2AMS10: Longitudinal Data Analysis 2021-2022
Analysis of variance

September 9, 2022

Edwin van den Heuvel
e.r.v.d.heuvel@tue.nl
Professor in Statistics

Mathematics & Computer Science

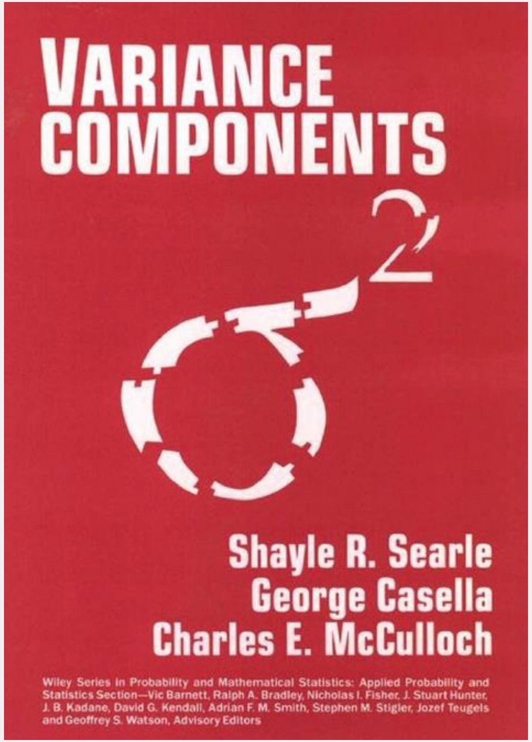
Analysis of Variance

Contents

- Introduction
- One-way ANOVA
  - Random effects model
  - Analysis of variance table
  - Parameter estimators
  - Confidence intervals balanced data
  - Confidence intervals unbalanced data
  - Predictions of random effects
  - Residuals
  - Likelihood-based estimation
  - Choice of estimation
- Two-way ANOVA models
  - Nested mixed effects model
  - Nested random effects model
  - Crossed mixed effects model
- Three-way ANOVA models
  - Expected mean squares
  - Moment-based estimators
  - Hypothesis testing
  - Unbalanced data

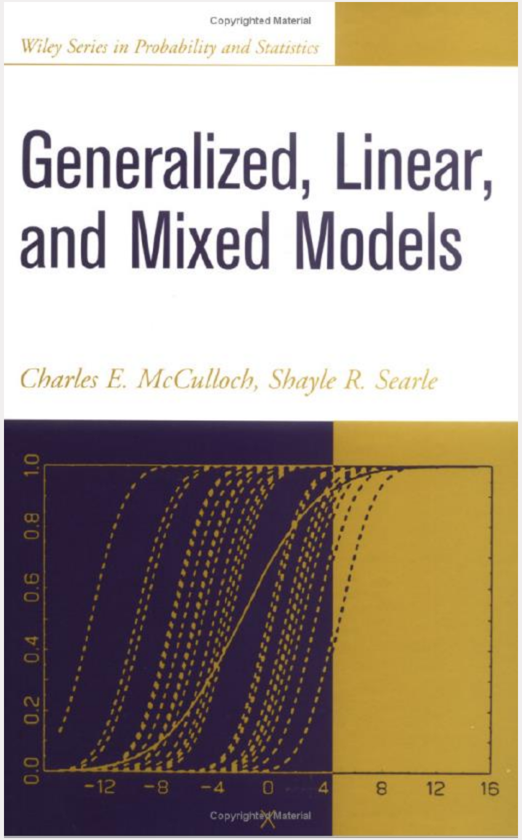
Analysis of Variance

Books for reference



- Topics are very similar with current presentation
  - Some difference in opinion
  - Book give more math detail
  - I added CI for ICC's

- Treats more than just ANOVA
  - Non-normal outcomes
  - Longitudinal data
  - Non-linear models



Introduction

Background and terminology

Statistics: Is concerned with the variability that is evident in any body of data.

- A traditional (but not obsolete) method for quantifying variability is the **analysis of variance** (ANOVA), developed by Sir Ronald Fisher
- Ronald Fisher wrote to George Snedecor in 1934:

  "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."

- Extensions of ANOVA are (linear, non-linear, and generalized linear) **mixed models** (also referred to as **multi-level models**)
  - ANOVA has limitations in describing complex correlations in longitudinal data
- The development of ANOVA and its extensions, which is still ongoing, have taken at least 80 years.

Introduction

Background and terminology

Applications of ANOVA are numerous:

- Agriculture experiments:..... maximizing crop yield and minimizing pesticides (Block designs, Latin squares, split-plot designs)
- Measurement reliability:..... quantifying noise in outcomes that are induced by the measurement system (medical diagnostics, bioassays, psychometry)
- Clinical trials:.........................quantifying effects of treatments (parallel group, stepped-wedge, cross-over designs)
- Quality improvement:..........minimizing variability between products within product specifications
- Heritability research:............quantifying how much variability in phenotypes is explained by genetics/generation

Introduction

Background and terminology

Data structure for ANOVA:

- The outcome is a continuous variable (blood pressure, crop yield, strength)
- It is typically observed under different conditions – also called **factors**
- **Factors:** are categorical variables (sex, field plots, type of product) and they have different **levels**

	Marital Status					
	Married			Not Married		
	Drug			Drug		
Sex	A	B	C	A	B	C
Male						
Female						

Extensions of ANOVA will

- Allow for numerical variables, like we use in linear regression
- Increase complexity for correlation structures present in longitudinal data

Introduction

Background and terminology

- Factors may influence the single outcome, referred to as an **effect**
  - Blood pressure is typically higher in men
  - Blood pressure varies with individuals
- The influence of a factor can be modeled as **fixed** or **random** effect
- Fixed effects: the levels of a factor enter the model as parameter
  - A fixed mean parameter for sex
- Random effects: the levels of a factor enter the model as a random variable
  - A random parameter for individuals
- Choosing between fixed or random effects is not always obvious
- Commonly used rules are
  - When all possible factor levels for the factor are included → fixed
  - When the included levels of a factor are a random draw of levels → random
  - When interest is in quantifying the mean differences between factor levels **(contrasts)** → fixed
  - When the interest is in quantifying the variability between factor levels **(variance components)** → random

Introduction

Background and terminology

<u>**Questions on fixed and random effects**</u>

- In a clinical trial on blood pressure reduction (hypertension treatment) the factor treatment is considered?
  A. Fixed
  B. Random
- In a clinical trial with longitudinal data, the factor individuals is considered?
  A. Fixed
  B. Random
- In a clinical trial with patients from different hospitals, the factor hospital is considered?
  A. Fixed
  B. Random

Introduction

Background and terminology

Crossed factors [notation AB]:

- A factor A is crossed with factor B, when the levels of factor A are identical across levels of factor B
- Example: Treatment (factor A) is crossed with hospitals (factor B)
  - The levels of treatment (e.g., new vs. control) are the same for each hospital
- The effect of a crossed factor A may change with levels of factor B $\rightarrow$ **Interaction effect**
  - Treatment effect can vary by hospital

Nested factors [notation $A(B)$]:

- A factor A is nested within factor B, when the levels of factor A within one level of factor B are different from the levels of factor A in another level of B
- Example: Children (factor A) are nested within schools (factor B)
  - The first child in school 1 is a different child from the first child in school 2
- There can be many levels of hierarchy
  - Children(Schools), Schools(Cities), Cities(Regions), Regions(Country)

Introduction

Background and terminology

<u>Goals of ANOVA:</u> Partition the total variability in observed data into variability by each crossed and nested factor and its possible interactions, including a part for the variability that is unexplained by the factors – the so-called **residual**

- Estimate effect sizes for factors (contrasts)
- Test if factors do or do not influence the outcome (hypothesis testing)
- Quantify variability between factors (variance components)
- Quantify the contribution of factors to the total variability (correlation analysis)

- ANOVA started with only fixed effects analysis
  - The two-sample t-test (with equal variances) is the simplest form of an ANOVA
- ANOVA was extended to also include
  - Random effects to address correlations
  - Combinations of random and fixed effects – called **mixed effects** analysis

TU/e

Introduction

Case study on primary school children

Study design information:

- 4106 children from primary schools (~ 11 years)
- 216 randomly selected schools (1 class/school)
- Classes can be multi-grade or single grade
- Measured both in grade 7 and in grade 8

Collected information:

- Language and arithmetic tests in grades 7 and 8
- Verbal and performal IQ (obtained in grade 7)
- Sex of child and minority status (born outside industrialized countries)
- Social economic status of child’s family
- Class size (not all children participated)

Research questions:

- Understanding cognitive development over time
  - How much is it changed?
  - Differences between certain subgroups (sex, minority)?
- How much are language and arithmetic scores affected by intelligence?
- What is the variation between children and schools/classes?

