T)}^2+\sigma_R^2]^2}\right], \qquad (2.23)\\
&\sum_{i=1}^{I}\sum_{J=1}^{J_i}\left[\frac{1}{K_{ij}\sigma_{C(T)}^2+\sigma_R^2}\right] + \frac{K-J}{\sigma_R^2} = \frac{SS_R}{\sigma_R^4} + \sum_{i=1}^{I}\sum_{J=1}^{J_i}\left[\frac{K_{ij}(\bar{y}_{ij.}-\mu_i)^2}{[K_{ij}\sigma_{C(T)}^2+\sigma_R^2]^2}\right],
\end{aligned}$$

where SS_R is the residual sums of squares $SS_R = \sum_{i=1}^{I}\sum_{J=1}^{J_i}\sum_{k=1}^{K_{ij}}(y_{ijk}-\bar{y}_{ij.})^2$. Note that it required a bit of algebraic manipulations to get the equations in (2.23), in particular the equation for variance components σ_R^2 (the third equation). We have made use of $y_{ijk}-\mu_i = y_{ijk}-\bar{y}_{ij.}+\bar{y}_{ij.}-\mu_i$ in the log likelihood function to simply the equations. Solving the variance components $\sigma_{C(T)}^2$ and σ_R^2 from the equations (2.23) can only be determined through numerical procedures. When the variance components are determined, the ML estimators for the fixed effects come in an explicit form, but they do depend on the variance components. We have seen something similar for the generalized least squares estimator.

For balanced data the equations in (2.23) do clean up somewhat. Indeed, when the number of children per class would have been constant across classes (i.e., $K_{ij} = K$), the ML solution for μ_i reduces to $\hat{\mu}_i = \bar{y}_{i..}$, the average IQ-score for class type i. Having averages as ML estimators for the fixed effects parameters holds true more generally. In balanced ANOVA models the ML estimators for the fixed effects are equal to the ordinary least squares estimators. Knowing that $\hat{\mu}_i = \bar{y}_{i..}$, we can substitute this ML solution in the two last equations of (2.23). The second equation then reduces to

$$SS_{C(T)} = IJ[K\sigma_{C(T)}^2+\sigma_R^2],$$

with $SS_{C(T)} = \sum_{i=1}^{I}\sum_{J=1}^{J}K(\bar{y}_{ij.}-\bar{y}_{i..})^2$ the sums of squares for class within class type. Now substituting $SS_{C(T)}/(IJ)$ for $K\sigma_{C(T)}^2+\sigma_R^2$ in the third equation of (2.23), we obtain the residual mean square $\hat{\sigma}_R^2 = SS_R/[IJ(K-1)] = MS_R$ as ML solution for σ_R^2. Using this solution, the ML solution for $\sigma_{C(T)}^2$ is then equal to $\hat{\sigma}_{C(T)}^2 = [SS_{C(T)}/(IJ)-MS_R]/K$. Thus the ML solutions for the variance components are also functions of the sums of squares, similar to the method of moments. However, The ML solution $\hat{\sigma}_{C(T)}^2$ is different from the mean squares approach, since $SS_{C(T)}/(IJ) = (J-1)MS_{C(T)}/J$ is not equal to $MS_{C(T)}$. Thus the ML solutions for the variance components (except for the residual variance component) is typically bias. Indeed, the expected value of the ML solution $\hat{\sigma}_{C(T)}^2$ is now equal to

$$\mathbb{E}[\hat{\sigma}_{C(T)}^2] = [1-J^{-1}]\sigma_{C(T)}^2 - \sigma_R^2/(JK) < \sigma_{C(T)}^2.$$

Thus the ML solutions typically underestimate the variance components. This is a feature of maximum likelihood estimation that is true for all variance components except for the residual variance.

As we already indicated above, the solution $\hat{\sigma}^2_{C(T)}$ becomes the ML estimator only when it is positive. In case $\hat{\sigma}^2_{C(T)} \leq 0$, the estimator $\hat{\sigma}^2_{C(T)}$ is truncated to zero and this solution is then implemented in the log likelihood (2.22) to make it possible to determine the ML estimators for μ_i and σ^2_R. Under the assumption that $\sigma^2_{C(T)} = 0$, the ML solutions for μ_i and σ^2_R are given by $\hat{\mu}_i = \bar{y}_{i..}$ and $\hat{\sigma}^2_R = [SS_{C(T)} + SS_R]/(IJK)$. Thus the ML solution for fixed effects parameters are not altered (although for unbalanced data the fixed effects estimators can alter), the ML solution for the residual variance component is changed to a solution that contains both the residual sums of squares and the sums of squares for class within class type. Since the sums of squares for class within class type can not be used for the estimation of $\sigma^2_{C(T)}$ it must end up in the estimation of another variance component.

Now that we have discussed the solutions of the ML equations in (2.23), we can formulate the ML estimators. Thus the ML estimators for model (2.11) with balanced data are now given by

$$\begin{array}{rcl} \hat{\mu}_i & = & \bar{y}_{i..} \\ \hat{\sigma}^2_{C(T)} & = & \max\left\{0, [SS_{C(T)}/(IJ) - MS_R]/K\right\} \\ \hat{\sigma}^2_R & = & \begin{cases} MS_R & \text{if} \quad \hat{\sigma}^2_{C(T)} > 0 \\ [SS_{C(T)} + SS_R]/(IJK) & \text{if} \quad \hat{\sigma}^2_{C(T)} \leq 0 \end{cases} \end{array} \tag{2.24}$$

Determining the expectations of the ML estimators for variance components is more complicated in general, due to the dependence of having negative ML solutions that are being truncated. However, it can be easily seen that the ML estimator for the residual variance is now biased too, since the unbiased estimator MS_R is altered when $\hat{\sigma}^2_{C(T)} \leq 0$. On the other hand, when sample sizes would be large enough negative variance component estimators would never occur when the true variance components are positive. Secondly, the bias of the ML solutions would also disappear due to large sample sizes.

Calculation of the standard errors of the ML estimators is based on the second derivative of the log likelihood estimators with respect to the model parameters. The second derivatives of the log likelihood function form a matrix with a dimension equal to the number of parameters. Thus for model (2.11) the matrix would be a 4×4 matrix. This matrix is referred to as the *Hessian matrix*. It may contain both the model parameters and the observed data. If we then take the expectation of the elements of the Hessian matrix and multiply these expectations with minus one, we obtain the so-called *Fisher information matrix*. The inverse Fisher information matrix represent the *asymptotic standard errors* of the ML estimators (Searle *et al.*, 2006; McCulloch and Searle, 2001).

To illustrate this, we consider our model (2.11) for balanced data, but note that the approach also applies to unbalanced data. The second derivatives are

$$\begin{array}{l} \dot{\ell}_{\mu_i,\mu_i} = -JK\lambda^{-1}, \\ \dot{\ell}_{\mu_r,\mu_s} = 0, \ r \neq s, \\ \dot{\ell}_{\mu_i,\sigma^2_{C(T)}} = -JK^2\lambda^{-2}(\bar{y}_{i..} - \mu_i) \\ \dot{\ell}_{\mu_i,\sigma^2_R} = -JK\lambda^{-2}(\bar{y}_{i..} - \mu_i) \end{array}$$

$$\dot{\ell}_{\sigma^2_{C(T)},\sigma^2_{C(T)}} = \tfrac{1}{2}IJK^2\lambda^{-2} - \sum_{i=1}^{I}\sum_{j=1}^{J}\left[K^3\lambda^{-3}(\bar{y}_{ij.}-\mu_i)^2\right],$$
$$\dot{\ell}_{\sigma^2_{C(T)},\sigma^2_{R}} = \tfrac{1}{2}IJK\lambda^{-2} - \sum_{i=1}^{I}\sum_{j=1}^{J}\left[K^2\lambda^{-3}(\bar{y}_{ij.}-\mu_i)^2\right],$$
$$\dot{\ell}_{\sigma^2_{R},\sigma^2_{R}} = \tfrac{1}{2}IJ\left[\lambda^{-2}+(K-1)\sigma_R^{-4}\right] - \sigma_R^{-6}SS_R - \sum_{i=1}^{I}\sum_{j=1}^{J}\left[K\lambda^{-3}(\bar{y}_{ij.}-\mu_i)^2\right],$$

with $\dot{\ell}_{a,b}$ defined by $\partial^2\ell(\boldsymbol{\mu},\boldsymbol{\sigma}|\boldsymbol{y})/(\partial a\partial b)$ and with $\lambda = K\sigma^2_{C(T)}+\sigma^2_R$. If we now take the expectations of these second derivatives and multiple them with minus one, the Fisher information matrix becomes equal to

$$I(\boldsymbol{\mu},\boldsymbol{\sigma}) = \begin{pmatrix} JK\lambda^{-1} & 0 & 0 & 0 \\ 0 & JK\lambda^{-1} & 0 & 0 \\ 0 & 0 & \frac{1}{2}IJK^2\lambda^{-2} & \frac{1}{2}IJK\lambda^{-2} \\ 0 & 0 & \frac{1}{2}IJK\lambda^{-2} & \frac{1}{2}IJ\left[\lambda^{-2}+(K-1)\sigma_R^{-4}\right] \end{pmatrix}.$$

Note that we have used that the expectation $\mathbb{E}(\bar{y}_{ij.}-\mu_i)^2$ is equal to λ/K and that the expectation of SS_R is given by $IJ(K-1)\sigma^2_R$. Taking the inverse of the Fisher information matrix, we obtain the following matrix:

$$I^{-1}(\boldsymbol{\mu},\boldsymbol{\sigma}) = \begin{pmatrix} \frac{\lambda}{JK} & 0 & 0 & 0 \\ 0 & \frac{\lambda}{JK} & 0 & 0 \\ 0 & 0 & \frac{2\sigma^4_R}{IJK^2}\left[\frac{1}{K-1}+\frac{\lambda^2}{\sigma^4_R}\right] & -\frac{2\sigma^4_R}{IJK(K-1)} \\ 0 & 0 & -\frac{2\sigma^4_R}{IJK(K-1)} & \frac{2\sigma^4_R}{IJ(K-1)} \end{pmatrix}. \qquad (2.25)$$

The inverse Fisher information matrix informs us that the asymptotic variances of the ML estimators are now given by

$$\begin{aligned} \mathsf{VAR}(\hat{\mu}_i) &= [K\sigma^2_{C(T)}+\sigma^2_R]/[JK] \\ \mathsf{VAR}(\hat{\sigma}^2_{C(T)}) &= 2[K\sigma^2_{C(T)}+\sigma^2_R]^2/[IJK^2] + 2\sigma^4_R/[IJK^2(K-1)] \\ \mathsf{VAR}(\hat{\sigma}^2_R) &= 2\sigma^4_R/[IJ(K-1)] \end{aligned}$$

The covariances between the fixed effects estimators are zero, the covariances between the fixed effects and random effects estimators are also zero, but the covariance between the two random effects estimators is equal to $\mathsf{COV}(\hat{\sigma}^2_{C(T)},\hat{\sigma}^2_R) = -2\sigma^2_R/[IJK(K-1)]$. Note that we do not have consider that the ML solutions for variance component estimators can be negative, since asymptotically this does not play a serious problem.

Clearly, the asymptotic variances and covariances still contain (some of the) model parameters, so they must be estimated to make them useful in practice. The variances and covariances can be estimated by substituting the ML estimators for the parameters in the inverse Fisher information matrix. With these estimates, asymptotic confidence intervals for the model parameters of the form (2.18) can be determined in the same way as we have discussed for the method of moments. The procedure MIXED of **SAS** has implemented the maximum likelihood estimators, the estimated variances and covariances, and the asymptotic confidence intervals for both balanced and unbalanced data (see Example 2). The confidence interval for the residual variance is based on the χ^2-distribution, the same as for the method of moments.

2.6.2 Restricted maximum likelihood estimation

We have seen that the maximum likelihood solutions for variance components can be biased. Indeed, the ML solution $\hat{\sigma}^2_{C(T)} = [SS_{C(T)}/(IJ) - MS_R]/K$ for the variance component $\sigma^2_{C(T)}$ of model (2.11) underestimates the variance component. The problem is that ML estimation does not correct for the number of fixed effects parameters that must be estimated too. This is a well known issue and occurs even in the simplest case where y_1, y_2,....., $y_n \sim N(\mu, \sigma^2)$ are independent and identically normally distributed. In this simple setting, the ML estimator for σ^2 is given by $\sum_{i=1}^{n}(y_i - \bar{y})^2/n$, while we would rather like to divide by the degrees of freedom $n-1$ instead of the sample size n if we wish to use an unbiased estimator. The use of the degrees of freedom $n-1$ would account for the loss of one degrees of freedom for estimation of μ.