Introduction

Case study on primary school children

	CLASS	CHILD	COMBI	SIZE	SSES	GIRL	MINORITY	SITTERS	CSES	IQV	IQP	PRE_LANG	POST_LANG	PRE_ARITH	POST_ARITH	DIFF_A
1	180	1	0	29	11	1	1	0	10	10.5	7.333333333	33	.	10	12	2
2	180	2	0	29	11	1	1	0	.	14	14.33333333	44	50	18	30	12
3	180	3	0	29	11	0	0	0	23	15	12.329999924	36	46	14	24	10
4	180	4	0	29	11	0	1	0	10	14.5	10	36	45	12	19	7
5	180	5	0	29	11	0	0	0	15	9.5	11	33	33	10	24	14
6	180	6	0	29	11	0	0	0	23	11	10	29	46	13	26	13
7	180	7	0	29	11	0	0	0	10	8	6.6659998894	19	20	8	9	1
8	180	8	0	29	11	0	1	0	10	9.5	9	22	30	8	13	5
9	180	9	0	29	11	0	1	0	23	9.5	10.329999924	20	30	7	13	6
10	180	10	0	29	11	0	0	0	10	13	14.329999924	44	57	17	30	13
11	180	11	0	29	11	0	1	1	13	9.5	8.6660003662	34	36	10	23	13
12	180	12	0	29	11	0	1	0	15	11	15	31	36	14	22	8
13	180	13	0	29	11	1	1	0	10	5.5	9	18	29	11	19	8
14	180	14	0	29	11	0	0	0	18	14	9.3330001831	36	40	10	23	13
15	180	15	0	29	11	0	1	0	15	9	9	31	41	10	18	8
16	180	16	0	29	11	0	1	0	20	10.5	13	34	47	12	22	10
17	180	17	0	29	11	0	1	0	10	10	8.6660003662	31	33	9	15	6
18	180	18	0	29	11	0	0	0	20	11	12.329999924	32	37	13	21	8
19	180	19	0	29	11	1	0	0	13	6.5	7.3330001831	23	29	9	13	4
20	180	20	0	29	11	1	1	1	10	4	9.3330001831	20	26	9	12	3
21	180	21	0	29	11	1	1	0	10	11	6.6659998894	27	37	7	11	4
22	180	22	0	29	11	1	1	0	15	11	11	31	40	16	23	7
23	180	23	0	29	11	1	0	1	10	11	11	24	27	16	17	1
24	180	24	0	29	11	1	0	0	13	12.5	10	35	43	9	16	7
25	180	25	0	29	11	1	1	0	10	11.5	11	32	39	14	24	10
26	180	26	0	29	11	1	1	0	10	10	6.6659998894	25	21	12	16	4
27	180	27	0	29	11	1	1	0	10	10	9.3330001831	33	42	17	11	-6
28	280	28	1	19	11	1	0	1	15	11.5	7.3330001831	22	21	3	6	3
29	280	29	1	19	11	0	0	0	15	10.5	9.3330001831	33	27	4	8	4
30	280	30	1	19	11	0	0	0	20	8	9.3330001831	22	16	11	9	-2
31	280	31	1	19	11	1	0	1	15	7.5	7.6659998894	25	31	9	9	0
32	280	32	1	19	11	0	1	1	10	7.5	11.329999924	27	21	11	18	7

Introduction

Exercise A1

<u>**Questions on the factors involved in the school data**</u>

- Mention all factors involved in the case study data
- Identify if these factors are nested or crossed
- Determine all possible interactions between factors
- Determine which factors should be treated as fixed and random

One-way ANOVA

Random effects model

Statistical model: Outcome y_{ij} of unit $j \in \{1,2,\dots,n_i\}$ in group $i \in \{1,2,\dots,m\}$

$$y_{ij} = \mu + a_i + e_{ij}$$

- With μ the mean value for the full population of all units
- With a_i the random effect of group i on the outcome, $a_i \sim N(0,\sigma_G^2)$
- With e_{ij} the residual (or error) for unit j in group i, $e_{ij} \sim N(0,\sigma_E^2)$
- With all random terms independent
- ANOVA is called **balanced** when sample sizes are equal: $n_1 = n_2 = \cdots = n_m$

- The variances σ_G^2 and σ_E^2 are referred to as **variance components (VC's)**
  - Between groups: σ_G^2
  - Within groups: σ_E^2
- Normality assumption is not essential for estimation, but is essential for
  - Hypothesis testing
  - Calculation of confidence intervals
- Extension to mixed models
  - Relaxing independence assumptions
  - Implementing **heteroscedasticity**: variance components depend on factors

One-way ANOVA

Random effects model

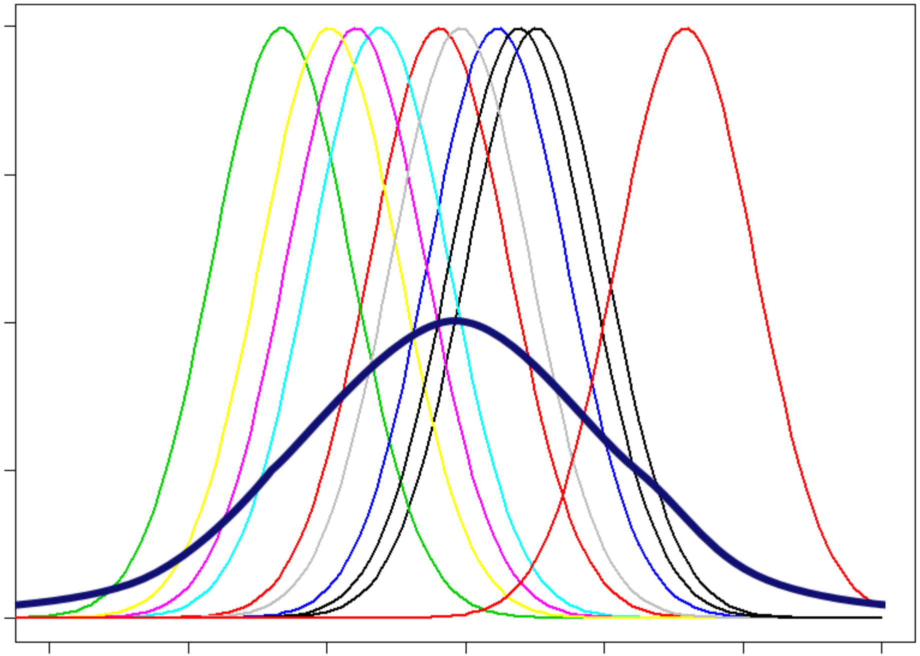
Example case study:

- A single outcome can be
  - An IQ score in grade 7
  - A language test score in one grade
  - An arithmetic test score in one grade
  - Difference in (language or arithmetic) test scores between grades
- Group would be class/school
  - School is a random draw from all schools in the Netherlands
  - Restricted to one level of multi-grade
  - No other factors in the case study is likely to be treated as random effect
- Units would be the children nested within class for a subgroup
  - All children in the class
  - All boys or all girls in the class
  - All minority or all non-minority children
  - Combinations (male minorities)
- Parameter interpretations:
  - Parameter μ: represents the mean in the Netherlands since schools are random
  - Parameter σ_G^2: represents variability in outcome between schools
  - Parameter σ_E^2: represents variability in outcome between children

One-way ANOVA
Random effects model

Visualization of statistical model:

- Each colored normal density would represent the outcomes of children in a single class
  - Is a theoretical representation, since classes will be typically finite
- Different colors would represent different classes
  - That the densities are not on top of each other indicates class differences
  - The normally distributed group effects induce an overall normal density



- In real data the normal densities would be replaced by histograms
- Thus, one-way ANOVA is a **model** for describing the data

One-way ANOVA

Random effects model

Example case study:

- Consider the difference in arithmetic test scores between the two grades
- Questions of interest:
  - What is the average change in score?
  - Are there class differences? And if so, how large are the class differences?
  - Are children different? And if so, how much are children different?
  - How large is the variation in change in test score in the Netherlands?
  - Do children in a class change similar? If so, how much are they alike?

SAS codes:

```
PROC MIXED DATA=ANALYSIS
   METHOD=TYPE3 CL COVTEST;
   CLASS CLASS;
   MODEL DIFF_A = /SOLUTION
   CL DDFM=SAT;
   RANDOM CLASS;
RUN;
```

- We put the data in a data set `ANALYSIS`, which requires a few data steps
- Will not explain all detail of the code here, but built it up slowly

One-way ANOVA

Random effects model

<u>Example case study:</u> (Outcome `diff_A`): Overview of all results

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
CLASS	196	27710	141.375199	Var(Residual) + 18.269 Var(CLASS)	MS(Residual)	3405	6.99	<.0001
Residual	3405	68826	20.213085	Var(Residual)	.	.	.	.

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CLASS	6.6320	0.7883	8.41	<.0001	0.05	5.0870	8.1770
Residual	20.2131	0.4899	41.26	<.0001	0.05	19.2863	21.2087

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	7.4608	0.2007	200	37.17	<.0001	0.05	7.0650	7.8566

<u>Additional information:</u>

- Intraclass correlation coefficient
  $I\hat{C}C = 0.247\ [0.207; 0.295]$
- Total variability
  $\hat{\sigma}_T^2 = 26.85\ [24.15; 28.72]$
- Differences between classes statistically significant

One-way ANOVA

Random effects model

Characterizations of statistical model:

- The outcome is normally distributed,
  $$y_{ij} \sim N(\mu, \sigma_G^2 + \sigma_E^2)$$
  - With $\sigma_{\text{TOT}}^2 \equiv \sigma_G^2 + \sigma_E^2$ the total variance
- Outcomes within one group are being correlated:
  $$\text{CORR}(y_{ir}, y_{is}) = \frac{\sigma_G^2}{\sigma_G^2+\sigma_E^2}, r \neq s$$
  - Which is referred to as the **Intraclass correlation coefficient (ICC)**
- Outcomes from different groups are uncorrelated
  $$\text{CORR}(y_{hr}, y_{is}) = 0, h \neq i$$

Intraclass correlation coefficient:

- Units within groups are more alike than units between groups
  - **Heterogeneity:** quantifies inconsistencies in outcomes across groups
  - **Heritability:** quantifies the genetic similarity between living beings
  - **Measurement reliability:** quantifies the agreement in repeated observations

Poor:	*ICC ≤ 0.40*
Fair:	*0.40 < ICC ≤ 0.60*
Good:	*0.60 < ICC ≤ 0.75*
Excellent:	*0.75 < ICC*

TU/e

One-way ANOVA

Analysis of variance table

- Most of the information is summarized in the **analysis of variance table**

Factor	df	SS	MS	EMS	Error Term	Error df	F-value	P-value
Between								
Residual								

- Columns up to EMS summarize variability in data
  - SS reports total variability for each factor/term with its own degrees of freedom (df)
  - $MS = SS/df$ reports the average variability
  - EMS reports what MS is estimating (expected)

df = degrees of freedom
SS = sums of squares
MS = mean squares
EMS = expected mean squares

- Columns after EMS provides hypothesis testing on each of the factor/terms
  - Test statistic is an (exact or approximate) F-test
  - The denominator in this F-test is the Error Term having an error degrees of freedom

One-way ANOVA

Analysis of variance table

Total sums of squares:

- The total variability in outcomes

$$SS_T = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

  - With n the **total sample size** defined by

$$n = \sum_{i=1}^{m} n_i$$

  - With $\bar{y}_{..}$ the **overall average** defined by

$$\bar{y}_{..} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ij}/n] = \sum_{i=1}^{m} [n_i/n] \bar{y}_{i.}$$

  - With $\bar{y}_{i.}$ the **group average** for group i defined by

$$\bar{y}_{i.} = \sum_{j=1}^{n_i} [y_{ij}/n_i]$$

- SS_T is partitioned in two parts

Within group sums of squares:

- Variability within group i:

$$SS_{W,i} = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

- Total variability within groups

$$SS_W = \sum_{i=1}^{m} SS_{W,i} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

Between group sums of squares

$$SS_B = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{m} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

- Partitioning of sums of squares

$$SS_T = SS_B + SS_W$$

One-way ANOVA

Analysis of variance table

Distributional considerations:

- Are needed to understand what the sums of squares are estimating
- The averages $\bar{y}_{i.}$ and $\bar{y}_{..}$ are normally distributed

$$\bar{y}_{i.} \sim N(\mu, \sigma_G^2 + \sigma_E^2/n_i)$$
$$\bar{y}_{..} \sim N(\mu, \textstyle\sum_{i=1}^m n_i^2 \sigma_G^2/n^2 + \sigma_E^2/n)$$

Distributional considerations:

- For balanced data ($n_i = n_0, \forall i$), the variance of $\bar{y}_{..}$ reduces to

$$\mathrm{VAR}(\bar{y}_{..}) = \sigma_G^2/m + \sigma_E^2/mn_0$$

- Covariances

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

Expected values of sums of squares

$$\mathbb{E}[SS_W] = \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \mathbb{E}(e_{ij} - \bar{e}_{i.})^2 = (n-m)\sigma_E^2$$
$$\mathbb{E}[SS_B] = \textstyle\sum_{i=1}^m \sum_{j=1}^{n_i} \mathbb{E}(a_i - \sum_{i=1}^m [n_i a_i/n] + \bar{e}_{i.} - \bar{e}_{..})^2$$
$$= [n - \textstyle\sum_{i=1}^m n_i^2/n]\sigma_G^2 + (m-1)\sigma_E^2$$

One-way ANOVA

Analysis of variance table

Degrees of freedom:

- For the variation within a single group
  - Estimator $S_i^2 = SS_{W,i}/(n_i - 1)$
  - Degrees of freedom $n_i - 1$
- For total variation within groups
  - Adding the degrees of freedom per group

    $$df_W = \sum_{i=1}^{m}(n_i - 1) = n - m$$
- The sums of squares between groups
  - Related to the calculation of a variance on the group averages
  - Degrees of freedom is number of groups minus one: $df_B = m - 1$

Mean Squares:

- Dividing the sums of squares by its degrees of freedom

$$MS_W = SS_W/df_W$$

$$MS_B = SS_B/df_B$$

Expected mean squares:

- Use of expected sums of squares

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

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

  - For balanced data ($n_i = n_0, \forall i$), the constant becomes equal to n_0

TU/e

One-way ANOVA

Analysis of variance table

Unbalanced data:

- Summarizing the variability and what it estimates

Factor	df	SS	MS	EMS
Between	$m-1$	$\sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	SS_B/df_B	$\frac{n-\sum_{i=1}^{m} n_i^2/n}{m-1}\sigma_G^2 + \sigma_E^2$
Residual	$n-m$	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	SS_W/df_W	σ_E^2

Balanced data: $(n_i = n_0, \forall i)$

- Summarizing the variability and what it estimates

Factor	df	SS	MS	EMS
Between	$m-1$	$n_0\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2$	SS_B/df_B	$n_0\sigma_G^2 + \sigma_E^2$
Residual	$m(n_0-1)$	$\sum_{i=1}^{m}\sum_{j=1}^{n_0}(y_{ij} - \bar{y}_{i.})^2$	SS_W/df_W	σ_E^2

One-way ANOVA

Exercise A2

Theoretical questions:

- Demonstrate that the correlation between y_{ir} and y_{is} is equal to the ICC
- Demonstrate that the sums of squares for the within group variability and between group variability add up to the total sums of squares
- Demonstrate that the variance of $\bar{y}_{..}$ is equal to $\sum_{i=1}^{m} n_i^2 \sigma_G^2 / n^2 + \sigma_E^2 / n$
- Demonstrate that
  - $\mathrm{COV}(y_{ij}, \bar{y}_{i.}) = \sigma_G^2 + \sigma_E^2 / n_i$
  - $\mathrm{COV}(\bar{y}_{i.}, \bar{y}_{..}) = [n_i \sigma_G^2 + \sigma_E^2]/n$
- Demonstrate that the expected mean squares are given by
  - $\mathbb{E}[MS_W] = \sigma_E^2$
  - $\mathbb{E}[MS_B] = \frac{n - \sum_{i=1}^{m} n_i^2 / n}{m-1} \sigma_G^2 + \sigma_E^2$

One-way ANOVA

Analysis of variance table

SAS procedure MIXED:

```
PROC MIXED
  DATA=name METHOD=TYPE3;
  CLASS group;
  MODEL outcome= /SOLUTION CL;
  RANDOM group;
RUN;
```

- Class statement: list categorical variables
- Model statement: list fixed effects terms (here only an intercept – thus no factors)
- Random statement: list random factors or parameters (here group)

Example case study:

- Outcome: difference in pre and post arithmetic test score (`diff_A`)
  - Group structure: School/class (`CLASS`)
  - Units: all children in the class

Type 3 Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	Expected Mean Square
CLASS	196	27710	141.375199	Var(Residual) + 18.269 Var(CLASS)
Residual	3405	68826	20.213085	Var(Residual)

- 197 classes with pre and post scores
- Average group size (# children per class) is 18.284 $= [3405 + 197]/197$
- Slightly higher than 18.269 in table

TU/e

One-way ANOVA

Analysis of variance table

Testing for a group effect:

- Null hypothesis: $H_0: \sigma_G^2 = 0$
  - The observations y_{ij} can be viewed as independent and identically distributed $y_{ij} \sim N(\mu, \sigma_E^2)$
  - The mean squares MS_W and MS_B both estimate σ_E^2 and should thus be similar
- Natural test statistic is an ***F*-test**

  $$F = MS_B / MS_W$$

  - Is F always F-distributed with df_B and df_W degrees of freedom under H_0?
  - No, but it is always approximately F

Within group mean squares:

- The residual mean squares is always chi-square distributed

  $$df_W MS_W / \sigma_E^2 \sim \chi^2_{df_W}$$

Between group sums of squares:

- For balanced data, the between group mean square is chi-square

  $$df_B MS_B / [n_0 \sigma_G^2 + \sigma_E^2] \sim \chi^2_{df_B}$$

  - MS_W and MS_B are independent
- For unbalanced data, MS_B is
  - Not chi-square distributed
  - But still independent of MS_W

One-way ANOVA

Analysis of variance table

Example case study: (Outcome `diff_A`)

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
CLASS	196	27710	141.375199	Var(Residual) + 18.269 Var(CLASS)	MS(Residual)	3405	6.99	<.0001
Residual	3405	68826	20.213085	Var(Residual)	.	.	.	.

- Thus, the null hypothesis $H_0: \sigma_G^2 = 0$ is rejected at significance level $\alpha = 0.05$.
- The test statistic $F = 6.99$ is large enough to be not concerned with the approximation of an F-distribution (also shown by extremely low P-value)
- The residual variance σ_E^2 is estimated at $\hat{\sigma}_E^2 = 20.21$
  - We use a hat to indicate the estimator for the parameter involved: $\boldsymbol{\hat{\sigma}_E^2 = MS_W}$
- The variance σ_G^2 can be estimated by $\hat{\sigma}_G^2 = 6.63 = [MS_B - MS_W]/18.269$
  - General estimator for σ_G^2 is $\boldsymbol{\hat{\sigma}_G^2 = (m-1)[MS_B - MS_W]/[n - \sum_{i=1}^{m} n_i^2/n\,]}$

One-way ANOVA

Parameter estimators

Variance component estimators:

- See previous page:

$$\hat{\sigma}_E^2 = MS_W$$
$$\hat{\sigma}_G^2 = [MS_B - MS_W]/C_n$$
$$C_n = [n - \textstyle\sum_{i=1}^{m} n_i^2/n\,]/(m-1)$$
$$= \bar{n} - \frac{1}{n(m-1)} \textstyle\sum_{i=1}^{m} (n_i - \bar{n})^2$$

  - With $\bar{n}$ the average group size
  - With $C_n = n_0$ for balanced data
- Estimation of intercept μ can be performed in two ways, using either
  - Ordinary least squares (OLS)
  - Generalized least squares (GLS)

Ordinary least squares:

- Minimize the least squares in μ

$$OLS = \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$
$$= (\boldsymbol{y} - \boldsymbol{\mu})^T (\boldsymbol{y} - \boldsymbol{\mu})$$

  - With $\boldsymbol{y}$ a vector with all observations
  - With $\boldsymbol{\mu}$ a vector of length n with μ
- The OLS solution is

$$\hat{\mu}_{\mathrm{OLS}} = \textstyle\sum_{i=1}^{m} \sum_{j=1}^{n_i} [y_{ij}/n]$$