Restricted or *residual maximum likelihood* (REML) estimation takes care of this problem by maximizing a different likelihood function where the fixed effects are essentially eliminated from the likelihood. If we would study the log likelihood function in (2.22) for balanced data, implement $\bar{y}_{ijk} - \bar{y}_{ij.} + \bar{y}_{i..} - \mu_i$ for $\bar{y}_{ijk} - \mu_i$, and implement $\bar{y}_{ij.} - \bar{y}_{i..} + \bar{y}_{i..} - \mu_i$ for $\bar{y}_{ij.} - \mu_i$, we obtain the following log likelihood function:

$$\begin{array}{l} -\frac{I}{2}[\log(2\pi) + \log(\lambda)] - \frac{1}{2\lambda}\sum_{i=1}^{I} JK(\bar{y}_{i..} - \mu_i)^2 \\ \quad -\frac{I(J-1)}{2}[\log(2\pi) + \log(\lambda)] - \frac{IJ(K-1)}{2}\log(\sigma^2_R) - \frac{SS_R}{2\sigma^2_R} - \frac{SS_{C(T)}}{2\lambda}, \end{array} \tag{2.26}$$

with $\bar{y}_{i..}$ the average IQ-score for class type i, SS_R and $SS_{C(T)}$ the sums of squares for the residual and the class within class type, respectively, and with $\lambda = K\sigma^2_{C(T)} + \sigma^2_R$ as defined before. The first line of the log likelihood function in (2.26) can be viewed as the log likelihood function for μ_1 and μ_2 given the averages $\bar{y}_{1..}$ and $\bar{y}_{2..}$, while the second line represents a log likelihood function for the variance components $\sigma^2_{C(T)}$ and σ^2_R given the sums of squares $SS_{C(T)}$ and SS_R. If we would just maximize the log likelihood function for the variance components, we obtain the solutions

$$\begin{array}{rcl} \hat{\sigma}^2_R & = & MS_R, \\ \hat{\sigma}^2_{C(T)} & = & [MS_{C(T)} - MS_R]/K, \end{array}$$

with $MS_R = SS_R/[IJ(K-1)]$ and $MS_{C(T)} = SS_{C(T)}/[I(J-1)]$ the mean squares. These solutions are identical to the moment estimators and are indeed unbiased for the variance components σ^2_R and $\sigma^2_{C(T)}$, respectively. However, these solutions are not yet the REML estimators, since we need to determine if the solutions for the variance components are negative, like we did for the ML estimators. In case negative estimates occur, the solutions are truncated to zero, the corresponding variance components are set to zero in the log likelihood function (2.26), and the remaining variance components are being estimated with this altered log likelihood function. Thus the REML estimators for the variance components of model (2.11) when the data is balanced are

$$\begin{array}{rcl} \hat{\sigma}^2_{C(T)} & = & \max\left\{0, [MS_{C(T)} - MS_R]/K\right\}, \\ \hat{\sigma}^2_R & = & \left\{ \begin{array}{lll} MS_R & \text{if} & \hat{\sigma}^2_{C(T)} > 0, \\ [SS_{C(T)} + SS_R]/[I(JK-1)] & \text{if} & \hat{\sigma}^2_{C(T)} \leq 0. \end{array} \right. \end{array} \tag{2.27}$$

If we now substitute these REML estimators into the log likelihood function (2.26) and then maximize this function with respect to the fixed effects estimators we obtain the REML estimator $\bar{y}_{i..}$ for μ_i. For balanced ANOVA models, the solution for the fixed effects parameters are independent of the variance component estimators. Thus for balanced ANOVA models the REML estimators for the fixed effects parameters are identical to the ML estimators (and to the generalized least squares estimators). However, for unbalanced ANOVA models they can be different from each other (see Example 2), since the two-step approach of the REML estimator affects the likelihood for the fixed effects.

Calculation of the standard errors are conducted in the same way as for ML estimation, i.e., through the inverse Fisher information matrix in (2.25). The full log likelihood function in (2.26) is being used for this. Estimation of the elements are conducted by substituting the REML estimators for the parameters in the inverse Fisher information matrix. Based on these asymptotic variances, confidence intervals can be obtained as we did for ML estimators. The use of Satterthwaite degrees of freedom is also recommended here when chi-square distributions are being used for testing null hypotheses and calculation of confidence intervals.

For unbalanced data, factorizing the likelihood like we did for the balanced ANOVA model (2.11), is more difficult (McCulloch and Searle, 2001). To explain REML estimators more generally, we will change to the matrix notation that we already used for the moment estimators in Section 2.5. Here we have discussed that the vector of all observations $\boldsymbol{y}$ is normally distributed with mean $\boldsymbol{X\beta}$ and variance-covariance matrix $\boldsymbol{V}$. The REML estimators maximize a likelihood function for linear combinations of the observations $\boldsymbol{A}^T\boldsymbol{y}$, with $\boldsymbol{A}$ a matrix of rank $N - r_{\boldsymbol{X}}$, N the total number of observations, and $r_{\boldsymbol{X}}$ the rank of matrix $\boldsymbol{X}$. The matrix $\boldsymbol{A}$ is chosen such that the fixed effects parameters are eliminated, i.e., $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$. With this condition, we obtain that $\boldsymbol{A}^T\boldsymbol{y}$ is normally distributed with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}$. Thus the distribution function of the linear transformed outcome variable has only unknown variance components, which can be estimated by maximizing the corresponding log likelihood function. After the variance components are estimated with the REML estimator $\hat{\boldsymbol{V}}$, the full log likelihood can be maximized over the fixed effects, where the unknown variance-covariance matrix $\boldsymbol{V}$ is being replaced by $\hat{\boldsymbol{V}}$. Thus REML estimation is a two-step procedure, where first the matrix $\boldsymbol{V}$ is being estimated, and then in the second step the fixed effects $\boldsymbol{\beta}$ are being estimated under the assumption that $\boldsymbol{V}$ is known and equal to $\hat{\boldsymbol{V}}$. The procedure MIXED of **SAS** has implemented the REML estimators and all associated statistics for both balanced and unbalanced data (see Example 2). This estimation approach is the default estimation method for procedure MIXED.

Example 2. Likelihood-based estimation of model (2.11)

Maximum likelihood approaches for ANOVA models can be implemented with the code in **SAS** code box 2.5.2, but with a different option for the estimation method. For ML estimation we must use METHOD = ML instead of METHOD = TYPE1 and for REML estimation we must use METHOD = REML. If we eliminate the option "METHOD =" from the codes all together, the procedure MIXED implements REML estimation.

Table 2.7 shows the estimates and their 95% confidence intervals for the parameters of model (2.11) using different estimation techniques (including the

Table 2.7: Parameter estimates and 95% confidence intervals for model (2.11) using different estimation techniques.

Parameter	TYPE 1/TYPE 3	REML	ML
μ_1	11.64 [11.43; 11.85]	11.74 [11.59; 11.89]	11.64 [11.42; 11.86]
$\mu_2 - \mu_1$	0.218 [-0.040; 0.475]	0.137 [-0.033; 0.307]	0.221 [-0.045; 0.486]
$\sigma^2_{C(T)}$	0.542 [0.397; 0.687]	0.522 [0.361; 0.822]	0.593 [0.457; 0.801]
σ^2_R	3.806 [3.638; 3.985]	4.105 [3.924; 4.300]	3.816 [3.648; 3.997]

moment-based estimates) with procedure MIXED.

The results show distinct differences between the estimation techniques. The REML estimation gives the highest residual variance and the highest average for the first class type, but the smallest estimate for the difference in IQ-score between the two class types and the smallest estimate for the variance component $\sigma^2_{C(T)}$. The fixed effects estimates and the estimate for the residual variance component with ML and TYPE 1 estimation are very close, but the estimates for the variance component $\sigma^2_{C(T)}$ is larger for ML than for TYPE 1. The TYPE 1 and TYPE 3 methods are identical for model (2.11), but they are mostly different. Which of these estimation methods is most reliable will be topic of discussion in Section 2.7.

2.7 Moment-based or likelihood-based estimation?

Having several methods of estimation for ANOVA models is not helping practitioners, since they would like to know which estimation method would perform best for the available data and for the ANOVA model they are studying. If we can show that one estimation method is superior to all others there would be no discussion any more. There is no mathematical proof that would demonstrate that one estimation method is superior to the others on finite data sets. Literature mostly seems to suggest the use of REML over ML (McCulloch and Searle, 2001) and likelihood-based methods over moment-based methods (Searle, 1995; Searle *et al.*, 2006) for estimation of variance components in particular when data is unbalanced. Although most of their arguments are valid, their arguments are incomplete, leading to an advise that may be incorrect in certain settings in my opinion.

2.7.1 Arguments and counter arguments for REML estimation

The ANOVA method or method of moments are difficult or impossible to use for linear mixed models that are not ANOVA models. This statement is used to argue the preference of likelihood-based methods over moment-based methods, since likelihood-based methods do not have this restriction. However, with the introduction of generalized estimating equations (Liang and Zeger, 1986), this argument has proven to be incorrect.

For balanced data the ANOVA method is superior to the other methods for estimation of variance components, since they are unbiased and they are "minimum variance". This means that there are no other unbiased methods that have a smaller variance than the ANOVA estimators. The REML **solutions** (but not

the estimators) are identical to the ANOVA estimators for balanced data and therefore inherit these nice properties of the ANOVA estimators for balanced data. However, the REML **estimators** do not have this property, since the estimators are changed whenever a negative estimate occurs. Thus truncation changes these properties and make the REML estimators biased.

That the REML estimators can never be negative is actually seen as another advantage over the ANOVA estimators that could become negative. However, this is a (very) weak argument, since ANOVA methods can be truncated to zero as well, which is in accord with usual practice (Swallow and Monahan, 1984). And this form of truncation is actually better than the truncation approach REML is applying. For ANOVA methods, truncation to zero of a variance component estimator does not alter the other variance component estimators. For REML estimation, truncation to zero of one variance component estimator directly affects other variance component estimators and therefore make the other variance component estimators biased too. REML estimators are biased for estimation of the residual variance component (see (2.27)), while the residual variance component estimator remain unbiased when truncation for ANOVA estimators is applied.

ML estimators do not take into account the number of fixed effects parameters that must be estimated too. This is an argument to choose REML over ML. Indeed, the REML **solutions** are unbiased, while the ML **solutions** underestimate the variance components. However, truncation of the variance component estimators typically increases the expectations, making the REML estimators too large on average (overestimation of the variance components). Truncation also increases the expected value of the ML solutions, but whether these ML estimators are then less biased than the REML estimators is unknown.

For unbalanced data the ANOVA estimators are non-unique since we can choose different forms of sums of squares. Unfortunately, there are no general theorems that indicate that any of these choices is superior to any of the other choices. Additionally, there is no proof that any of these choices is better than the REML estimators in general or the other way around. Therefore, uniqueness of the REML estimators is used as an argument to choose REML over other methods. However, lack of evidence is not a good argument for choosing one over the other. It merely suggests that we do not know which method to use for unbalanced data and that more research is needed to determine the best approach for each practical situation.

Model building is more complicated with REML than with ML (or moment-based estimators), even though we did not discuss this topic here. Comparing two ANOVA models with different fixed effects models is non-trivial with REML estimators, due to the linear transformation that is used to eliminate the fixed effects for variance component estimation. The two different fixed effects lead to different transformations, which makes comparisons of the likelihoods difficult (Verbeke and Molenberghs, 2000). For model selection it is recommended to use ML and use REML when the final model is obtained (Verbeke and Moelenberghs, 2000), although there exists counter arguments for this practice (Verbyla, 2019; Gurka, 2006).

2.7.2 Advise on the use of estimation methods

The moment-based estimators do not make any distributional assumptions, except for the existence of the first two moments, while likelihood-based methods require an assumption of the distribution of the random effects (typically normal distributions). Literature has shown that ANOVA models can be used successfully for many types of non-normally distributed data and may be efficient with respect to other methods (see for instance Van den Heuvel and IJzerman-Boon, 2013). We do not agree with the argument that uniqueness is a solid reason for the use of likelihood-based methods over moment-based methods. We should use the best possible estimation technique for the situation at hand.

Our advise is roughly based on the sample sizes or degrees of freedom for the different factors that are considered in the ANOVA model.

- ▷ For balanced ANOVA models we recommend the use of the moment-based estimation. The moment-based estimators are in some way optimal and are therefore recommended. We also suggest to truncate any negative variance component estimator to zero if this would make more sense.

- ▷ We still recommend the moment-based estimation technique for unbalanced data sets that has at least one random factor with a small number of levels (say less than 10, see Figure 2.1) over likelihood-based estimation. The moment-based estimator does not alter any of the variance component estimators when one of the variance components is estimated negatively, a characteristic that is not shared by likelihood-based estimators. If the number of levels for a random factor is small, the variance component estimator is less precise and the probability for having a negative estimator may be (relatively) large (see Figure 2.1), while in reality the underlying variance component is not zero. The best estimator for this variance component is most likely equal to zero, but we do not want to convolute any of the other variance component estimators by the lack of precision for this one variance component. We recommend the use of TYPE 3 estimation over TYPE 1, unless the order of factors for entering the model is obvious and TYPE 1 makes more sense.

- ▷ We recommend ML estimation for unbalanced data sets with large numbers of levels (i.e., degrees of freedom) for each random factor (say more than 100 levels). When the number of levels are large for each random factor, ML and REML variance component estimators are almost identical and there is no real advantage of using REML over ML estimation. Likelihood-based estimation has some asymptotic optimality properties and due to the large sample sizes, it is less likely that variance component estimators are incorrectly set to zero. Thus ML is preferred over the moment-based estimation.

- ▷ When the unbalanced data set contains one or more random factors with a medium number of levels (10 to 100), we recommend the use of REML. The ML estimators for the variance components may be somewhat underestimated and the degrees of freedom for the standard errors of the fixed effects are too large. Thus confidence intervals on fixed effects with ML estimators may be too small. The REML estimator would correct for

these issues and is therefore preferred over ML estimation. It is difficult to say if they are always better to moment-based estimators, but in this way we stay aligned with recommendations in literature.

2.8 Goodness-of-fit of the ANOVA model

It is recommended to verify the model assumptions for any data analysis you do to make sure that your inference is trustworthy. This investigation is often referred to as an evaluation of the *goodness-of-fit* of the implemented model. Unfortunately, investigating all possible violations can be elaborate and complicated since violation of the model assumptions can happen in many different ways. This makes the reporting of the results of the evaluation also important. Readers can then determine what aspects of the goodness-of-fit have been evaluated. Here we will discuss a few suggestions for the evaluation of the goodness-of-fit that can be useful in practice, without trying to be complete.