  - Take derivative of OLS with respect to μ
  - Set derivative equal to zero
  - Solve equation for μ

One-way ANOVA

Parameter estimators

Generalized least squares:

- Minimize the weighted least squares

$$GLS = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$$

  - With $\boldsymbol{V}$ the variance-covariance matrix of the vector $\boldsymbol{y}$

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}_m \end{pmatrix}, \quad \boldsymbol{V}_i = \begin{pmatrix} \sigma_G^2 + \sigma_E^2 & \sigma_G^2 & \sigma_G^2 \\ \sigma_G^2 & \ddots & \sigma_G^2 \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_E^2 \end{pmatrix}$$

- The GLS solution depends on σ_G^2 and σ_E^2:

$$\mu_{\text{GLS}} = \frac{\sum_{i=1}^{m} [\bar{y}_{i.}/(\sigma_G^2 + \sigma_E^2/n_i)]}{\sum_{i=1}^{m} [1/(\sigma_G^2 + \sigma_E^2/n_i)]}$$

- Thus, $\hat{\mu}_{\text{GLS}}$ is obtained by substituting the variance component estimators in μ_{GLS}

The GLS estimator

Is unequal to the OLS estimator, except when

- The between variance component estimator is zero ($\hat{\sigma}_G^2 = 0$)
- When the data is balanced ($n_i = n_0, \forall i$)

Procedure MIXED of SAS uses the GLS estimator

One-way ANOVA

Parameter estimators

Example case study: (Outcome `diff_A`)

- Variance component estimators

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CLASS	6.6320	0.7883	8.41	<.0001	0.05	5.0870	8.1770
Residual	20.2131	0.4899	41.26	<.0001	0.05	19.2863	21.2087

$\hat{\sigma}_G^2$
$\hat{\sigma}_E^2$

Confidence Intervals:
Are based on different concepts
- Asymptotic inference
- Exact distributions

- Fixed effects estimators

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	7.4608	0.2007	196	37.17	<.0001	0.05	7.0650	7.8567

$\hat{\mu}_{\mathrm{GLS}}$

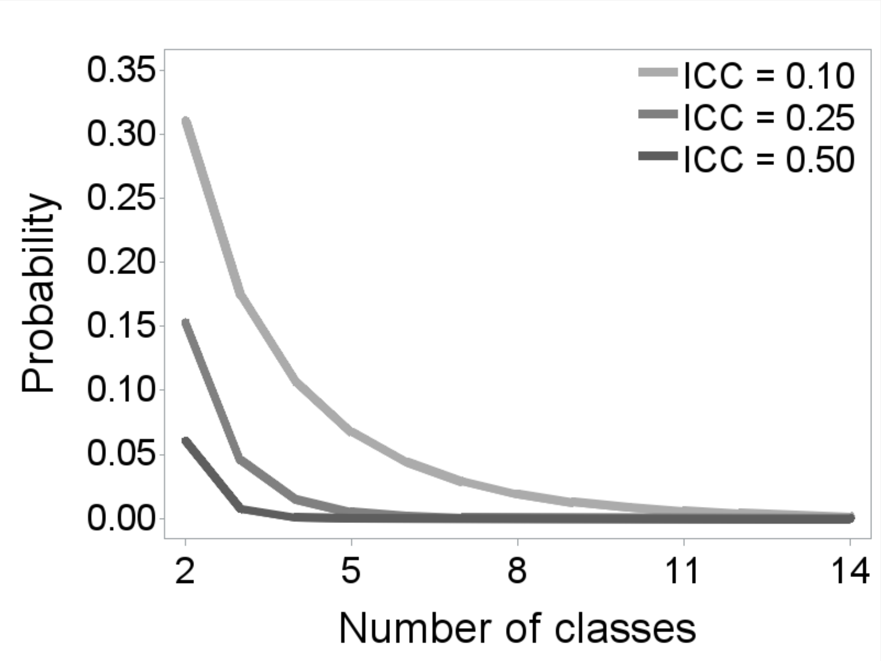
The arithmetic average of all observations is $\hat{\mu}_{\mathrm{OLS}} = 7.673$

One-way ANOVA

Parameter estimators

Variance components estimators:

- The variance component estimator can become negative
  $$P(\hat{\sigma}_G^2 \leq 0) = P\left(\frac{MS_B}{MS_W} \leq 1\right)$$
  $$\approx F_{df_B, df_W}(\sigma_E^2 / [C_n \sigma_G^2 + \sigma_E^2])$$
  - With $F_{d_1,d_2}(x)$ the F-distribution with d_1 and d_2 degrees of freedom
  - The probability is only exact in case of balanced data
  - The probability depends on the sample sizes and the ratio σ_G^2 / σ_E^2 or the ICC
- Probability can be reasonably large



One-way ANOVA

Confidence intervals for balanced data

Overall mean μ:

- OLS solution is normally distributed

$$\mathbb{E}(\hat{\mu}_{\text{OLS}}) = \mathbb{E}\left(\frac{1}{n_0 m}\sum_{i=1}^{m}\sum_{j=1}^{n_0} y_{ij}\right) = \mu$$

$$\text{VAR}(\hat{\mu}_{\text{OLS}}) = [n_0\sigma_G^2 + \sigma_E^2]/[n_0 m]$$

  - Estimator of the variance: $MS_B/[n_0 m]$
  - The OLS estimator and the variance estimator are independent
- Thus $\sqrt{n_0 m}[\hat{\mu}_{\text{OLS}} - \mu]/\sqrt{MS_B}$ has a t-distribution, which leads to

$$\hat{\mu}_{\text{OLS}} \pm t_{m-1}^{-1}(1-\alpha/2)\sqrt{MS_B/[n_0 m]}$$

  - With $t_d^{-1}(q)$ the qth quantile of the t-distribution with d degrees of freedom

Within group variance σ_E^2:

- Based on the chi-square distribution

$$UCL = df_W MS_W/\chi_{df_W}^{-2}(1-\alpha/2)$$

$$LCL = df_W MS_W/\chi_{df_W}^{-2}(\alpha/2)$$

  - $\chi_d^{-2}(q)$ the qth quantile of the chi-square distribution with d degrees of freedom
- The confidence interval on the residual variance works for unbalanced data

Both confidence intervals are

- Implemented in PROC MIXED of SAS
- Exact under the conditions of the ANOVA model (no approximations)

One-way ANOVA

Confidence intervals for balanced data

Between group variance σ_G^2:

- No exact confidence interval exists
- SAS uses an asymptotic approach

$$\hat{\sigma}_G^2 \pm z_{1-\alpha/2} V(\hat{\sigma}_G^2)$$

  - With $\hat{\sigma}_G^2 = [MS_B - MS_W]/n_0$
  - With z_q the qth quantile of the normal distribution
  - With $V(\hat{\sigma}_G^2)$ the estimated standard error of estimator $\hat{\sigma}_G^2$

$$\mathrm{VAR}(\hat{\sigma}_G^2) = \frac{2[n_0\sigma_G^2+\sigma_E^2]^2}{n_0^2\,(m-1)} + \frac{2\sigma_E^4}{n_0^2 m(n_0-1)}$$

$$V^2(\hat{\sigma}_G^2) = \frac{2MS_B^2}{n_0^2\,(m-1)} + \frac{2MS_W^2}{n_0^2 m(n_0-1)}$$

Chi-square approximation:

- Works for positive estimates only

$$UCL = df_G\hat{\sigma}_G^2/\chi_{df_G}^{-2}(1-\alpha/2)$$

$$LCL = df_G\hat{\sigma}_G^2/\chi_{df_G}^{-2}(\alpha/2)$$

  - $\chi_d^{-2}(q)$ the qth quantile of the chi-square distribution with d degrees of freedom
  - $df_G = 2[\hat{\sigma}_G^2/V(\hat{\sigma}_G^2)]^2$
- Approach is based on Satterthwaite's work in 1946 (see next slide)
  - It is not implemented in SAS for the ANOVA estimators, but uses this Satterthwaite approach for likelihood-based estimator

One-way ANOVA

Confidence intervals for balanced data

Satterthwaite degrees of freedom:

- Assume we have the following:
  - T_n an unbiased estimator for the parameter θ, i.e., $\mathbb{E}[T_n] = \theta$
  - τ_n^2 the variance of T_n: $\mathrm{VAR}(T_n) = \tau_n^2$
  - $\hat{\tau}_n^2$ an estimator for τ_n^2
- Approximating T_n with a chi-square distribution: $df_{T_n} T_n/\theta \sim \chi^2_{df_{T_n}}$
- The degrees of freedom must satisfy

$$\mathbb{E}[df_{T_n} T_n/\theta] = df_{T_n}$$
$$\mathrm{VAR}\left(df_{T_n} T_n/\theta\right) = 2df_{T_n}$$

- This results in $df_{T_n} = 2T_n^2/\hat{\tau}_n^2$

Examples Satterthwaite:

- Applied to VC's (previous slide on $\hat{\sigma}_G^2$)
- Applied to linear combinations of mean squares:
  - $T_n = \omega_B MS_B + \omega_W MS_W$
  - $\theta \equiv \mathbb{E}[T_n] = \omega_B \mathbb{E}[MS_B] + \omega_W \mathbb{E}[MS_W]$
  - $\tau_n^2 \equiv \mathrm{VAR}(T_n)$

$$= \omega_B^2 \mathrm{VAR}(MS_B) + \omega_W^2 \mathrm{VAR}(MS_W)$$
$$= 2\omega_B^2 \frac{(\mathbb{E}[MS_B])^2}{df_B} + 2\omega_W^2 \frac{(\mathbb{E}[MS_W])^2}{df_W}$$