For any ANOVA model there exists three types of assumptions that could be violated separately or in combination:

- ▷ **Independence:** The first assumption is that all random terms are assumed mutually independent. Investigating this assumption is the most complicated one since it requires a comparison of the ANOVA model with a linear mixed effects model that falls outside the class of ANOVA models. For instance, if we would like to assume that residuals are correlated or there exists a correlation between residuals and random effects, we are steping outside the class of ANOVA models.

- ▷ **Normality:** The second assumption is that all random terms in the ANOVA model are assumed normally distributed. Verifying this assumption is relevant when inference is based on either hypothesis testing or confidence intervals.

- ▷ **Homoscedasticity:** The third assumption is that variance components are assumed *homoscedastic*. As we already mentioned in Section 2, the variability in IQ scores between classes can depend on the type of class. Investigating homoscedastiticty is important when the focus is on estimating variance components, like inference on correlation coefficients, but *heteroscedasticity* can also affect the standard errors of fixed effects parameters.

An investigation of independence is postponed to Section 3 where mixed models is shortly being explained. We address the assumption of homoscedasticity first, because heteroscedasticity can cause a violation of normality. Indeed, if we would consider the one-way fixed effects ANOVA model in (2.1), where each group with its own mean $\mu + \alpha_i$ would also have its own variance $\sigma^2_{E,i}$, the histogram of the full data may show many non-normal shapes, depending on the values of the means and the variances. Even if we would eliminate the differences in means for the histogram, by studying the differences $y_{ij} - \bar{y}_{i\cdot}$, we may still see highly non-normal shapes (like a high peak in the middle of the data with heavy and sparse tails) due to the different variances. Thus, before normality is investigated, we better first investigate homoscedasticity

and make model corrections if this assumption is violated. Then the next step is to evaluate the assumption of normality.

2.8.1 Heteroscedasticity

In the two-way nested ANOVA model, both variance components $\sigma^2_{C(T)}$ and σ^2_R could in principle depend on other variables, like sex and type of class. However, it is uncommon to explain differences in the variance components with variables that are not included as factor in the ANOVA model. If the factor sex would influence the variance components, then the factor sex should also be included in the ANOVA model to create different means for sex. Statisticians would like to keep their models hierarchical: model first moments (different means for sex) before modeling second moments (different variances). Thus for the (unbalanced) two-way nested ANOVA model in (2.11), we will not study the effect of sex on $\sigma^2_{C(T)}$ and σ^2_R since it is not part of the ANOVA model.

The factor type of class is part of the ANOVA model and it could influence the variability. In the procedure MIXED of **SAS**, it is easy to estimate separate variance components for different levels of a (combination of) categorical factor(s). The RANDOM statement has the option to add "GROUP = `variable`". The variance components that are requested in the RANDOM statement are then calculated for each level of the included `variable` in the GROUP statement. Thus adding "GROUP = `COMBI`" will lead to an estimate of two variance components $\sigma^2_{C(T)}$: one for schools with single grades ($\sigma^2_{C(T)}(S)$) and one for schools with multi-grade ($\sigma^2_{C(T)}(M)$) classes. In this model the variance σ^2_R between children within classes is not altered.

Investigating if such an extension would make sense is easiest with the likelihood-based estimation techniques. The likelihood function can be used to compare models. If the likelihood function would increase substantially, by splitting the variance component $\sigma^2_{C(T)}$ into two separate variance components $\sigma^2_{C(T)}(S)$ and $\sigma^2_{C(T)}(M)$, there is a signal that the variability between classes is different between type of class. If the likelihood function hardly increases, it is clear that the two variance components $\sigma^2_{C(T)}(S)$ and $\sigma^2_{C(T)}(M)$ for the between class variability can be considered the same for both class types. Thus the likelihood function can be used to test between these two *hierarchical ANOVA models*. Hierarchical means that the model with less parameters can be created from the model with more parameters by making restrictions to (a subset of) the parameters. Indeed, if in the "larger" model the two variance components $\sigma^2_{C(T)}(S)$ and $\sigma^2_{C(T)}(M)$ are set equal to $\sigma^2_{C(T)}$, we obtain the "smaller" model. Thus by comparing these two hierarchical ANOVA models we would be able to test the null hypothesis on homoscedasticity $H_0 : \sigma^2_{C(T)}(S) = \sigma^2_{C(T)}(M) = \sigma^2_{C(T)}$.

Likelihood ratio test: The test statistic for testing hierarchical models is called the *likelihood ratio test* (LRT) statistic. It can be used to test any two hierarchical models, whether the reduction in the parameters is coming from variance components or from fixed effects parameters. Thus the likelihood ratio test could also be used to test the null hypothesis $H_0 : \mu_1 = \mu_2$ in ANOVA model (2.11). To formulate the likelihood ratio test statistic in its generic form, assume that we have a statistical model with the vector of parameters $\boldsymbol{\theta}$, including fixed effects and/or random effects parameters, and with log likelihood function

$\ell(\boldsymbol{\theta}|\boldsymbol{y})$. The maximum likelihood estimator is then denoted by $\hat{\boldsymbol{\theta}}$. Thus for the null hypothesis on homoscedasticity of model (2.11), the set of parameters is given by $\boldsymbol{\theta}^T = (\mu_1, \mu_2, \sigma^2_{C(T)}(S), \sigma^2_{C(T)}(M), \sigma^2_R)$ and its likelihood function is given by (2.22), with the obvious adaption that the variance component $\sigma^2_{C(T)}$ would change with the index i. Now consider the hierarchically smaller model where the parameters are restricted or constraint in some way. Let's call this restricted set of parameters $\boldsymbol{\theta}_0$. For the null hypothesis of homoscedasticity the reduced set is given by $\boldsymbol{\theta}_0^T = (\mu_1, \mu_2, \sigma^2_{C(T)}, \sigma^2_{C(T)}, \sigma^2_R)$. Then the log likelihood function for the statistical model with this reduced set of parameters is given by $\ell(\boldsymbol{\theta}_0|\boldsymbol{y})$ and its maximum likelihood estimator is denoted by $\hat{\boldsymbol{\theta}}_0$. The likelihood ratio test statistic for the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is then given by

$$\mathsf{LRT} = -2[\ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})]. \tag{2.28}$$

If the LRT would be large, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is rejected, otherwise the null hypothesis is not rejected. To understand what is large for the LRT, it is important to know the distribution of the likelihood ratio test under the null hypothesis. Under certain regularity conditions (Searle *et al.*, 2006), the distribution of the LRT is approximately chi-square distributed with a degrees of freedom equal to the difference in the number of parameters between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, say r_0. Thus for the null hypothesis on homoscedasticity of model (2.11), the number of degrees of freedom would be equal to $r_0 = 1$. Using the chi-square distribution, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is rejected whenever $\mathsf{LRT} > \chi^{-2}_{r_0}(1-\alpha)$, with $\chi^{-1}_d(q)$ the qth quantile of a chi-square distribution with d degrees of freedom and α the level of significance.

Unfortunately, for specific restrictions on the parameters, that are related to testing a parameter at its boundary, the LRT is not approximately chi-square distributed (see Verbeke and Molenbergh, 2000), but rather distributed as a mixture of a chi-square distributed variable and a degenerated random variable taking its value at this boundary. For example, the the likelihood ratio test LRT in (2.28) for testing the null hypothesis $H_0 : \sigma^2_{C(T)} = 0$ in model (2.11), is distributed as $Z \cdot X$, with Z having a Bernoulli distribution with probability $p = 0.5$, $Z \sim \text{Ber}(0.5)$, and X having a chi-square distribution with 1 degrees of freedom, $X \sim \chi^2_1$. For this special case, the null hypothesis $H_0 : \sigma^2_{C(T)} = 0$ is rejected at significance level α whenever $\mathsf{LRT} > \chi^{-2}_{r_0}(1-2\alpha)$. Thus, we are still using the quantile of the chi-square distribution, but at twice the significance level.

Example 3. Heteroscedastiticy of the between class variability

To illustrate the the likelihood ratio test on the IQ data, we tested the null hypothesis $H_0 : \sigma^2_{C(T)}(S) = \sigma^2_{C(T)}(M)$ with maximum likelihood estimation. We used the programming codes in the SAS code box (), and added the option "GROUP = `COMBI`" to the RANDOM statement. Then the full RANDOM statement would become "RANDOM `CLASS(COMBI)`/GROUP = `COMBI`;". The parameter estimators are provided in Table 2.8. We also copied the ML estimates from Table 2.7 into Table 2.8 and added the $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ values for both models. These values can be found in the output session "**Fit Statistics**" of the **SAS** output. To obtain both $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ and $-2\ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y})$, we need to run the ANOVA model twice: once under the heteroscedastic setting and once under the homoschedastic setting.

Table 2.8: Parameter estimates of the heteroscedastic ANOVA model (2.11) and the likelihood values.

Parameter	Heteroscedasticity		Homoscedasticity	
	Estimate	CI	Estimate	CI
μ_1	11.62	[11.34; 11.90]	11.64	[11.42; 11.86]
$\mu_2 - \mu_1$	0.242	[-0.062; 0.546]	0.221	[-0.045; 0.486]
$\sigma^2_{C(T)}(S)$	0.377	[0.270; 0.564]	0.593	[0.457; 0.801]
$\sigma^2_{C(T)}(S)$	1.117	[0.756; 1.782]		
σ^2_R	3.811	[3.643; 3.990]	3.816	[3.648; 3.997]
$-2\ell(\hat{\boldsymbol{\theta}}\|\boldsymbol{y})$	16591.5		16606.3	

Calculating the likelihood ratio test statistic from the likelihood values in Table 2.8, results in $\mathsf{LRT} = 16606.3 - 16591.5 = 14.8$. For testing at a significance level of $\alpha = 0.05$, the LRT value should be compared with the 95%th quantile of the chi-square distribution with just one degrees of freedom, which is $\chi_1^{-2}(0.95) = 3.8415$. The LRT is much larger than this quantile value, thus the null hypothesis $H_0 : \sigma^2_{C(T)}(S) = \sigma^2_{C(T)}(M)$ is being rejected at the significance level of $\alpha = 0.05$. Thus, we conclude that heteroscedasticity in the variability between classes is present, with a larger variability for schools with multi-grade classes than for schools with single grade classes.

The heteroscedasticity also affects the confidence intervals slightly, in particular the confidence interval for the difference in IQ score between type of class. The confidence interval is widened with the heteroscedastic model compared to the homoscedastic model. This may not be surprising with the large difference in variability between classes for type of class.

2.8.2 Residuals

The assumption of normality in ANOVA models is implemented in different ways: both the residuals and the random effects are assumed normally distributed. To evaluate normality of all random terms in the model is not straightforward. The reason is that the random terms are unmeasured and that they have to be predicted to be able to investigate them (see the BLUPs in (2.17)). Since the predictions of the random effects are complicated functions of the model parameter estimates, the distribution of the predictions is unknown and often not normally distributed. Thus evaluation of normality of the random effects should be conducted with the care and is often ommitted.

For the residuals this may be slightly different, because we can define different residuals in random and mixed effects ANOVA models. For model (2.11), the two types of residuals are defined by

$$\begin{aligned} \text{Marginal}: &\quad \hat{e}^M_{ij} = y_{ijk} - \hat{\mu}_i, \\ \text{Conditional}: &\quad \hat{e}^C_{ij} = y_{ijk} - \hat{y}_{ij}, \end{aligned} \tag{2.29}$$

with $\hat{\mu}_i$ the fixed effects estimator of μ_i and $\hat{y}_{ij}$ the BLUP given in (2.17). The difference in the two residuals is that the marginal residuals are not corrected for the random effects but the conditional residuals are. Thus if $\hat{y}_{ij}$ would be very precise, the conditional residual is close to e_{ij}, but the marginal residual is

never close to e_{ij}. Indeed, if the estimator $\hat{\mu}_i$ is close to μ_i, then the marginal residual is close to $a_{j(i)} + e_{ij}$.

The advantage of the marginal residuals is that their distribution is more likely closer to a normal distribution. The fixed effect estimator $\hat{\mu}_i$ will be normally distributed for balanced design (if the assumptions of normality holds) and approximately normally distributed for unbalanced designs. Thus the marginal residuals will be approximately normally distributed. However, plotting the marginal residuals in a probability plot, may still demonstrate violation of normality, even if the assumption of normality holds. This has to do with the relative sizes of the variance components. If the variance component for variability between classes is large, say substantially larger than the variability within classes, then a small number of classes may provide a disjoint set of data points with hardly any overlap. The data shows a mixture of normally shaped histograms that are partly separated. The advantage of the conditional residuals is that it is closer to the true residuals and that it is not convoluted with multiple other sources of randomness. Nevertheless, since it makes use of the predicted random effects, its distribution often deviates from normality. It is common to investigate both the marginal and conditional residuals.

Example 4. Evaluation of the marginal and conditional residuals

Procedure MIXED of **SAS** can generate both residuals simply by adding the option "RESIDUAL" to the MODEL statement. The final MODEL statement would then become "MODEL `IQV = COMBI` / SOLUTION CL DDFM = SAT OUTP = PRED RESIDUAL;". We show the standard graphical output of the marginal and conditional residuals of model (2.11) in Figures 2.2 and 2.3, where we have implemented the heteroscedastic model (see Example 3).

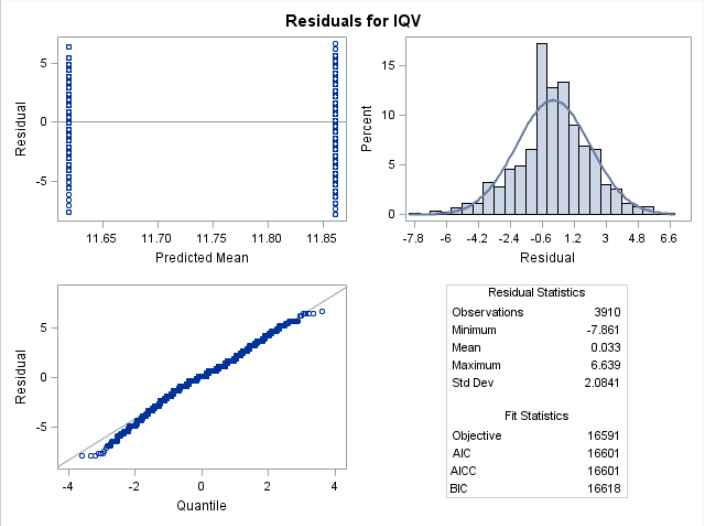


Figure 2.2: Marginal residuals of heteroscedastic ANOVA model (2.11).