  - Since MS_W and MS_B are independent and chi-square distributed

$$df_{T_n} = \frac{[\omega_B MS_B + \omega_W MS_W]^2}{\omega_B^2 MS_B^2/df_B + \omega_W^2 MS_W^2/df_W}$$

One-way ANOVA

Confidence intervals for balanced data

Intraclass correlation coefficient:

- Estimator of ICC

$$I\hat{C}C = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2+\hat{\sigma}_E^2} = \frac{F-1}{F+n_0-1}$$

  - With $F = MS_B/MS_W$
- F is F-distributed when multiplied by

$$\frac{\sigma_E^2}{n_0\sigma_G^2+\sigma_E^2} = \frac{ICC^{-1}-1}{ICC^{-1}+n_0-1}$$

- Thus, exact confidence interval is

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

  - With F_L and F_U the lower and upper $\alpha/2$ quantiles of the F distribution with df_B and df_W degrees of freedom

Sum of variance components:

- Total variation in outcome:

$$\sigma_T^2 = \sigma_G^2 + \sigma_E^2$$

- Estimator:

$$\hat{\sigma}_T^2 = \hat{\sigma}_G^2 + \hat{\sigma}_E^2$$
$$= [MS_B + (n_0 - 1)MS_W]/n_0$$

- Estimated standard error defined by

$$V^2(\hat{\sigma}_T^2) = \frac{2MS_B^2}{n_0^2\,(m-1)} + \frac{2(n_0-1)^2MS_W^2}{n_0^2m(n_0-1)}$$

- Approximate confidence interval

$$UCL = df_T\hat{\sigma}_T^2/\chi_{df_T}^{-2}(1-\alpha/2)$$
$$LCL = df_T\hat{\sigma}_T^2/\chi_{df_T}^{-2}(\alpha/2)$$

  - With $df_T = 2[\hat{\sigma}_T^2/V(\hat{\sigma}_T^2)]^2$ (Satterthwaite)

One-way ANOVA

Confidence intervals for unbalanced data

Model parameters:

- The GLS estimator for μ is approximately normally distributed

$$\hat{\mu}_{\mathrm{GLS}} \pm t^{-1}_{df_{\mathrm{GLS}}}(1-\alpha/2)V(\hat{\mu}_{\mathrm{GLS}})$$

  - With $V(\hat{\mu}_{\mathrm{GLS}})$ the estimated standard error of the GLS estimator $\hat{\mu}_{\mathrm{GLS}}$
  - With df_{GLS} the degrees of freedom of the standard error $V(\hat{\mu}_{\mathrm{GLS}})$
  - SAS has different options for df_{GLS}
- Confidence intervals for variance components are the same as for balanced data

Degrees of freedom for fixed effects:

- Choices for procedure MIXED:
  - RESIDUAL – only suitable for special cases [gives degrees of freedom from residuals]
  - CONTAIN – default for random effects [Finds the error term like in ANOVA tables]
  - BETWITHIN – default for repeated values [separates between and within variance and uses degrees of freedom for within]
  - SATTERTHWAITE – recommended [approximates with chi-square distribution]
  - KENWARDROGER – recommended [inflates the standard error and then uses Satterthwaite degrees of freedom]

One-way ANOVA

Confidence intervals for unbalanced data

Example case study: (Outcome `diff_A`)

- Fixed effects estimators

```
MODEL outcome= /SOLUTION CL DDFM = Option;
```

Solution for Fixed Effects									
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	
Intercept	7.4608	0.2007	3601	37.17	<.0001	0.05	7.0673	7.8543	RESIDUAL
Intercept	7.4608	0.2007	196	37.17	<.0001	0.05	7.0650	7.8567	CONTAIN
Intercept	7.4608	0.2007	3601	37.17	<.0001	0.05	7.0673	7.8543	BETWITHIN
Intercept	7.4608	0.2007	200	37.17	<.0001	0.05	7.0650	7.8566	SATTERTHWAITE
Intercept	7.4608	0.2007	200	37.17	<.0001	0.05	7.0650	7.8566	KENWARDROGERS

There are not always differences between options, it depends on the ANOVA model

One-way ANOVA

Confidence intervals for unbalanced data

Intraclass correlation coefficient:

- Estimator of ICC is now

$$\hat{ICC} = \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \hat{\sigma}_E^2} = \frac{F-1}{F+C_n-1}$$

  - With $F = MS_B/MS_W$ approximately F-distributed when multiplied with

$$\frac{\sigma_E^2}{C_n\sigma_G^2 + \sigma_E^2} = \frac{ICC^{-1}-1}{ICC^{-1}+C_n-1}$$

- Approximate confidence interval is

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

  - With F_L and F_U the lower and upper $\alpha/2$ quantiles of the F distribution with df_B and df_W degrees of freedom

Example case study: (`diff_A`)

- Manual calculation (95% confidence):
  - $MS_W = 20.213085$
  - $MS_B = 141.375199$
  - $C_n = 18.269$
  - $df_W = n - m = 3405$
  - $df_B = m - 1 = 196$
  - $F_L = 0.80766$
  - $F_U = 1.21496$
  - $F = MS_B/MS_W = 6.99424$
  - $\hat{ICC} = [F-1]/[F+C_n-1] = 0.24705$
  - $LCL = 0.20658$
  - $UCL = 0.29539$

Collect this information from ANOVA table

```
QUANTILE('F',q,
DFB,DFW)
```

One-way ANOVA

Confidence intervals for unbalanced data

Sum of variance components:

- Total variation in outcome:

$$\sigma_T^2 = \sigma_G^2 + \sigma_E^2$$

- Estimator is now:

$$\hat{\sigma}_T^2 = \hat{\sigma}_G^2 + \hat{\sigma}_E^2$$
$$= [MS_B + (C_n - 1)MS_W]/C_n$$

- Estimated standard error defined by

$$V^2(\hat{\sigma}_T^2) = \frac{2MS_B^2}{C_n^2\,(m-1)} + \frac{2(C_n-1)^2 MS_W^2}{C_n^2(n-m)}$$

- Approximate confidence interval

$$LCL = df_T\hat{\sigma}_T^2/\chi_{df_T}^{-2}(1-\alpha/2)$$
$$UCL = df_T\hat{\sigma}_T^2/\chi_{df_T}^{-2}(\alpha/2)$$

  - With $df_T = 2[\hat{\sigma}_T^2/V(\hat{\sigma}_T^2)]^2$ (Satterthwaite)

Example case study: (`diff_A`)

- Manual calculation (95% confidence):
  - $MS_W = 20.213085$ $\quad df_W = 3405$
  - $MS_B = 141.375199$ $\quad df_B = 196$
  - $C_n = 18.269$

ANOVA table again

  - $\hat{\sigma}_T^2 = 26.8452$
  - $V^2(\hat{\sigma}_T^2) = 0.82550$
  - $df_T = 1746.01$
  - $\chi_{df_T}^{-2}(0.975) = 1863.71$
  - $\chi_{df_T}^{-2}(0.025) = 1632.10$
  - $LCL = 25.1498$
  - $UCL = 28.7189$

```
QUANTILE(
'CHISQ',
q,DF_T)
```

One-way ANOVA

Confidence intervals for unbalanced data

Example case study: (Outcome `diff_A`): Overview of all results

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
CLASS	196	27710	141.375199	Var(Residual) + 18.269 Var(CLASS)	MS(Residual)	3405	6.99	<.0001
Residual	3405	68826	20.213085	Var(Residual)	.	.	.	.

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CLASS	6.6320	0.7883	8.41	<.0001	0.05	5.0870	8.1770
Residual	20.2131	0.4899	41.26	<.0001	0.05	19.2863	21.2087

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	7.4608	0.2007	200	37.17	<.0001	0.05	7.0650	7.8566

Additional information:

- Intraclass correlation coefficient
  $I\hat{C}C = 0.247\ [0.207; 0.295]$
- Total variability
  $\hat{\sigma}_T^2 = 26.85\ [24.15; 28.72]$
- Differences between classes statistically significant

One-way ANOVA

Exercise A3

Data analytic questions:

- Perform a one-way ANOVA on the difference in language test scores for all children with class as a random effect
- Fit the ANOVA model using the ANOVA estimators
  - Estimate the model parameters
  - Obtain the confidence intervals
  - Calculate the intraclass correlation coefficient and a 95% confidence interval
  - Calculate the total variability and a 95% confidence interval

One-way ANOVA

Predictions of random effects

The average progression in a class:

- It is tempting to estimate the average progression in a single class with $\bar{y}_{i.}$
- This would be the OLS estimator for a fixed effects ANOVA model

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

  - With α_i the effect of class and $\mu_i = \mu + \alpha_i$ the class mean
- However, we assumed a random effect $a_i \sim N(0, \sigma_G^2)$ for class effects
  - With a_i an unobserved random variable
  - How to determine $\mu + a_i$ given data $\bar{y}_{i.}$?

- We normally estimate a parameter and not a random variable
- Joint distribution of a_i and $\bar{y}_{i.}$

$$\begin{pmatrix} \mu + a_i \\ \bar{y}_{i.} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_G^2 & \sigma_G^2 \\ \sigma_G^2 & \sigma_G^2 + \sigma_E^2/n_i \end{pmatrix}\right)$$

- **Best linear unbiased predictor (BLUP)**

$$\mathbb{E}(\mu + a_i | \bar{y}_{i.}) = \mu + \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2/n_i}(\bar{y}_{i.} - \mu)$$

- Estimation of BLUP

$$\hat{y}_i = \hat{\mu}_{\text{GLS}} + \frac{\hat{\sigma}_G^2}{\hat{\sigma}_G^2 + \hat{\sigma}_E^2/n_i}(\bar{y}_{i.} - \hat{\mu}_{\text{GLS}})$$

  - The BLUPs are closer to $\hat{\mu}_{\text{GLS}}$ than the averages $\bar{y}_{i.}$

TU/e

One-way ANOVA

Predictions of random effects

Example case study: (`diff_A`)

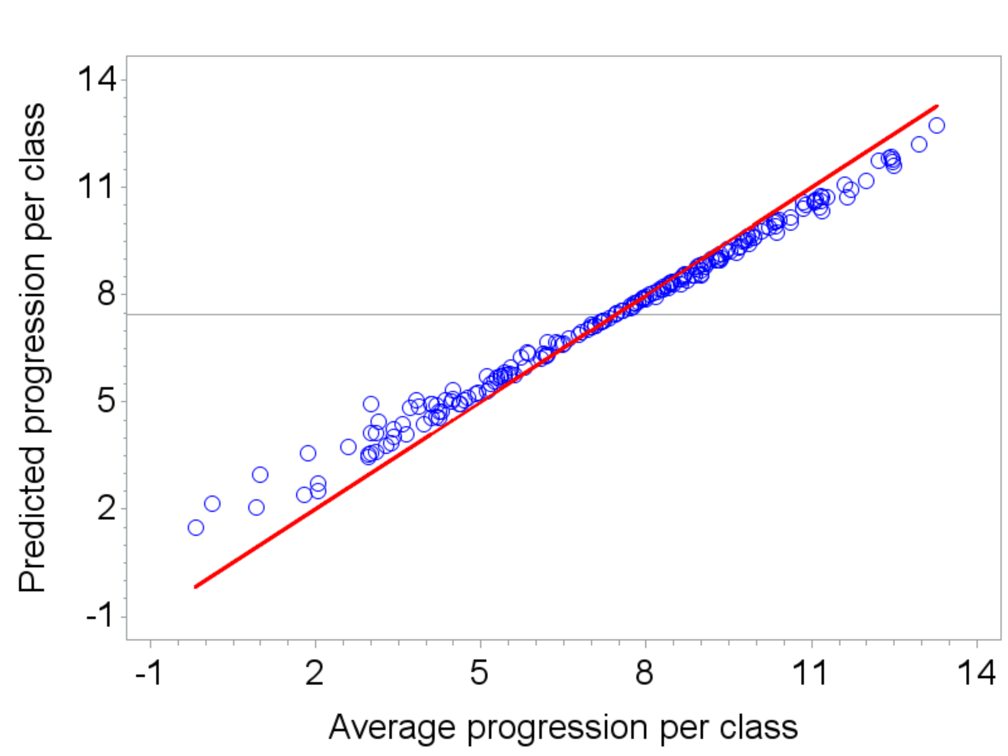
- The class progression using the BLUPs can be obtained with option `OUTP = name` in the model statement
  - `name` is the name of the data set of your choice that contains the information
  - It contains a BLUP for every record, but for children within class they are constant
  - The variable name for the BLUPs is `Pred`
- The average BLUPs: $\hat{\mu}_{\mathrm{GLS}} = 7.4608$
  - $\mathbb{E}[\mathbb{E}(\mu + a_i | \bar{y}_{i.})] = \mu$
  - $5.6365 = \mathrm{VAR}\big(\mathbb{E}(\mu + a_i | \bar{y}_{i.})\big) < \mathrm{VAR}(a_i) = 6.6320$

Order	Lowest		Highest	
	$\bar{y}_{i.}$	$BLUP_i$	$\bar{y}_{i.}$	$BLUP_i$
1	-0.182	1.476	13.27	12.73
2	0.933	2.036	12.95	12.22
3	0.125	2.149	12.45	11.85
4	1.800	2.415	12.41	11.81
5	2.032	2.519	12.47	11.78
6	2.047	2.734	12.22	11.74
7	1.000	2.960	12.50	11.71
8	2.958	3.466	12.00	11.60
9	2.950	3.547	11.59	11.19
10	1.857	3.557	11.71	11.09

- Most extreme cases
  - BLUPs are closer to $\hat{\mu}_{\mathrm{GLS}}$
  - Ranking with BLUPs and averages can change due to sample size

One-way ANOVA

Predictions of random effects



- The horizontal line represents the estimated mean $\hat{\mu}_{\mathrm{GLS}} = 7.4608$
- The red line represents $y = x$
- The blue circles represent results for classes (not children)
- BLUPs are closer to the mean
- When sample sizes within class are large, the BLUPs and averages will be almost identical

One-way ANOVA

Predictions of random effects

<u>**SAS Codes:**</u>

- Collecting the random effects

```
PROC MIXED DATA=ANALYSIS
     METHOD=TYPE3 CL COVTEST;
     CLASS CLASS;
     MODEL DIFF_A = /SOLUTION
     CL DDFM=SAT OUTP=PRED;
     RANDOM CLASS;
RUN;
```

- The data set `PRED` contains variable `PRED` on an individual level and not on class level

- Creating a data set at class level

```
PROC MEANS DATA=PRED NOPRINT;
     BY CLASS;
     VAR DIFF_A PRED;
     OUTPUT OUT=BLUPS
     MEAN=DIFF_A PRED;
     WHERE DIFF_A NE .;
RUN;
PROC SORT DATA=BLUPS;
     BY DIFF_A;
RUN;
```

One-way ANOVA

Predictions of random effects

```
GOPTIONS HTEXT=2.5;
SYMBOL1 I=NONE V=CIRCLE C=BLUE HEIGHT=2;
SYMBOL2 I=JOIN V=NONE   C=RED WIDTH=2;
AXIS1 ORDER=(-1 TO 14 BY 3) OFFSET=(0.5 CM, 0.5 CM)
LABEL=(HEIGHT=2.5 "Average progression per class");
AXIS2 ORDER=(-1 TO 14 BY 3) OFFSET=(0.5 CM, 0.5 CM)
LABEL=(HEIGHT=2.5 ANGLE=90 "Predicted progression per
class");
PROC GPLOT DATA=BLUPS;
    PLOT PRED*DIFF_A DIFF_A*DIFF_A/OVERLAY HAXIS=AXIS1
    VAXIS=AXIS2 VREF=7.4608;
RUN;QUIT;
```

One-way ANOVA

Residuals

Two types of residuals:

- **Conditional residuals**

  $$\hat{e}_{ij} = y_{ij} - \hat{y}_i$$

  - With $\hat{y}_i$ the BLUP for class i
  - Both fixed effects and random effects are subtracted from the observation
  - Residuals (`resid`) stored with `OUTP=`
- **Marginal residuals**

  $$\hat{e}_{ij} = y_{ij} - \hat{\mu}_{\text{GLS}}$$

  - Only the fixed effects are subtracted from the observations, not the random effects
  - Residuals (`resid`) stored with `OUTPM=`
- Residuals are also standardized

Standardization of residuals:

- By its own variance:

  $$\hat{e}_{ij}/\sqrt{\text{VAR}(\hat{e}_{ij})}$$

  - Called **standardized residuals**
- By the variance of the corresponding observation

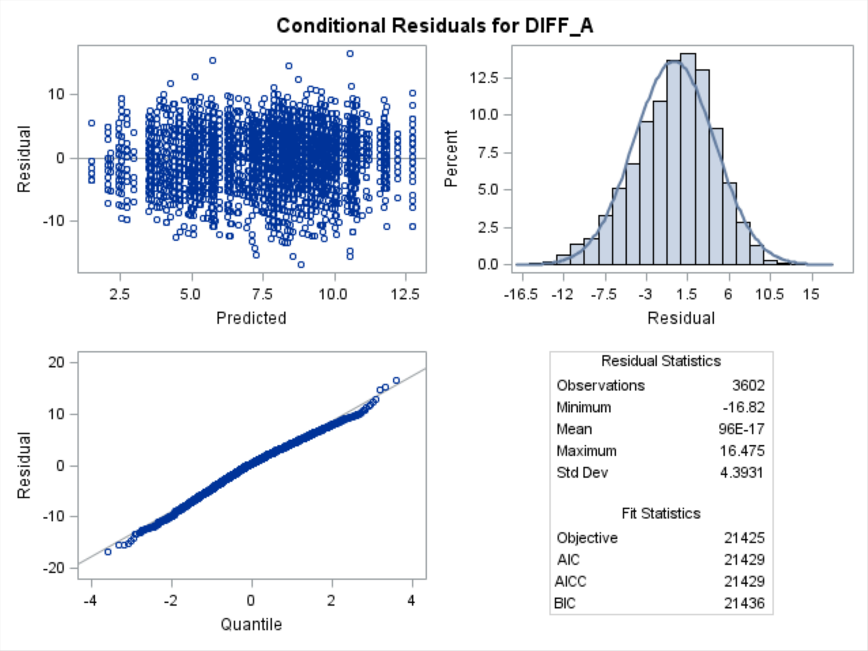
  $$\hat{e}_{ij}/\sqrt{\text{VAR}(y_{ij})}$$

  - Called **Pearson's residuals**
- Standardization helps with detecting possible outliers
- Residuals are visualized with option `RESIDUAL` in model statement