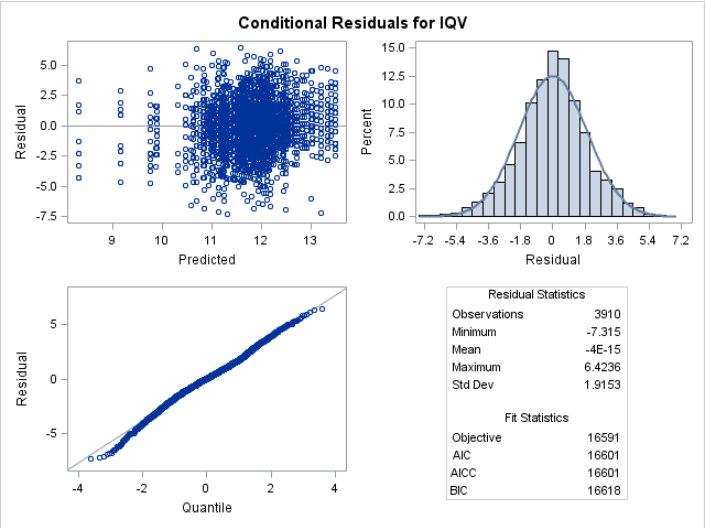


Figure 2.3: Conditional residuals of heteroscedastic ANOVA model (2.11).

When the marginal residuals are plotted against the predicted values (in this case $\hat{\mu}_i$) we obviously have only two sets of observations, since type of class has only two levels. The histogram of the marginal residuals seem somewhat different from normal, but this could be the cause of the imbalances in class sizes. The probability plot of the marginal residuals shows a distribution that is close to normal. The conditional residuals against the predicted class IQ's shows a large spherical set of observations with a few classes further away from the majority of predictions. This may suggest that there are *outlier* classes in the data. Both the histogram and the probability plot shows that the conditional residuals are close to normal. Thus, the residual analysis does not show any concerns with the assumption of normality.

2.9 Higher-order ANOVA models

In Section 2.3 we have tried to keep the ANOVA model relatively simple, but the shool data may request much more complicated ANOVA models. For instance, if we wish to investigate the progression in the language or arithmetic test scores for boys and girls separately, we have to add more factors to the two-way nested ANOVA model (2.11): the factors sex and grade. The post test score was observed in grade 8 and the pre test score in grade 7, thus the factor grade represents a factor time.

One way to look at this progression is to calculate the difference between post and pre scores (or alternatively the ratio of pre and post scores) to eliminate the factor time. We could then formulate a three-way mixed effects ANOVA model for this situation (which is studied in more detail in Exercise 4). Alternatively, we could also study an ANOVA model that analyzes the post and pre scores jointly, including a fourth factor in the ANOVA model. This four-way ANOVA model is of course more elaborate and sophisticated than the three-way model. The question is what can this more sophisticated ANOVA model bring that is

not present in the three-way model if we wish to study progression? To answer this question we will study both models in more detail.

2.9.1 A three-way ANOVA model

The three-way ANOVA model for the outcome "difference in post and pre test scores" can be formulated as follows. The numerical outcome is y_{ijkl} and it represents the difference in arithmetic score of child $l \in \{1, 2, ..., L_{ijk}\}$ in class $k \in \{1, 2, ..., K_{ij}\}$ within type of class $j \in \{1, 2\}$ and having gender $i \in \{1, 2\}$. The ANOVA model with all possible interaction terms can be formulated as

$$y_{ijkl} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + c_{k(j)} + (\beta c)_{ik(j)} + e_{ijkl}, \tag{2.30}$$

with μ an overall average progression, β_i the effect of sex, with $\beta_2 = 0$, γ_j the effect of class type, with $\gamma_2 = 0$, $(\beta\gamma)_{ij}$ the interaction effect of sex and class type, with $(\beta\gamma)_{12} = (\beta\gamma)_{21} = (\beta\gamma)_{22} = 0$, $c_{k(j)}$ the effect of class within class type, with $c_{k(j)}$ i.i.d. normally distributed $c_{k(j)} \sim N(0, \sigma^2_{C(T)})$, $(\beta c)_{ik(j)}$ the interaction effect of sex and class within class type, with $(\beta c)_{ik(j)}$ i.i.d. normally distributed $(\beta c)_{ik(j)} \sim N(0, \sigma^2_{SC(T)})$, and e_{ijkl} the residual, with e_{ijkl} i.i.d. normally distributed $e_{ijkl} \sim N(0, \sigma^2_R)$.

Model interpretations: The restrictions on the fixed effects are needed to make the three-way ANOVA model in (2.30) is identifiable, as we also discussed for the two-way nested model in (2.11). For the two fixed effects factors sex and type of class we can only estimate 4 mean parameters (since both sex and type of class have two levels). These men parameters can be referred to as μ_{ij}, and they need to be equal to $\mu_{ij} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij}$ in model (2.30). However, the right-hand side has nine parameters, thus we need to put five parameters equal to zero. For unbalanced data it is most convenient to choose just one of the fixed effects parameters of each factor separately equal to zero (e.g., $\beta_2 = 0.$ and $\gamma_2 = 0$). For the interaction effects the number of parameters that can be set freely is the product of the number of degrees of freedom for the factors involved. For the interaction effects of sex and type of class the number of parameters unequal to zero is $1 = (2 - 1) \times (2 - 1)$. We have chosen the parameter $(\beta\gamma)_{11}$ to be unequal to zero because β_1 and γ_1 were unequal to zero. Thus the parametrization we have chosen in model (2.30) is

$$\mu_{22} = \mu, \quad \mu_{12} = \mu + \beta_1, \quad \mu_{21} = \mu + \gamma_1, \mu_{11} = \mu + \beta_1 + \gamma_1 + (\beta\gamma)_{11}.$$

Model (2.30) also has an interaction effect between a fixed effects factor (sex) and a random effects factor (class within type of class). Interaction effects between factors with at least one random effects factor are always treated as random effects. That is why the interaction effects $(\beta c)_{ik(j)}$ are assumed random, since it is an interaction of the factor "sex" with the random effects factor "class within type of class". This interaction term is also nested within type of class. It represents how differences between children from different sexes vary across classes within type of class. The difference between two children from different sexes can be viewed as an effect of sex and this effect varies with anothe factor class within type of class. Indeed, the difference $y_{2jkl_2} - y_{1jkl_1}$ between a boy and girl is normally distributed with mean $\beta_2 - \beta_1 + (\beta\gamma)_{2j} - (\beta\gamma)_{1j}$ and with variance $\sigma^2_{SC(T)} + \sigma^2_R$. The mean is affected by the difference in sex and the interaction

effect of sex and class type, while the variance is affected by the interaction between sex and class within type of class. To compare this difference between boys and girls, we look at a difference between two children of the same sex $y_{ijkl_2} - y_{ijkl_1}$. The distribution of this difference is normal with mean 0 and variance σ_R^2 and does not inclued the effect of sex and the two interactions effects.

Intraclass correlation coefficients: The random interaction effects also influence the correlations between children within a class. If we calculate the correlation of the outcome between children of the same sex within one class we obtain

$$\mathsf{CORR}(y_{ijkl_1}, y_{ijkl_2}) = \frac{\sigma^2_{C(T)} + \sigma^2_{SC(T)}}{\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R}.$$

This correlation is the same for boys and girls. If we calculate the correlation in the outcome of children of different sexes in the same class we obtain

$$\mathsf{CORR}(y_{1jkl_1}, y_{2jkl_2}) = \frac{\sigma^2_{C(T)}}{\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R},$$

which is smaller than or equal to the correlation in outcome of children of the same sexes if we assume that variance components are non-negative.

This ordering in correlation coefficients would probably make sense, since we may expect that boys and girls interact within their own sexes more than they would interact between the sexes. However, if the opposite is true, we expect a higher correlation in outcome between children of opposite sexes than a correlation in outcome between children of the same sex. This would then create a negative estimate of the variance component $\sigma^2_{SC(T)}$. It may not indicate a lack of interaction ($H_0 : \sigma_{SC(T)} = 0$), but a sign that the outcome in children within a class is stronger correlated between children of opposite sexes than in children of the same sex ($H_0 : \sigma_{SC(T)} < 0$). Looking at the variance components of an ANOVA model in this way is seldom done, becasue the hypothesis of $H_0 : \sigma_{SC(T)} < 0$ can not be tested with the current model. It requires a reformulation of the ANOVA model to be able to investigate if the correlation in outcome between children of different sexes is larger than the correlation in outcome of children of the same sexes. Nevertheless, the example shows that negative variance components may actually have a particular interpretation.

Estimation for balanced data: Using the calculation rules for the expected mean squares discussed in Section 2.4.5, we obtain the following results:

$$\begin{aligned}
\mathbb{E}[MS_S] &= Q(\beta_1) + L\sigma^2_{SC(T)} + \sigma^2_R \\
\mathbb{E}[MS_T] &= Q(\gamma_1) + IL\sigma^2_{C(T)} + L\sigma^2_{SC(T)} + \sigma^2_R \\
\mathbb{E}[MS_{ST}] &= Q((\beta\gamma)_{11}) + L\sigma^2_{SC(T)} + \sigma^2_R \\
\mathbb{E}[MS_{C(T)}] &= IL\sigma^2_{C(T)} + L\sigma^2_{SC(T)} + \sigma^2_R \\
\mathbb{E}[MS_{SC(T)}] &= L\sigma^2_{SC(T)} + \sigma^2_R \\
\mathbb{E}[MS_R] &= \sigma^2_R
\end{aligned} \tag{2.31}$$

Thus the variance component estimators are now easily established: $\hat{\sigma}^2_R = MS_R$, $\hat{\sigma}^2_{SC(T)} = [MS_{SC(T)} - MS_R]/L$, and $\hat{\sigma}^2_{C(T)} = [MS_{C(T)} - MS_{SC(T)}]/[IL]$. The expected mean squares also show that hypothesis testing is straightforward using F-tests. For instance, the null hypothesis $H_0 : \beta_1 = 0$ can be tested with

$F_S = MS_S/MS_{SC(T)}$ and the null hypothesis $H_0 : \gamma_1 = 0$ can be tested with $F_T = MS_T/MS_{C(T)}$. In exercise 5 it is shown that the null hypothesis $H_0 : \gamma_1 = 0$ can not be tested with an exact F-test when the factor S is taken random, even for balanced data.[11]

Furthermore, as we know from Section 2.4, the fixed effects parameters for balanced data are estimated by using the averages.[12] With the fixed effects restrictions we imposed on model (2.30), the following estimators for the fixed effects are given by

$$\begin{array}{rcl} \hat{\mu} & = & \bar{y}_{22..}, \\ \hat{\beta}_1 & = & \bar{y}_{12..} - \bar{y}_{22..}, \\ \hat{\gamma}_1 & = & \bar{y}_{21..} - \bar{y}_{22..}, \\ (\hat{\beta\gamma})_{11} & = & \bar{y}_{11..} - \bar{y}_{12..} - \bar{y}_{12..} + \bar{y}_{22..}. \end{array} \tag{2.32}$$

The estimators in (2.32) are all unbiased and normally distributed. The estimator $\hat{\mu}$ is normally distributed with mean μ and variance $[\sigma^2_{C(T)} + \sigma^2_{SC(T)}]/K + \sigma^2_R/[LK]$. This variance can not be estimated by a single mean squares and requires a linear combination of two mean squares:

$$\hat{\mathsf{VAR}}(\hat{\mu}) = [MS_{C(T)} + (I-1)MS_{SC(T)}]/[IKL].$$

Thus to obtain an appropriate degrees of freedom for this standard error, we may resort to the Satterthwaite approach in Section 2.4.4. This is necessary to make sure that an approximate confidence interval on μ of the form in (2.18) has an appropriate coverage probability. This issue of being unable to estimate the standard error of the fixed effects estimator $\hat{\mu}$ with a single mean square, is not present in the other fixed effects estimators. The variances of $\hat{\beta}_1$ and $\hat{\gamma}_1$ are given by $\mathsf{VAR}(\hat{\beta}_1) = \mathsf{VAR}(\hat{\gamma}_1) = 2[L\sigma^2_{SC(T)} + \sigma^2_R]/[LK]$ and they can be estimated by $2MS_{SC(T)}/[KL]$. The variance of estimator $(\hat{\beta\gamma})_{11}$ is given by $\mathsf{VAR}((\hat{\beta\gamma})_{11}) = 4[L\sigma^2_{SC(T)} + \sigma^2_R]/[LK]$ and can be estimated by $4MS_{SC(T)}/[KL]$. If we would have taken alternative restrictions on the fixed effects, we would obtain different estimators and therefore also different standard errors. For balanced data, the restrictions $\beta_1 + \beta_2 + \cdots + \beta_I = 0$, $\gamma_1 + \gamma_2 + \cdots + \gamma_J = 0$, $(\beta\gamma)_{i1} + (\beta\gamma)_{i2} + \cdots + (\beta\gamma)_{iJ} = 0$, and $(\beta\gamma)_{1j} + (\beta\gamma)_{2j} + \cdots + (\beta\gamma)_{Ij} = 0$ are more natural and provides more symmetry. It is arbitrary to choose just one of the parameters for a particular level of a fixed effects factor equal to zero. The overall mean μ is then estimated by the overall average $\hat{\mu} = \bar{y}_{....}$,the effect β_i is estimated by $\hat{\beta}_i = \bar{y}_{i...} - \bar{y}_{....}$, the effect γ_j is estimated by $\hat{\gamma}_j = \bar{y}_{.j..} - \bar{y}_{....}$, and the interaction effect $(\beta\gamma)_{ij}$ is estimated by $(\hat{\beta\gamma})_{ij} = \bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....}$.[13]

[11] Clearly, when the the factor S represents sex, as we have seen in the case study for the school data, we should not treat factor S as a random effect.