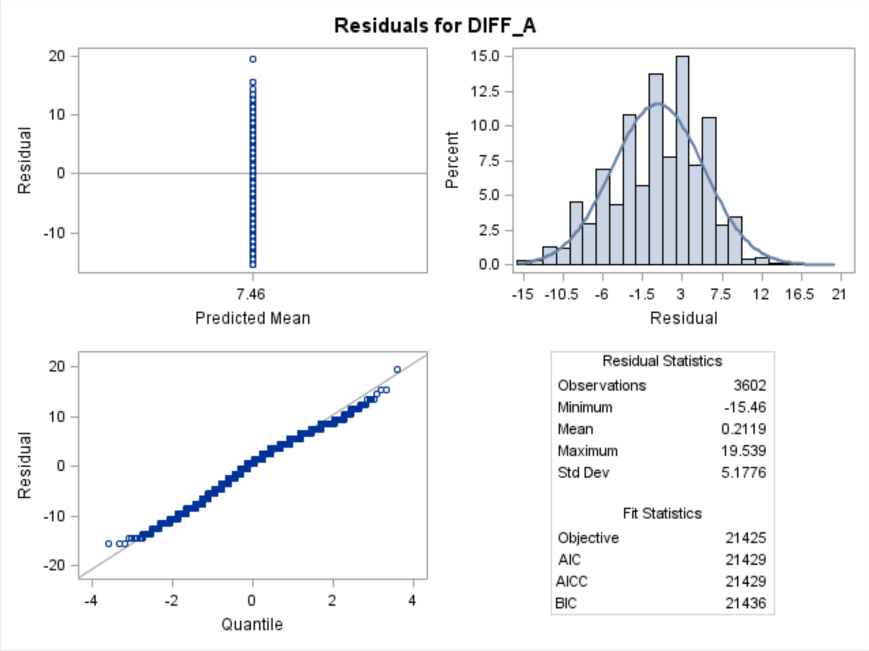
One-way ANOVA

Residuals

Case study: conditional residuals



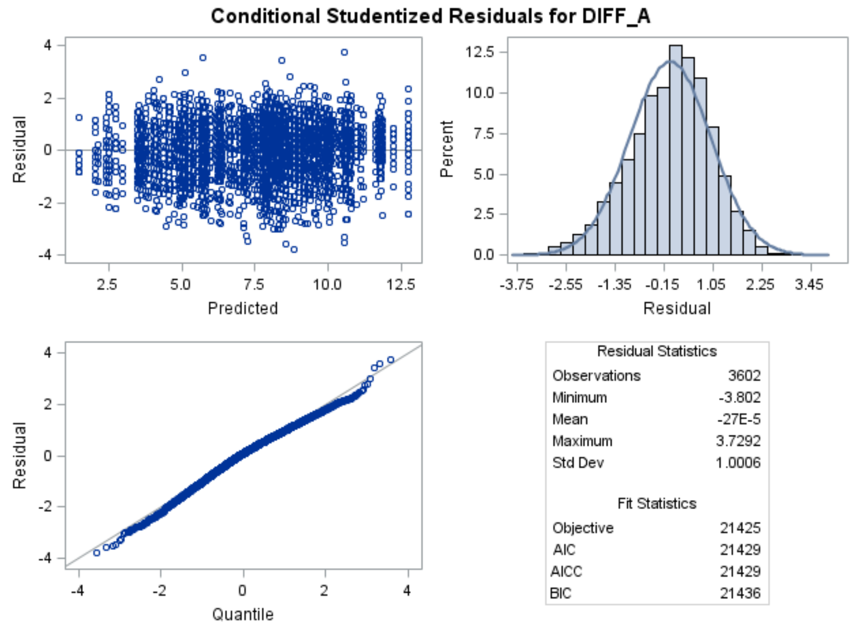
Case study: marginal residuals



One-way ANOVA

Residuals

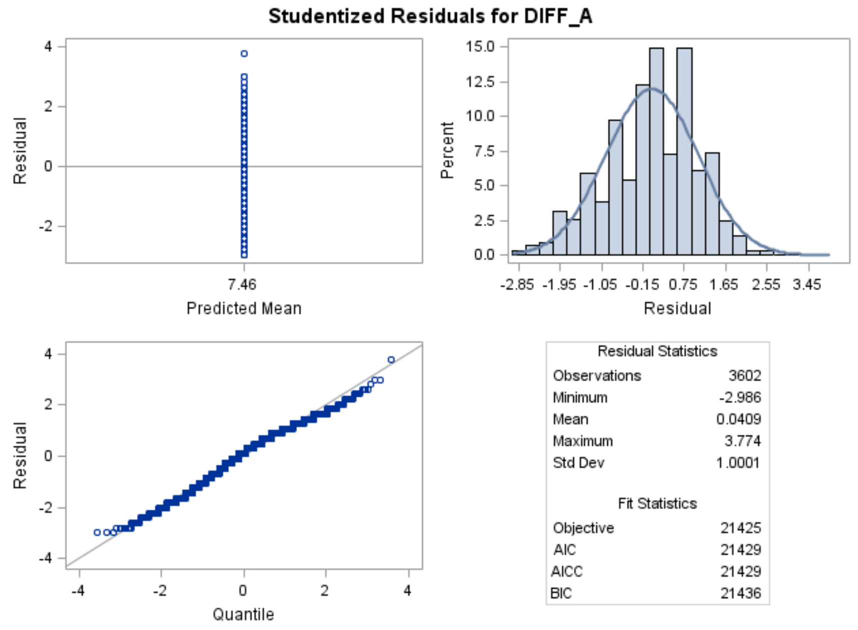
Case study: conditional residuals



Residual Statistics	
Observations	3602
Minimum	-3.802
Mean	-27E-5
Maximum	3.7292
Std Dev	1.0006

Fit Statistics	
Objective	21425
AIC	21429
AICC	21429
BIC	21436

Case study: marginal residuals

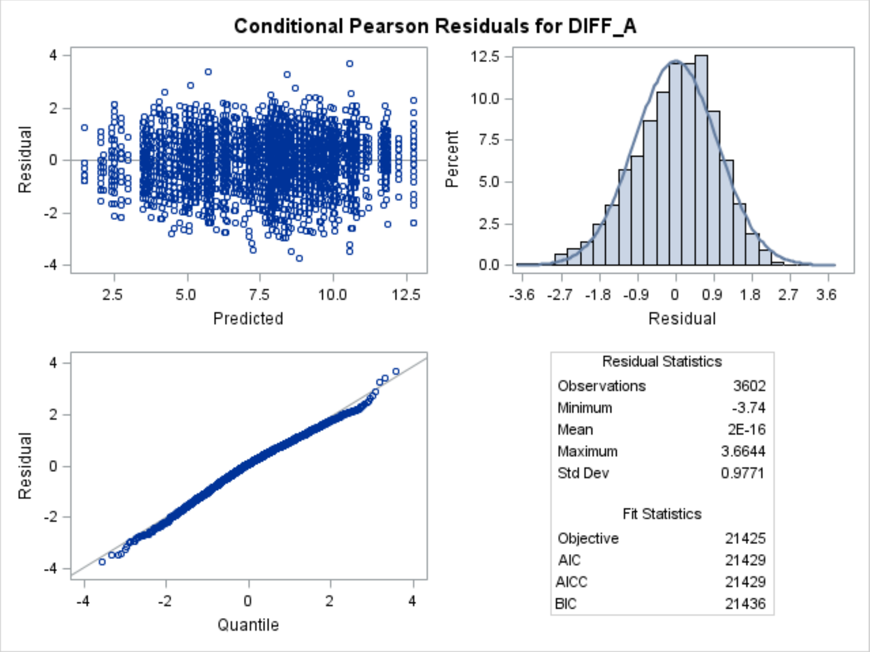


Residual Statistics	
Observations	3602
Minimum	-2.986
Mean	0.0409
Maximum	3.774
Std Dev	1.0001

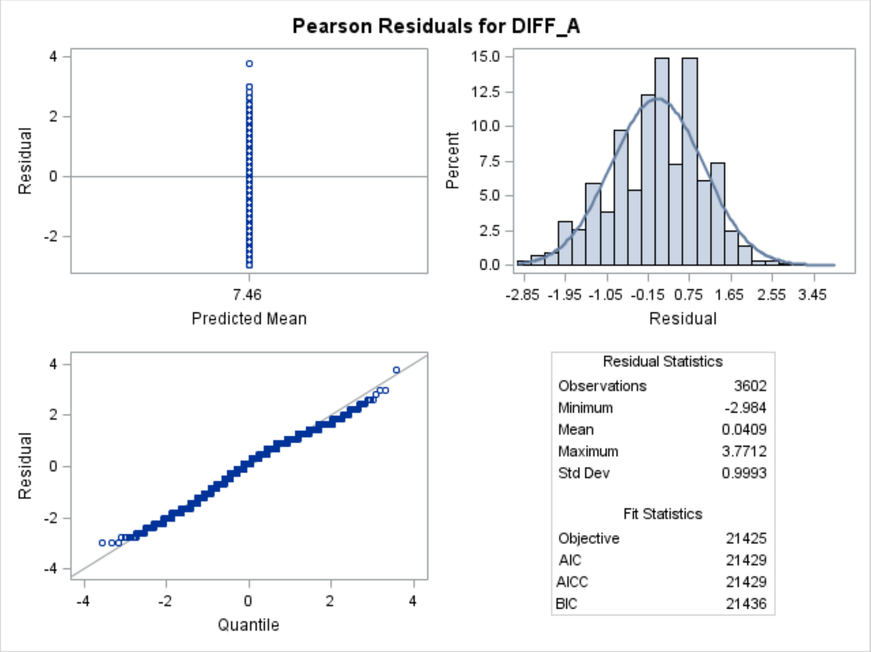
Fit Statistics	
Objective	21425
AIC	21429
AICC	21429
BIC	21436

One-way ANOVA
Residuals

Case study: conditional residuals



Case study: marginal residuals



One-way ANOVA

Residuals

Distribution of the residuals:

- Evaluating residuals from mixed effects ANOVA models is complicated
- Conditional residuals
  - Behave normal due to the elimination of the random effects
  - Behave non-normal due to the VC estimates included in the BLUP's
- Marginal residuals
  - Behave normal since fixed effects estimates are approximately normal
  - Behaves non-normal due to imbalances in sample sizes of random effects

Purpose of residuals:

- Conditional residuals:
  - Presence of outlying observations
  - Homoscedasticity of variances (possibly with respect to other variables)
  - Normality of residuals
  - Check for trends in data
- Marginal residuals:
  - Check for non-linearities of fixed effects with respect to variables
  - Check for trends in data
- **Be cautious:** residuals are correlated leading to unintended effects

One-way ANOVA

Exercise A4

<u>Data analytic questions:</u>

- Perform a one-way ANOVA on the difference in language test scores for all children with class a random effect
- Investigate the BLUPs for each class
  - Study the most extreme BLUPs and averages
  - Make a plot of the predictions against the averages
- Investigate the residuals
  - Conditional residuals
  - Marginal residuals
  - Do you think that the conditions of normality are satisfied?