[12] For the fixed effects of model (2.30) with the imposed restrcitions on the fixed effects we only need to know the averages $\bar{y}_{ij..} = \sum_{k=1}^{K}\sum_{l=1}^{L} y_{ijkl}/(KL)$.

[13] Thus for model (2.30) with these more symmetric imposed restrictions, we need to calculate multiple averages: $\bar{y}_{ij..} = \sum_{k=1}^{K}\sum_{l=1}^{L} y_{ijkl}/(KL)$, $\bar{y}_{.j..} = \sum_{i=1}^{I} \bar{y}_{ij..}$, $\bar{y}_{i...} = \sum_{j=1}^{J} \bar{y}_{ij..}$, and $\bar{y}_{....} = \sum_{i=1}^{I} \bar{y}_{i...} = \sum_{j=1}^{J} \bar{y}_{.j..}$. Note that we had to calculate these averages anyhow for the calculation of the mean squares.

These estimators are normally distributed:

$$\begin{aligned}
\hat{\mu} &\sim N(\mu, [IL\sigma^2_{C(T)} + L\sigma^2_{SC(T)} + \sigma^2_R]/[IJKL]),\\
\hat{\beta}_i &\sim N(\beta_i, (I-1)[L\sigma^2_{SC(T)} + \sigma^2_R]/[IJKL]),\\
\hat{\gamma}_j &\sim N(\gamma_1, (J-1)[IL\sigma^2_{C(T)} + L\sigma^2_{SC(T)} + \sigma^2_R]/[IJKL]),\\
(\hat{\beta\gamma})_{ij} &\sim N((\beta\gamma)_{ij}, (I-1)(J-1)[L\sigma^2_{SC(T)} + \sigma^2_R]/[IJKL]).
\end{aligned}$$

The variances of these estimators can now be estimated by a single mean square. For balanced data, it is recommended to use the more symmetric restrictions on the fixed effects, because the confidence intervals on the fixed effects parameters are now all exact, and not an approximation, under the model assumptions. Unfortunately, most software packages do not implement this recommended restriction, since it would be an inappropriate restrictions for unblanced data.

2.9.2 A four-way ANOVA model

For the four-way ANOVA model the numerical outcome is y_{hijkl} and it represents the arithmetic score of child $l \in \{1, 2, ..., L_{hijk}\}$ in class $k \in \{1, 2, ..., K_{hij}\}$ within type of class $j \in \{1, 2\}$ having gender $i \in \{1, 2\}$ and observed at grade $h \in \{1, 2\}$. The ANOVA model with all possible crossed and nested terms can now be formulated as

$$\begin{aligned}
y_{hijkl} = \mu &+ \alpha_h + \beta_i + \gamma_j + (\alpha\beta)_{hi} + (\alpha\gamma)_{hj} + (\beta\gamma)_{ij} + (\alpha\beta\gamma)_{hij}\\
&+ c_{k(j)} + d_{l(ijk)} + (\alpha c)_{hk(j)} + (\beta c)_{ik(j)} + (\alpha\beta c)_{hik(j)}\\
&+ (\alpha d)_{hl(ijk)} + e_{hijkl}
\end{aligned} \tag{2.33}$$

with α_h the effect of grade, with $\alpha_2 = 0$, $(\alpha\beta)_{hi}$ the interaction effect of grade and sex, with $(\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$, $(\alpha\gamma)_{hj}$ the interaction effect of grade and class type, with $(\alpha\gamma)_{12} = (\alpha\gamma)_{21} = (\alpha\gamma)_{22} = 0$, $(\alpha\beta\gamma)_{hij}$ the interaction effect of grade, sex, class type, with $(\alpha\beta\gamma)_{122} = (\alpha\beta\gamma)_{212} = (\alpha\beta\gamma)_{221} = (\alpha\beta\gamma)_{112} = (\alpha\beta\gamma)_{121} = (\alpha\beta\gamma)_{112} = 0$, $d_{l(ijk)}$ the effect of child within sex and class within type of class, with $d_{l(ijk)}$ i.i.d. normally distributed $d_{l(ijk)} \sim N(0, \sigma^2_{I(STC)})$, $(\alpha c)_{hk(j)}$ the interaction effect of grade and class within class type, with $(\alpha c)_{ik(j)}$ i.i.d. normally distributed $(\alpha c)_{ik(j)} \sim N(0, \sigma^2_{GC(T)})$, $(\alpha\beta c)_{hik(j)}$ the interaction effect of grade, sex and class within class type, with $(\alpha\beta c)_{hik(j)}$ i.i.d. normally distributed $(\alpha\beta c)_{ik(j)} \sim N(0, \sigma^2_{GSC(T)})$, $(\alpha d)_{hl(ijk)}$ the interaction effect of grade and child within sex and class within type of class, with $(\alpha d)_{hl(ijk)}$ i.i.d. normally distributed $(\alpha d)_{hl(ijk)} \sim N(0, \sigma^2_{GI(STC)})$, e_{hijkl} the residual, with e_{hijkl} i.i.d. normally distributed $e_{hijkl} \sim N(0, \sigma^2_R)$, and with all other terms defined by ANOVA model (2.30).

In model (2.33) we have three fixed effects factors: grade, sex, and type of class. All three factors have two levels, so we could estimate eight different means (if we include all possible interactions). If we carefully look at the restrictions we imposed on the fixed effects α_h,β_i, γ_j, $(\alpha\beta)_{hi}$, $(\alpha\gamma)_{hj}$, $(\beta\gamma)_{ij}$, and $(\alpha\beta\gamma)_{hij}$, we have selected the parameters α_1, β_1, γ_1, $(\alpha\beta)_{11}$, $(\alpha\gamma)_{11}$, $(\beta\gamma)_{11}$, and $(\alpha\beta\gamma)_{111}$ to be non-negative. This leads indeed to eight parameters if we also include the average μ. Thus model identifiability seems to be guaranteed. Note that for balanced data alternative restrictions are more appropriate.

Furthermore, it should be noted that the unit of analysis has changed from class within type of class to child (or individual) within sex and class within

type of class, since we are now analyzing repeated measures on each child. In our previous ANOVA models we had no repeats on individuals, so the residual represented variability between children. Now the residual represents variability within one child. However, the interaction term of grade and child within sex and class within type of class is *confounded* with this residual. Thus the two variance components $\sigma^2_{GI(STC)}$ and σ^2_R can not be estimated separately. Estimation of these two variance components would only be possible if we would have at least two readings of the outcome on children in the same grade (i.e., true repeats at the same moment in time and not just over time). The lack of real repeats imply that we can only estimate $\sigma^2_{GI(STC)} + \sigma^2_R$, leading to six variance components in the ANOVA model (2.33). This also implies that we will treat the term $(\alpha d)_{hl(ijk)} + e_{hijkl}$ as just one residual term $\tilde{e}_{hijkl}$ having two interpretations, but we will rename this as the residual e_{hijkl} again having variance σ^2_R.

Intraclass correlation coefficients: Thus model (2.33) has six variance components (that can be estimated), and several correlation coefficients can be calculated. Similar to model (2.30), we could calculate correlations between children within class, for children of the same sex as well as for children with a different sex. The correlation between outcomes of children with the same sex within one class observed in the same grade is given by

$$\mathsf{CORR}(y_{hijkl_1}, y_{hijkl_2}) = \frac{\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_{GC(T)} + \sigma^2_{GSC(T)}}{\sigma^2_{C(T)} + \sigma^2_{I(SCT)} + \sigma^2_{SC(T)} + \sigma^2_{GC(T)} + \sigma^2_{GSC(T)} + \sigma^2_R}.$$

For children of different sexes in the same class and grade the correlation is the same, but the variance components $\sigma^2_{SC(T)}$ and $\sigma^2_{GSC(T)}$ should be eliminated from the numerator. Alternatively, we can also calculate correlations in the outcomes obtained over time for the same children or for different children in the same class. The correlation between two outcomes in grade 7 and 8 on the same children (within children correlation) is equal to

$$\mathsf{CORR}(y_{1ijkl}, y_{2ijkl}) = \frac{\sigma^2_{C(T)} + \sigma^2_{I(SCT)} + \sigma^2_{SC(T)}}{\sigma^2_{C(T)} + \sigma^2_{I(SCT)} + \sigma^2_{SC(T)} + \sigma^2_{GC(T)} + \sigma^2_{GSC(T)} + \sigma^2_R}.$$

Here we see that the variance component for differences between children ($\sigma^2_{I(SCT)}$) is playing an important role for the calculation of the correlation.

Understanding the effect of time: Based on ANOVA model (2.33), we can now determine the difference between pre and post test scores. This difference in the post and pre score for a single child l is then equal to

$$\begin{aligned} y_{2ijkl} - y_{1ijkl} = \alpha_2 - \alpha_1 + (\alpha\beta)_{2i} - (\alpha\beta)_{1i} + (\alpha\gamma)_{2j} - (\alpha\gamma)_{1j} \\ +(\alpha\beta\gamma)_{2ij} - (\alpha\beta\gamma)_{1ij} + (\alpha c)_{2k(j)} - (\alpha c)_{1k(j)} \\ +(\alpha\beta c)_{2ik(j)} - (\alpha\beta c)_{1ik(j)} + e_{2ijkl} - e_{1ijkl} \end{aligned} \tag{2.34}$$

This difference has the exact same terms as the terms in model (2.30), except that the terms in the difference (2.34) are represented by differences in effects of model (2.33) and that they all contain the symbol α (except for the residual). Nevertheless, it should be clear that $\alpha_2 - \alpha_1$ in (2.34) will play the role of μ in (2.30), $(\alpha\beta)_{2i} - (\alpha\beta)_{1i}$ in (2.34) will play the role of β_i in (2.30), $(\alpha c)_{2k(j)} - (\alpha c)_{1k(j)}$ in (2.34) will play the role of $c_{k(j)}$ in (2.30), etc. Furthermore, the variance component $\sigma^2_{C(T)}$ in model (2.30) is identical to twice the variance

component $2\sigma^2_{GC(T)}$ in model (2.33). Thus, a random effect of class within type of class for the difference in arithmetic test scores is twice the random effects of the interaction between grade and class within type of class. Moreover, all effects in model (2.30) are formed by the interaction effects of grade with the other factors in model (2.33). Thus, if our focus is on progression, the four-way ANOVA model for the arithmetic post and pre scores does not provide more information than would describe the three-way ANOVA model on the difference on pre and post scores. The assumptions of normality and homoscedasticity in the four-way ANOVA model does not play any role in this evaluation.

2.9.3 Confidence intervals on sums of variance components

For the three-way ANOVA model in (2.30), we have three variance components: $\sigma^2_{C(T)}$, $\sigma^2_{SC(T)}$, and σ^2_R. The variance of the outcome in the children in the population is then determined by $\sigma^2_{\text{TOT}} = \sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R$.[14] For the four-way ANOVA model in (2.33) we have six variance components $\sigma^2_{C(T)}$, $\sigma^2_{SC(T)}$, $\sigma^2_{GC(T)}$, $\sigma^2_{I(SCT)}$, $\sigma^2_{GSC(T)}$, and σ^2_R and the total variability in the outcome of all children in garde seven or eight is now $\sigma^2_{\text{TOT}} = \sigma^2_{C(T)} + \sigma^2_{I(SCT)} + \sigma^2_{SC(T)} + \sigma^2_{GC(T)} + \sigma^2_{GSC(T)} + \sigma^2_R$. Thus depending on the ANOVA model, we may need to add many variance components to quantify the variability in certain (subgroups of) populations. Approximate confidence interval for sums of variance components has been discussed in literature, but here we discuss the approach of Van den Heuvel (2010). This approach works for all ANOVA models and for balanced and unbalanced data, without calculating the linear combinations of means squares. More generically, the approach of Van den heuvel (2010) uses REML estimators and is therefore also suitable for other models than ANOVA models where multiple variance components need to be added. The approach is based on the concept of Satterthwaite (1946).

Let's assume we have P variance components from an ANOVA model, say σ^2_1, σ^2_2,..., and σ^2_P, and let $\sigma^2_{\text{TOT}} = \sum_{r=1}^{P} \sigma^2_r$ be the total variance of interest. There could be other variance components present in the ANOVA model, but we are only interested in the ones we just listed. We will assume that the variance components are estimated with REML, say $\hat{\sigma}^2_r$, and that the standard error is defined by $\mathsf{VAR}(\hat{\sigma}^2_r) = \tau^2_r$. Furthermore, it is well-known that variance component estimators are not independent (see formula (2.25)), so we also define the covariance between REML estimators: $\mathsf{COV}(\hat{\sigma}^2_r, \hat{\sigma}^2_s) = \tau_{rs}$. Thus, $\tau_{rr} = \tau^2_r$. As we have discussed in Section 2.6, the standard errors can be estimated as well and they are denoted by $\hat{\tau}_{rs}$.

The REML estimator for σ^2_{TOT} is now obtained by substituting the REML estimators $\hat{\sigma}^2_r$ in σ^2_{TOT}, leading to $\hat{\sigma}^2_{\text{TOT}} = \sum_{r=1}^{P} \hat{\sigma}^2_r$. The variance of $\hat{\sigma}^2_{\text{TOT}}$ is now equal to $\mathsf{VAR}(\hat{\sigma}^2_{\text{TOT}}) = \sum_{r=1}^{P} \sum_{s=1}^{P} \tau_{rs}$. This variance can be estimated by substituting the estimators $\hat{\tau}_{rs}$ for τ_{rs} in $\mathsf{VAR}(\hat{\sigma}^2_{\text{TOT}})$. Thus, the estimated variance

[14] This variance represents the variability in the full population under the assumption that (1) the model (2.30) accurately describes the outcome and (2) the sampling of children followed a proper two-stage cluster sampling approach. This last assumption means that the schools are collected by a simple random sample of schools and that the classes within schools (of grade 7) were collected with simple random sampling. The fact that not all chidren in the collected classes may have participated, indicates that we should be careful in making statistical inference.

of $\hat{\sigma}^2_{\text{TOT}}$ is equal to $\hat{\tau}^2_{\text{TOT}} = \sum_{r=1}^{P}\sum_{s=1}^{P}\hat{\tau}_{rs}$. Assuming that $df_{\text{TOT}}\hat{\sigma}^2_{\text{TOT}}/\sigma^2_{\text{TOT}}$ is aprroximately chi-square distributed with df_{TOT} degrees of freedom results in Satterthwaite degrees of freedom: $df_{\text{TOT}} = 2[\hat{\sigma}^4_{\text{TOT}}]/\hat{\tau}^2_{\text{TOT}}$. Then, an approximate $100\%(1-\alpha)$ confidence interval on σ^2_{TOT} is obtained by

$$\left[df_{\text{TOT}}\hat{\sigma}^2_{\text{TOT}}/\chi^{-2}_{df_{\text{TOT}}}(1-\alpha/2),\ df_{\text{TOT}}\hat{\sigma}^2_{\text{TOT}}/\chi^{-2}_{df_{\text{TOT}}}(\alpha/2)\right], \tag{2.35}$$

with $\chi^{-2}_d(q)$ the qth quantile of a chi-square distribution of with d degrees of freedom. This quantile can be obtained in **SAS** in a data step where we make use of the function QUANTILE("CHISQ", q, d), see **SAS** code box below.

SAS code box: Critical Value χ^2-Distribution

```
DATA name;
  df = value;
  alpha = value;
  critical_low = QUANTILE('CHISQ', alpha/2, df);
  critical_up = QUANTILE('CHISQ', 1-alpha/2, df);
RUN;
PROC PRINT DATA = name;
RUN;
```

The **DATA** step is used to create a data set `name`. You can provide your favorite name here. The next row makes a new variable df and the value after the equality sign is the degrees of freedom df_{TOT}. Then we create a variable alpha, which we would like to be equal to the significance level α. Then we are ready to create the two critical values critical_low and critical_up of the χ^2 distribution with df_{TOT} degrees of freedom. The function in **SAS** is QUANTILE and the distribution we want to use is CHISQ. Then the input of this function is $\alpha/2$ and $1-\alpha/2$ for the lower and upper tail, respectively, and df_{TOT} degrees of freedom. Then to learn about the value, we print the data set using the **SAS** code presented here. The statement **PROC PRINT** is a procedure that prints a data set (here DATA = `name`) into **SAS** output window.

To calculate the confidence interval in (2.35), we have to program this ourselves. Most software packages do not provide the confidence intervals for functions of the variance components. For the calculation of (2.35) we need the estimated covariances $\hat{\tau}_{rs}$. These values can be obtained by an option in **SAS**. In the statement of the procedure (the first line in the programming codes in **SAS** code box 2.5.2) we just add ASYCOV (next to or before the "METHOD = " option). The term stands for asymptotic covariance. The **SAS** output then produces these covariances. If we wish to store the covariances in a **SAS** data set, we need to add a new statement to the **PROC MIXED** programming codes. This statement can be put in front of the **PROC MIXED** programming statements. If we use statement "ODS OUTPUT ASYCOV = `name`;" the covariance estimates are stored in the **SAS** data set `name`. We can also store the variance component estimates in a separate **SAS** data set with the statement "ODS OUTPUT COVPARMS = `name`;" All output of **SAS** procedures is automatically stored in separate data files. You can only access these data files after you have called them and given it your own name.

After collecting the variance component estimates and their estimated covariances, we can create a date set with all information in one row in separate columns. We can use PROC TRANSPOSE to put results in a column into a row, but often it is easier to just use a few data steps. When all info is in one row, the confidence interval in (2.35) can then be formulated by just typing the formulae in a data step (see Example 5).

2.9.4 Confidence intervals on intraclass correlation coefficients

The three-way and four-way ANOVA models in (2.30) and (2.33), respectively, show that the intraclass correlation coefficients are of the form $\mathsf{ICC} = \sigma_G^2/[\sigma_G^2 + \sigma_E^2]$, with $\sigma_G^2 = \sum_{r=1}^{Q} \sigma_r^2$, $\sigma_E^2 = \sum_{r=Q+1}^{P}$ and $P > Q$. We have assumed again that the ANOVA model contains the variance components σ_1^2, σ_2^2,..., and σ_P^2. An estimator of the ICC can be obtained by substituting the estimators $\hat{\sigma}_r^2$ for the σ_r^2 into σ_G^2 and σ_E^2. Similar to the confidence interval on sums of variance components, we approximate the distribution of these estimators $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ with a chi-square distribution:

$$\begin{array}{rcl} df_G\hat{\sigma}_G^2/\sigma_G^2 & \sim & \chi^2_{df_G} \\ df_E\hat{\sigma}_E^2/\sigma_E^2 & \sim & \chi^2_{df_E} \end{array}$$

The degrees of freedom are obtained by Satterthwaite approach:

$$df_G = 2[\hat{\sigma}_G^2/\hat{\tau}_G]^2 \quad \text{and} \quad df_E = 2[\hat{\sigma}_E^2/\hat{\tau}_E]^2,$$

with $\hat{\tau}_G^2 = \sum_{r=1}^{Q}\sum_{s=1}^{Q}\hat{\tau}_{rs}$, $\hat{\tau}_G^2 = \sum_{r=1}^{Q}\sum_{s=1}^{Q}\hat{\tau}_{rs}$, and $\hat{\tau}_{rs}$ the estimate for the covariance $\tau_{rs} = \mathsf{COV}(\hat{\sigma}_r^2, \hat{\sigma}_s^2)$ of the variance component estimators $\hat{\sigma}_r^2$ and $\hat{\sigma}_s^2$. Assuming that $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ are approximately independent, the distribution of the estimator $\hat{\mathsf{ICC}} = \hat{\sigma}_G^2/[\hat{\sigma}_G^2 + \hat{\sigma}_E^2]$, can now be approximated with the distribution of $\sigma_G^2 F/[\sigma_G^2 F + \sigma_E^2]$, where $F = [\hat{\sigma}_G^2/\sigma_G^2]/[\hat{\sigma}_E^2/\sigma_E^2]$ is approximately F-distributed with df_G and df_E degrees of freedom. Using the approximate distribution for the $\hat{\mathsf{ICC}}$, we obtain the following approximate $100\%(1-\alpha)$ confidence interval on ICC:

$$\left[\frac{\hat{\sigma}_G^2 F^{-1}_{df_G,df_E}(\alpha/2)}{\hat{\sigma}_G^2 F^{-1}_{df_G,df_E}(\alpha/2) + \hat{\sigma}_E^2}, \frac{\hat{\sigma}_G^2 F^{-1}_{df_G,df_E}(1-\alpha/2)}{\hat{\sigma}_G^2 F^{-1}_{df_G,df_E}(1-\alpha/2) + \hat{\sigma}_E^2}\right], \tag{2.36}$$

with $F^{-1}_{d_1,d_2}(q)$ the qth quantile of the F-distribution with d_1 and d_2 degrees of freedom. This quantile can be obtained in **SAS** in a data step where we make use of the function QUANTILE("F", q, d_1, d_2). The confidence interval in (2.36) has been discussed in Demtrashvili *et al.*, (2016) as one of the options for calculating confidence intervals. It should be noted that the interval in (2.36) does not reduce to the interval in (2.6) when we would study one-way random effects ANOVA, since (2.36) is based on the chi-square distribution for the variance component estimates and (2.6) is based on the chi-square distribution of the mean squares. The approach in (2.36) works best when σ_G^2 is based on several variance components.

Example 5. Confidence intervals on functions of variance components

Let's assume that we are interested in the verbal IQ (`IQV`) for all seven grade children as the outcome and we wish to study differences between sex and type

of class (multi-grade versus single grade classes). After reading in the **SAS** data set `schooldata` and renaming it to `ANALYSIS`, the following programming statements can be used to fit the three-way ANOVA model in (2.30) and storing the information on the variance component estimators:

```
ODS OUTPUT COVPARMS=VCS;
ODS OUTPUT ASYCOV=COVS;
PROC MIXED DATA=ANALYSIS METHOD=REML ASYCOV CL;
    CLASS CLASS COMBI GIRL;
    MODEL IQV = GIRL COMBI GIRL*COMBI / SOLUTION CL DDFM=SAT;
    RANDOM CLASS(COMBI) GIRL*CLASS(COMBI);
RUN;
```

The variance component estimates in the **SAS** data set `VCS` are provided under each other in one column with the name `Estimate`. The column `CovParm` indicates which variance component is represented by each row. We use three data steps to put the variance component estimates in three separate **SAS** data sets:

```
DATA VC_CT;
    SET VCS;
    WHERE COVPARM = "CLASS(COMBI)";
    VC_CT = ESTIMATE;
    KEEP VC_CT;
RUN;
DATA VC_GCT;
    SET VCS;
    WHERE COVPARM = "CLASS*GIRL(COMBI)";
    VC_GCT = ESTIMATE;
    KEEP VC_GCT;
RUN;
DATA VC_R;
    SET VCS;
    WHERE COVPARM = "Residual";
    VC_R = ESTIMATE;
    KEEP VC_R;
RUN;
```

The **SAS** data set `COVS` contains the covariance estimates of the variance component estimators. The data in `COVS` is organized as a matrix. It has a column `CovParm`, like we have seen in the **SAS** data set `VCS`, that indicates the variance component that is represented by each row. The **SAS** data `COVS` also has columns `CovP1`, `CovP2`, and `CovP3`. There are three columns, since there are three variance components to be estimated. The value in the first row of column `CovP1`, represents the estimated variance of the variance component estimator that is listed in the first row (here the value is $\hat{\tau}_1^2 = \mathsf{VAR}(\hat{\sigma}^2_{C(T)})$). The value in the third row of column `CovP2` represent the covariance of the variance component estimators that are listed in the second and third row (here the value is

$\hat{\tau}_{23} = \mathsf{COV}(\hat{\sigma}^2_{SC(T)}, \hat{\sigma}^2_R)$). To get all the variances and covariances, we will use again three data steps. The first data step collects the variances and covariances of the first row, the second data set collects the second row, and the third data set collects the third row. Since the covariance matrix is symmetric, we only need to collect the upper (or lower) part of the matrix.

```
DATA COVS_COMBI;
    SET COVS;
    WHERE COVPARM = "COMBI";
    TAU_TT = COVP1;
    TAU_TC = COVP2;
    TAU_TR = COVP3;
    KEEP TAU_TT TAU_TC TAU_TR;
RUN;
DATA COVS_CLASS;
    SET COVS;
    WHERE COVPARM = "CLASS(COMBI)";
    TAU_CC = COVP2;
    TAU_CR = COVP3;
    KEEP TAU_CC TAU_CR;
RUN;
DATA COVS_ERROR;
    SET COVS;
    WHERE COVPARM = "Residual";
    TAU_RR = COVP3;
    KEEP TAU_RR;
RUN;
```

The six created SAS data sets contains all the information necessary to calculate the total variance and the two intraclass correlation coefficients with their 95% confidence interval. Before we calculate all the results in a data step, we first have to put the six SAS data sets into one data set. The following steps provides a SAS data set with all the information we need.

```
DATA PARMS;
    MERGE VC_T VC_CT VC_R COVS_COMBI COVS_CLASS COVS_ERROR;
RUN;
DATA PARMS;
    SET PARMS;
/* TOTAL VARIANCE */;
    VC_TOT = VC_T + VC_CT + VC_R;
    SE2_TOT = TAU_TT + TAU_CC + TAU_RR + 2*TAU_TC + 2*TAU_TR + 2*TAU_CR;
    DF_TOT = 2*(VC_TOT**2)/SE2_TOT;
    LCL_TOT = DF_TOT*VC_TOT/QUANTILE('CHISQ',0.975,DF_TOT);
    UCL_TOT = DF_TOT*VC_TOT/QUANTILE('CHISQ',0.025,DF_TOT);
```

```
/* ICC WITHIN CLASS SAME SEX */;
    ICC1 = (VC_T + VC_CT)/VC_TOT;
    VC_G1 = VC_T + VC_CT;
    SE2_G1 = TAU_TT + TAU_CC + 2*TAU_TC;
    DF_G1 = 2*(VC_G1**2)/SE2_G1;
    VC_E1 = VC_R;
    SE2_E1 = TAU_RR;
    DF_E1 = 2*(VC_E1**2)/SE2_E1;
    FL_G1 = QUANTILE('F',0.025,DF_G1,DF_E1);
    FU_G1 = QUANTILE('F',0.975,DF_G1,DF_E1);
    LCL_ICC1 = VC_G1*FL_G1/(VC_G1*FL_G1+VC_E1);
    UCL_ICC1 = VC_G1*FU_G1/(VC_G1*FU_G1+VC_E1);
/* ICC WITHIN CLASS DIFFERENT SEX */;
    ICC2 = VC_T/VC_TOT;
    VC_G2 = VC_T;
    SE2_G2 = TAU_TT;
    DF_G2 = 2*(VC_G2**2)/SE2_G2;
    VC_E2 = VC_CT + VC_R;
    SE2_E2 = TAU_CC + TAU_RR + 2*TAU_TC;
    DF_E2 = 2*(VC_E2**2)/SE2_E2;
    FL_G2 = QUANTILE('F',0.025,DF_G2,DF_E2);
    FU_G2 = QUANTILE('F',0.975,DF_G2,DF_E2);
    LCL_ICC2 = VC_G2*FL_G2/(VC_G2*FL_G2+VC_E2);
    UCL_ICC2 = VC_G2*FU_G2/(VC_G2*FU_G2+VC_E2);
RUN;
```

The final **SAS** data set has many columns and only one row. Many of these columns are only supportive, used for the final calculations or at possibly for verification purposes if this would be required. These columns may not be needed in the ouput of the **SAS** output window. With the procedure PRINT only the variables of interest can be printed in the **SAS** output window. The following code prints the total variance, the two intraclass correlation coefficients, and their 95% confidence intervals.

```
PROC PRINT DATA=PARMS;
    VAR VC_TOT LCL_TOT UCL_TOT
        ICC1 LCL_ICC1 UCL_ICC1 ICC2 LCL_ICC2 UCL_ICC2;
RUN;
```

The results of the analysis are $\hat{\sigma}^2_{\text{TOT}} = 4.41[4.19; 4.66]$, $\hat{\text{ICC}}_1 = 0.139[0.106; 0.174]$, and $\hat{\text{ICC}}_2 = 0.134\ [0.101;\ 0.169]$. Thus the standard deviation in the verbal IQ among seventh grade children is approximately equal to 2.10 at an average of 11.51 [11.25; 11.77] for boys in multi-grade classes. The correlation among children of the same sex within a class is equal to 13.9% [10.6; 17.4], and the correlation between children of different sexes within an class is equal to 13.4% [10.1; 16.9].

2.10 Mixed effects models: Extending ANOVA models

ANOVA models are a subset of the much larger class of mixed effects models. Mixed effects models or just *mixed models* form a large class of statistical models that can describe data with complex correlation structures. Structures that are more elaborate than what can be described by ANOVA models. Mixed models describe how a set of random variables with certain distribution functions are functionally or mathematically related to each other. They model both fixed and random effects of a set of independent variables on a set of dependent variables. Contrary to ANOVA models, the variables do not have to be restricted to categorical variables alone. This generalization to any type of variable also extends the definitions for fixed and random effects.

A fixed effect or fixed effect parameter is a quantification of the direct influence of an independent variable on a dependent variable. For instance, a fixed effect parameter could be the slope of a linear regression analysis of for instance body mass index on blood pressure. In ANOVA models, fixed effects would be differences between means, but in mixed models any type of parameter can be chosen as a fixed effect. A random effect is an effect that can randomly vary with the levels of a categorical variable. For instance, the slope for body mass index on blood pressure can be different for each individual. For the random effects we still need a categorical variable (here individual), but again the parameter can be more than just a mean. Thus any single fixed effect parameter could potentially be changed to many random parameters that may vary with the levels of a categorical variable. In this situation, the fixed effect still exists as the mean of these individual coefficients.

This large class of mixed models can be divided into *linear mixed models*, *non-linear mixed models*, and *generalized linear mixed models*. In a linear mixed model the fixed and random effects parameters are all linearly related to the (conditional) expectation of the dependent variable, while in non-linear mixed models these relations are typically non-linear. Generalized linear mixed models exists of a linear function in the fixed and random parameters, the so-called *linear predictor*, and a link function that connects the expectation of the dependent variable with the linear predictor. Thus mixed effects models can be viewed as an extension of linear, non-linear, and logistic regression models. Indeed, traditional regression analyses only include fixed effects, while mixed effects models extent these traditional models with random effects. Including random effects in these regression analyses typically helps us modeling correlations among multiple values of the (set of) dependent variables. In the following sections we will describe the linear, non-linear, and mixed effects models.

Exercises

1. Assume the unbalanced one-way random effects model

   (a) Demonstrate that the correlation between y_{ir} and y_{is} is equal to the ICC define in (2.5).

   (b) Demonstrate that the sums of squares for the within group variability and between group variability add up to the total sums of squares: $SS_T = SS_B + SS_W$.

(c) Demonstrate that the variance of $\bar{y}_{..}$ is equal to $\sum_{i=1}^{m}[n_i^2\sigma_G^2/n^2] + \sigma_E^2/n$.

(d) Demonstrate that

i. $\mathsf{COV}(y_{ij}, \bar{y}_{i.}) = \sigma_G^2 + \sigma_E^2/n_i$

ii. $\mathsf{COV}(\bar{y}_{i.}, \bar{y}_{..}) = [n_i\sigma_G^2 + \sigma_E^2]/n$

(e) Demonstrate that the expected mean squares are given by

i. $\mathbb{E}[MS_W] = \sigma_E^2$

ii. $\mathbb{E}[MS_B] = (n - \sum_{i=1}^{m}[n_i^2/n])\sigma_G^2/(m-1) + \sigma_E^2$

2. Assume a balanced two-way crossed random effects model without interaction effects.

(a) Determine all possible the correlations between the outcomes.

(b) Determine the sums of squares and mean squares.

(c) Determine the expected mean squares.

(d) Determine the variances of the fixed effects estimators and determine the estimators of these variances.

(e) Determine the variances of the variance component estimators and determine the estimators of these variances.

(f) Determine the test statistics for the null hypotheses on the random effects.

3. Assume a balanced two-way crossed random effects model with interaction effects.

(a) Determine all possible the correlations between the outcomes

(b) Determine the sums of squares and mean squares.

(c) Determine the expected mean squares.

(d) Determine the variances of the fixed effects estimators and determine the estimators of these variances.

(e) Determine the variances of the variance component estimators and determine the estimators of these variances.

(f) Determine the test statistics for the null hypotheses on the random effects.

4. Assume a three-way balanced ANOVA model for the IQ data:

$$y_{hijk} = \mu + \alpha_h + \beta_i + (\alpha\beta)_{hi} + c_{j(i)} + (\alpha c)_{hj(i)} + e_{hijk},$$

with y_{hijk} the response of child $k \in \{1, 2, ..., K\}$ in class $j \in \{1, 2, ..., J\}$ of class type $i \in \{1, 2, ..., I\}$ having sex $h \in \{1, 2, ..., H\}$, with μ an overall average, with the fixed effects parameters satisfying $\alpha_1 + \alpha_2 + \cdots + \alpha_H = 0$, $\beta_1 + \beta_2 + \cdots + \beta_I = 0$, $(\alpha\beta)_{h1} + (\alpha\beta)_{h2} + \cdots + (\alpha\beta)_{hI} = 0$, and $(\alpha\beta)_{1i} + (\alpha\beta)_{2i} + \cdots + (\alpha\beta)_{Hi} = 0$, with $c_{j(i)}$ i.i.d. normally distributed $c_{j(i)} \sim N(0, \sigma_{C(T)}^2)$, with $(\alpha c)_{hj(i)}$ i.i.d. normally distributed $(\alpha c)_{hj(i)} \sim N(0, \sigma_{SC(T)}^2)$, and with e_{hijk} i.i.d. normally distributed $e_{hijk} \sim N(0, \sigma_R^2)$. Note that interaction terms with random effects factors are always random, thus $(\alpha c)_{hj(i)}$ is random since $c_{j(i)}$ is random.

(a) Determine the correlations between the IQ-scores for children within a class, within and between sexes.

(b) Determine the OLS estimators for the fixed effects.

(c) Determine the sums of squares and mean squares.

(d) Determine the expected mean squares.

(e) Determine the variances of the fixed effects estimators and determine the estimators of these variances.

(f) Determine the variances of the variance component estimators and determine the estimators of these variances.

(g) Determine the test statistics for the null hypotheses on the fixed effects and random effects.

(h) Fit the ANOVA model on the unbalanced data and compare the parameter estimates and the calculated test statistics when you use the following estimation techniques

i. TYPE 1: class type before sex
ii. TYPE 1: sex before class type
iii. TYPE 3
iv. ML
v. REML

(i) Investigate heteroscedasticity of the variance components for the two fixed effects factors sex and type of class.

(j) Investigate the normality of the residuals for the ANOVA models (possibly corrected for heteroscedasticity).

5. Consider the balanced ANOVA model in exercise 4, but assume now that α_h becomes a random effect a_h, with $a_h \sim N(0, \sigma_S^2)$, not representing the factor sex anymore, but something else. As mentioned in Exercise 4, interaction terms with random effects factors are always random too. Thus the term $(\alpha\beta)_{hi}$ changes to $(a\beta)_{hi}$ with $(a\beta)_{hi} \sim N(0, \sigma_{ST}^2)$ now being random and $(\alpha c)_{hj(i)}$ changes to $(ac)_{hj(i)}$ with $(ac)_{hj(i)} \sim N(0, \sigma_{SC(T)}^2)$.

(a) Determine the sums of squares and mean squares.

(b) Determine the expected mean squares.

(c) Determine the variances of the fixed effects estimators and determine the estimators of these variances.

(d) Determine the variances of the variance component estimators and determine the estimators of these variances.

(e) Determine the test statistics for the null hypotheses on the fixed effects and random effects.

6. For the school data we formulated the likelihood function in (2.22) for IQ-scores following model (2.11). This question requests that you calculate all the elements to obtain the ML estimators.

(a) Calculate the likelihood equations in (2.23).

(b) Solve the likelihood equations and provide the ML estimators in (2.24).

(c) Determine the second derivatives of the log likelihood function.

(d) Determine the Fisher matrix and the inverse Fisher information matrix in (2.25).

7. For ANOVA model (2.11) with balanced data, we also calculated the REML estimators. This question requests that you calculate them yourself.

(a) Determine the partitioning (2.26) of the full log likelihood function in (2.22) for balanced data.

(b) Select the proper REML log likelihood function and derive the likelihood equations for the variance components.

(c) Solve the likelihood equations and formulate the REML estimators in (2.27) for the variance components.

3 Linear mixed effects models

In Section 2 we have discussed many different ANOVA models. They form only a subset of the *linear mixed models*. The difference between the two sets is typically coming from the assumptions on the random effects. In ANOVA models we assume that the random effects are all independent, but in linear mixed effects models we may not alway want to assume this. To illustrate the difference, we will consider the IQ-score of children again. We may be interested in differences in IQ-scores between the two sexes and if such differences would depend on the class type (single or multi-grade classes). An ANOVA model for this setting has been proposed in Section 2.9 and in Exerise 4 of Chapter 2.

$$y_{hijk} = \mu + \alpha_h + \beta_i + (\alpha\beta)_{hi} + c_{j(i)} + (\alpha c)_{hj(i)} + e_{hijk}, \tag{3.1}$$

with y_{hijk} the IQ-score of child $k \in \{1, 2, ..., K_{hij}\}$ in class $j \in \{1, 2, ..., J_i\}$ of class type $i \in \{1, 2\}$ having sex $h \in \{1, 2\}$, and with

- ▷ μ an overall average IQ-score,
- ▷ α_h the effect of sex, with $\alpha_2 = 0$,
- ▷ β_i the effect of class type, with $\beta_2 = 0$,
- ▷ $(\alpha\beta)_{hi}$ the interaction effect of sex and class type, with $(\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$,
- ▷ $c_{j(i)}$ the effect of class within class type, with $c_{j(i)}$ i.i.d. normally distributed $c_{j(i)} \sim N(0, \sigma^2_{C(T)})$),
- ▷ $(\alpha c)_{hj(i)}$ the interaction effect of sex and class within class type, with $(\alpha c)_{hj(i)}$ i.i.d. normally distributed $(\alpha c)_{hj(i)} \sim N(0, \sigma^2_{SC(T)})$),
- ▷ e_{hijk} the residual, with e_{hijk} i.i.d. normally distributed $e_{hijk} \sim N(0, \sigma^2_R)$).

Note that the restrictions on the fixed effects are needed to make sure that the model is identifiable (i.e., not overparametrized). There are only four combinations of class type and sex, implying that we can only estimate four means μ_{hi}. With the imposed restrictions we obtain that $\mu_{11} = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$, $\mu_{12} = \mu + \alpha_1$, $\mu_{21} = \mu + \beta_1$, and $\mu_{22} = \mu$. Thus four effect parameters determine the four means. Furthermore, as we already indicated in Exercise 4, interaction terms between fixed and random effects factors are treated as random.

ANOVA model (3.1) induces two different correlation coefficients. One correlation coefficient for children within a class of the same sex and one correlation coefficient for children within a class of different sexes. These two correlations are given by

$$\begin{array}{rcl} \rho_{WS} & = & [\sigma^2_{C(T)} + \sigma^2_{SC(T)}]/[\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R], \\ \rho_{BS} & = & \sigma^2_{C(T)}/[\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R]. \end{array} \tag{3.2}$$

The sum of variance components $\sigma^2_{C(T)} + \sigma^2_{SC(T)} + \sigma^2_R$ represents the total variability in IQ-scores for children. This variability is the same for both sexes in model (3.1), but this may potentially be incorrect. The correlation in IQ-scores

between children of the same sex may be different for boys and girls due to differences in social behavior. To investigate this we could eliminate the random effect of $c_{j(i)}$ from model (3.1) and now assume that the two random interaction effects $(\alpha c)_{1j(i)}$ and $(\alpha c)_{2j(i)}$ are bivariately normally distributed with means zero and variance-covariance matrix given by

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The two correlation coefficients in (3.2) are now changed to three correlation coefficients:

$$\begin{aligned} \rho_{WS,h} &= \sigma_h^2/[\sigma_h^2 + \sigma_R^2], \\ \rho_{BS} &= \rho_{12}\sigma_1\sigma_2/\sqrt{[\sigma_1^2 + \sigma_R^2][\sigma_2^2 + \sigma_R^2]}. \end{aligned}$$

In case the two variance components σ_1^2 and σ_2^2 are different, boys and girls will have a different correlation in IQ-scores, but when they are equal, we return to the two correlation coefficients in (3.2), albeit in a different parametrization. Thus by eliminating $c_{j(i)}$ and making the random effects $(\alpha c)_{1j(i)}$ and $(\alpha c)_{2j(i)}$ bivariate normal, we have extended the ANOVA model in (3.1) to a linear mixed effects model that is not a variance component model anymore.

3.1 General formulation of linear mixed models

The linear mixed model in its generic form was presented in the landmark paper of Laird and Ware (1982). For each unit i, there exists a vector of repeated outcomes, typically collected over time. This vector is denoted by $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^T$, with n_i the number of repeated outcomes. This number may be different from unit to unit. For longitudinal data, these repeated outcomes are observed at time points $\boldsymbol{t}_i = (t_{i1}, t_{i2}, ..., t_{in_i})^T$. Each unit can have their own unique set of time points. The general linear mixed effects model is now written by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{3.3}$$

with $i \in \{1, 2, ..., m\}$ and

- ▷ $\boldsymbol{X}_i$ an observed matrix of covariates (containing continuous and/or categorical variables) for unit i,
- ▷ $\boldsymbol{\beta}$ a set of unknown fixed effects parameters (including a potential intercept) that is common across units,
- ▷ $\boldsymbol{Z}_i$ an observed matrix describing the structure of the random effects for unit i that typically contains categorical covariates or the time points of the repeated observations,
- ▷ $\boldsymbol{b}_i$ a vector of random effects for unit i to determine unit-specific parameters,
- ▷ $\boldsymbol{e}_i$ a vector of residuals for unit i.

It is assumed that $\boldsymbol{b}_i \sim N(0, \boldsymbol{G}_i)$ has a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{G}_i$, $\boldsymbol{e}_i \sim N(0, \boldsymbol{R}_i)$ has a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{R}_i$, and $\boldsymbol{b}_i$

and $\boldsymbol{e}_i$ are independent. The variance-covariance matrices $\boldsymbol{G}_i$ and $\boldsymbol{R}_i$ depend on unit i, because each unit may have its own number n_i of repeated observations. Thus, the matrices are of the same structure, but they only differ in their dimensions. Thus, when the number of repeated observations are the same across all units, we can ignore the index i of these matrices. For ANOVA models, the variance-covariance matrices are typically diagonal matrices, indicating that the random effects and the residuals are independent. Finally, for longitudinal data, the matrix X_i may contain the time points at which the repeated observations of an unit is collected. Thus this matrix may differ from unit to unit due to its dimension, but also due to the values of the covariates. The vector of unknown parameters $\boldsymbol{\beta}$ will be the same for each unit.

Based on the independence assumption of the random effects and the residuals, we can now easily determine the moments of the vector $\boldsymbol{y}_i$. The first two moments are determined by

$$\begin{array}{rcl} \mathbb{E}(\boldsymbol{y}_i) & = & \mathbb{E}[\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i] = \boldsymbol{X}_i\boldsymbol{\beta} \\ \mathsf{VAR}(\boldsymbol{y}_i) & = & \mathbb{E}[\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i][\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i]^T \\ & = & \mathbb{E}[\boldsymbol{Z}_i\boldsymbol{b}_i\boldsymbol{b}_i^T\boldsymbol{Z}_i^T] + \mathbb{E}[\boldsymbol{e}_i\boldsymbol{e}_i^T] \\ & = & \boldsymbol{Z}_i\boldsymbol{G}_i\boldsymbol{Z}_i^T + \boldsymbol{R}_i \end{array}$$

Note that we only made use of the independence assumption and that these moments would remain true even if we do not assume normality of the random effects and residuals. Now also including the assumption of normality of the random effects and residuals, we may conclude that the vector of repeated observations has a multivariate normal distribution. We know that sums of normally distributed random variables remain normally distributed. Thus, $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$ has a multivariate distribution with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{V}_i \equiv \boldsymbol{Z}_i\boldsymbol{G}_i\boldsymbol{Z}_i^T + \boldsymbol{R}_i$. It should be noted that the matrix $\boldsymbol{V}_i$ may still depend on the unit i even if for each unit the same number of information is collected. The reason is that the matrix $\boldsymbol{Z}_i$ may contain unit-specific information, like the time points at which the repeated observations have been collected.

In the linear mixed effects models it is assumed that the data from different units are independent. Thus, the final linear mixed effects model for all the outcome data $\boldsymbol{y} \equiv (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_m)^T$ can now be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e},$$

with $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{b}$, and $\boldsymbol{e}$ defined by

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \vdots \\ \boldsymbol{Z}_m \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_m \end{pmatrix}.$$

The procedure **MIXED** in **SAS** (link documentation) has been especially developed to be able to fit the linear mixed models to longitudinal or repeated data. It has the following structure:

```
PROC MIXED DATA = NAME OPTIONS;
  CLASS FACTORS;
```

```
  MODEL OUTCOME = FACTORS COVARIATES
    / SOLUTION CL DDFM = SAT RESIDUAL OPTIONS;
  RANDOM FACTORS COVARIATES / SUBJECT = UNIT OPTIONS;
  REPEATED FACTORS / SUBJECT = UNIT OPTIONS;
  ESTIMATE 'LABEL' FIXED EFFECTS | RANDOM EFFECTS;
  LSMEANS FACTORS / OPTIONS;
RUN;
```

In the first line of the programming statements we need to tell on what data the model is applied to. The OPTIONS are multiple as we already has seen in the fit of ANOVA models. It contains for instance the method of estimation and options to obtain confidence intervals and standard errors on the covariance parameters. The second line is necessary to tell the procedure which variables are categorical. We call these variables factors. Then the third line is the MODEL statement. This statement is aligned with the matrix $\boldsymbol{X}_i$ in the linear mixed model. Thus, the MODEL statement focuses on the fixed effects parameters $\boldsymbol{\beta}$. We typically include SOLUTION and CL to make sure we obtain estimates of the fixed effects parameters. Although there are many choices for the degrees of freedom, we mostly choose Satterthwaite degrees of freedom for the analysis. This is provided by DDFM = SAT as we already know from the ANOVA models. The RANDOM statement focuses on the random effects $\boldsymbol{b}_i$ and it is used to describe the $\boldsymbol{Z}_i$ matrix in the linear mixed model (3.1). The SUBJECT statement is needed to inform the procedure what the unit i in the analysis is. Thus the variable `UNIT` should be part of the CLASS statement. There are many OPTIONS available. One of these options is the choice of the form of matrix $\boldsymbol{G}_i$ using the statement TYPE = . Another option, is to use GROUP = `FACTOR` to calculate the variance-covariance matrix $\boldsymbol{G}_i$ for each level of the variable `FACTOR`. The REPEATED statement is included to make choices on the variance-covariance matrix $\boldsymbol{R}_i$ of the residuals. For longitudinal data the `FACTOR` is typically the variable time, since the residuals are related to the noise at each time point. The OPTIONS are similar to the OPTIONS in the random statement. The ESTIMATE statement can be used to calculate the so called contrasts. For instance, if we wish to calculate a difference between the first and the average of the second and third time point, we could use the ESTIMATE statement. Calculation of such differences are called *contrasts* as we already have seen for the ANOVA models. Contrasts can be calculated for both fixed and random effects. Finally, the LSMEANS statement can be used calculate standard differences between levels of the factors. The OPTIONS contain approaches to deal with multiple testing.

3.2 Subject-specific linear mixed models

3.3 Marginal linear mixed models

4 Non-linear mixed models

To illustrate the differences, consider the Mitscherlich function $\mathbb{E}(y|x, \boldsymbol{\beta}) = \beta_1 + \beta_2 \exp\{\beta_3 x\}$, with y the dependent variable, x the independent variable, and $\boldsymbol{\beta} =$

$(\beta_1, \beta_2, \beta_3)^T$ unknown parameters (either fixed or random). The Mitscherlich function is a non-linear function in the parameters (Box and Lucas, 1959), while the Mitscherlich function becomes linear in the parameters when $\gamma = 1$ occurs and it becomes generalized linear when $\alpha = 0$ occurs:

$$\begin{aligned} \beta_3 = 1: &\quad \mathbb{E}(y|x, \boldsymbol{\beta}) = \beta_1 + \beta_2 \exp\{x\}, \\ \beta_1 = 0: &\quad \log(\mathbb{E}(y|x, \boldsymbol{\beta})) = \log(\beta_2) + \beta_3 x, \end{aligned}$$

Within the class of linear mixed models, there exists a smaller class of linear mixed models, which are called *ANOVA models* or *variance component models* (Searle *et al.*, 2006). They form a smaller class, because they impose an independence restriction on the random effects parameters (Khuri and Sahai, 1985; Searle *et al.*, 2006; Van den Heuvel, 2010).

5 Generalized linear mixed models

References

1. Asar Ö, Bolin D, Diggle PJ, Wallin J. Linear mixed effects models for non-Gaussian continuous repeated measurement data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2020; **69**(5):1015-1065.

2. Bonett DG, Wright TA. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, 2000; 65(1):23-8.

3. Box GE, Lucas HL. Design of experiments in non-linear situations. *Biometrika*, 1959, **46**(1/2):77-90.

4. Brandsma HP, Knuver JWM. Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, 1989; **13**(7):777 - 788.

5. Crump SL. The estimation of variance components in analysis of variance. *Biometrics Bulletin*, 1946; **2**(1):7-11.

6. Demetrashvili N, Wit EC, van den Heuvel ER. Confidence intervals for intraclass correlation coefficients in variance components models. *Statistical methods in medical research*, 2016; **25**(5):2359-76.

7. Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, 1986; 1:67-82.

8. Efron B, Morris C. Stein's paradox in statistics. *Scientific American*, 1977; **236**(5):119-27.

9. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. *Longitudinal data analysis*. CRC press, 2008.

10. Ganguli M. A note on nested sampling. *Sankhyā: The Indian Journal of Statistics*, 1941; **5**(4):449-452.

11. Geraci M, Farcomeni A. A family of linear mixed-effects models using the generalized Laplace distribution. *Statistical methods in medical research*, 2020; **29**(9):2665-2682.

12. Graybill FA. An Introduction to Linear Statistical Models. McGraw-Hill, New York, 1961.

13. Gurka MJ. Selecting the best linear mixed model under REML. *The American Statistician*, 2006; **60**(1):19-26.

14. Henderson CR. Estimation of variance and covariance components. *Biometrics*, 1953; **9**(2):226-252.

15. Kaptein MC, Van den Heuvel ER. *Statistics for Data Scientists: An introduction to probability, statistics, and data analysis*. Springer, 2021.

16. Khuri AI, Sahai H. Variance components analysis: a selective literature survey. *International Statistical Review*, 1985; **53**(3):279-300.

17. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*, 1982; **38**(4):963-974.

18. Levy PS, Lemeshow S. *Sampling of populations: methods and applications*. John Wiley & Sons; 2013.

19. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986; **73**(1):13-22.

20. Littell RC. Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics*. 2002; **7**(4):472-490.

21. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*, 1986; **73**(1):13-22.

22. Littell RC. Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics*. 2002; **7**(4):472-490.

23. McCulloch CE, Searle SR, Generalized, Linear, and Mixed Models, Wiley, 2001.

24. Nijhuis MB, Van den Heuvel ER. Closed-form confidence intervals on measures of precision for an interlaboratory study. *Journal of Biopharmaceutical Statistics*, 2007; **17**(1):123-42.

25. Nooraee N, Molenberghs G, van den Heuvel ER. GEE for longitudinal ordinal data: comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN. *Computational Statistics & Data Analysis*, 2014; **77**:70-83.

26. Regis M, Brini A, Nooraee N, Haakma R, van den Heuvel ER. The t linear mixed model: model formulation, identifiability and estimation. *Communications in Statistics-Simulation and Computation*, 2019; **23**:1-25.

27. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 1946; **2**(6):110-114.

28. Scheffé H. The analysis of variance, John Wiley & Sons, 1959.

29. Searle SR. Linear Models. John Wiley & Sons, New York, 1971.

30. Searle SR. An overview of variance component estimation. *Metrika*, 1995; **42**(1):215-230.

31. Searle SR, Casella G, McCulloch CE, Variance Components, John Wiley & Sons, 2006.

32. Sheynin O. On the History of the Principle of Least Squares. *Archive of History of Exact Sciences*, 1993; **46**(1):39-54.

33. Snijders TAB, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage Publications, 1999.

34. Swallow WH, Monahan JF. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, 1984; **26**(1):47-57.

35. Van den Heuvel ER. A comparison of estimation methods on the coverage probability of Satterthwaite confidence intervals for assay precision with unbalanced data. *Communications in Statistics—Simulation and Computation*, 2010; **39**(4):777-794.

36. Van den Heuvel ER, IJzerman-Boon PC. A comparison of test statistics for the recovery of rapid growth-based enumeration tests. *Pharmaceutical statistics*, 2013; **12**(5):291-299.

37. Verbeke G, Molenberghs G. *Linear Mixed Models For Longitudinal Data*, New York: Springer-Verlag, 2000.

38. Verbyla AP. A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics*, 2019; **61**(1):39-50.

39. Wald A. A note on the analysis of variance with unequal class frequencies. *The Annals of Mathematical Statistics*, 1940; **11**(1):96-100.

40. Zwanenburg RJ, Bocca G, Ruiter SA, Dillingh JH, Flapper BC, van den Heuvel ER, van Ravenswaaij-Arts C. Is there an effect of intranasal insulin on development and behaviour in Phelan-McDermid syndrome? A randomized, double-blind, placebo-controlled trial. *European Journal of Human Genetics*, 2016; **24**(12):1696-701.