One-way ANOVA

Likelihood-based estimation

Advantages ANOVA estimators:

- Distributional assumptions?
  - Not for estimation, except the existences for finite second moments
  - Normality and independence for testing and confidence intervals
  - Extended to generalized estimating equations (GEE) – discussed later
- Variance components are estimated separately from fixed effects
- For balanced data, the estimators have nice distributional properties
  - Including being minimal variance

Disadvantages ANOVA estimators:

- Variance component estimates can become negative
  - Truncation to zero is often practiced
  - But this leads to bias
- For unbalanced data
  - Estimation technique is not unique, only for fully nested models (discussed later)
  - Optimality for estimation is unknown
  - Distributional properties of estimators are unknown in general
- Can not estimate all types of correlations under normality

One-way ANOVA

Likelihood-based estimation

Maximum likelihood estimation:

- The outcomes $(y_{i1}, y_{i2}, \dots, y_{in_i})^T$ in group i are i.i.d. normal when conditioned on a_i

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

- Thus, the likelihood function L is

$$\prod_{i=1}^m \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \frac{1}{\sigma_E}\phi\left(\frac{y_{ij}-\mu-z}{\sigma_E}\right)\right] \frac{1}{\sigma_G}\phi\left(\frac{z}{\sigma_G}\right) dz$$

- Maximizing this function with respect to μ, σ_G^2, and σ_E^2 leads to three likelihood equations (derivatives that are set equal to zero)

- Likelihood equations

$$\mu = \frac{\sum_{i=1}^m \sigma_E^2 \bar{y}_{i.}/[\sigma_G^2+\sigma_E^2/n_i]}{\sum_{i=1}^m \sigma_E^2/[\sigma_G^2+\sigma_E^2/n_i]}$$

$$\sum_{i=1}^m \left[\frac{n_i^2(\bar{y}_{i.}-\mu)^2}{[n_i\sigma_G^2+\sigma_E^2]^2}\right] = \sum_{i=1}^m \left[\frac{n_i}{n_i\sigma_G^2+\sigma_E^2}\right]$$

$$\sum_{i=1}^m \left[\frac{n_i(\bar{y}_{i.}-\mu)^2}{[n_i\sigma_G^2+\sigma_E^2]^2}\right] + \frac{SS_W}{\sigma_E^4} = \sum_{i=1}^m \left[\frac{1}{n_i\sigma_G^2+\sigma_E^2}\right] + \frac{n-m}{\sigma_E^2}$$

  - With SS_W the sums of squares for the within group variability

- The ML solutions $\hat{\mu}$, $\hat{\sigma}_G^2$, and $\hat{\sigma}_E^2$ can only be determined numerically for unbalanced data

TU/e

One-way ANOVA

Likelihood-based estimation

Maximum likelihood estimation:

- The ML estimator for μ is equal to the generalized least squares $\hat{\mu}_{\text{GLS}}$
  - This is true for all fixed effects
  - For balanced and unbalanced data

Balanced data:

- The **ML solutions** are

$$\hat{\mu} = \bar{y}_{..}$$
$$\hat{\sigma}_G^2 = [(m-1)MS_B/m - MS_W]/n_0$$
$$\hat{\sigma}_E^2 = MS_W$$

  - Thus, the ML solution underestimates σ_G^2
- ML estimations maximizes the likelihood under constraints $\sigma_G^2 \geq 0$
- Thus, when $mMS_W \geq (m-1)MS_B$, the ML estimator $\hat{\sigma}_G^2 = 0$
- It also affects the estimator for σ_E^2

$$\hat{\sigma}_E^2 = [SS_B + SS_W]/[mn_0]$$

  - The sums of squares for the between group variability ends up in the within group variability
- The **ML estimators** are

$$\hat{\mu} = \bar{y}_{..}$$
$$\hat{\sigma}_G^2 = \max\{0, \frac{(m-1)MS_B/m - MS_W}{n_0}\}$$
$$\hat{\sigma}_E^2 = \begin{cases} MS_W & \text{if} \quad \hat{\sigma}_G^2 > 0 \\ [SS_B + SS_W]/[mn_0] & \text{if} \quad \hat{\sigma}_G^2 = 0 \end{cases}$$

One-way ANOVA

Likelihood-based estimation

Maximum likelihood estimation:

- The ML estimators of the variance components are biased
- The standard errors are based on the Fisher information matrix
  - Calculate the second derivatives of the log likelihood function w.r.t. μ, σ_G^2, and σ_E^2
  - Multiply them with minus one
  - Calculate the expected value to form the 3×3 Fisher information
  - Invert the Fisher information matrix
  - The diagonals form the asymptotic variances of the ML estimators

Standard errors of ML estimators:

- Inverse Fisher information matrix for balanced data is equal to

$$\begin{pmatrix} \frac{\lambda}{mn_0} & 0 & 0 \\ 0 & \frac{2\sigma_E^4}{mn_0^2}\left[\frac{1}{n_0 - 1} + \frac{\lambda^2}{\sigma_E^4}\right] & \frac{-2\sigma_E^4}{mn_0(n_0 - 1)} \\ 0 & \frac{-2\sigma_E^4}{mn_0(n_0 - 1)} & \frac{2\sigma_E^4}{m(n_0 - 1)} \end{pmatrix}$$

  - With $\lambda = n_0\sigma_G^2 + \sigma_E^2$
- Thus, $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ are negatively correlated but independent of $\hat{\mu}$

TU/e

One-way ANOVA

Likelihood-based estimation

Example case study: (Outcome `diff_A`)

- Maximum likelihood: `METHOD=ML`
- Covariance parameters: `ASYCOV`
- Full SAS code:

```
PROC MIXED DATA=name
     METHOD=ML CL ASYCOV;
     CLASS CLASS;
     MODEL DIFF_A = /SOLUTION
           CL DDFM=SAT;
     RANDOM CLASS;
RUN;
```

Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
CLASS	6.7259	0.05	5.3840	8.6434
Residual	20.2159	0.05	19.2890	21.2115

Asymptotic Covariance Matrix of Estimates			
Row	Cov Parm	CovP1	CovP2
1	CLASS	0.6550	-0.01539
2	Residual	-0.01539	0.2400

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	7.4601	0.2019	195	36.95	<.0001	0.05	7.0619	7.8584

One-way ANOVA

Exercise A5

Theoretical questions ML:

- Derive the likelihood equations
- Determine the ML solutions when the data is balanced
- Determine the ML estimators when the data is balanced
- Determine the Fisher information matrix
- Determine the inverse Fisher information matrix

Data analytical questions ML:

- Fit one-way random effects model with ML (see exercise A3)
  - Estimate the model parameters
  - Obtain the confidence intervals
  - Calculate the intraclass correlation coefficient and a 95% confidence interval
  - Calculate the total variability and a 95% confidence interval

One-way ANOVA

Likelihood-based estimation

Disadvantage maximum likelihood:

- The ML solutions are biased
  - Does not use correct degrees of freedom for the between group sums of squares
  - Is a general problem for simpler and more complex ANOVA models
- Let's assume that $y_1, y_2, \ldots, y_m$ are i.i.d. normal $y_i \sim N(\mu, \sigma_E^2)$, the MLE is

$$\hat{\mu} = \sum_{i=1}^{m} y_i / m$$

$$\hat{\sigma}_E^2 = \sum_{i=1}^{m} [(y_i - \bar{y}_.)^2 / m]$$

  - MLE does not use $m - 1$
  - Although MS_W for the random effects model was correct when $\hat{\sigma}_G^2 > 0$

Restricted maximum likelihood (REML):

- Log likelihood y_i i.i.d. $N(\mu, \sigma_E^2)$

$$-\sum_{i=1}^{m} [\log(\sigma_E) + (y_i - \mu)^2 / (2\sigma_E^2)]$$

$$= -\frac{1}{2} \log(\sigma_E^2) - m(\bar{y}_. - \mu)^2 / (2\sigma_E^2)$$

$$- \frac{m-1}{2} \log(\sigma_E^2) - \sum_{i=1}^{m} [(y_i - \bar{y}_.)^2 / (2\sigma_E^2)]$$

- REML likelihood for variance:

$$- \frac{m-1}{2} \log(\sigma_E^2) - \sum_{i=1}^{m} [(y_i - \bar{y}_.)^2 / (2\sigma_E^2)]$$

  - Does not contain any fixed effects
- Maximizing for σ_E^2 gives:

$$\hat{\sigma}_E^2 = \sum_{i=1}^{m} [(y_i - \bar{y}_.)^2 / (m - 1)]$$

  - Maximizing full likelihood for μ given $\hat{\sigma}_E^2$ is still $\hat{\mu} = \sum_{i=1}^{m} y_i / m$

One-way ANOVA

Likelihood-based estimation

Restricted maximum likelihood:

- Uses a likelihood function for a linear transformation of the data such that
  - Fixed effects are eliminated from model
  - Variance components (VC's) remain
- Maximizes REML likelihood for VC's
- Then full likelihood is maximized for the fixed effects, given VC estimators
  - Still generalized least squares estimators
  - Full likelihood theory remains appropriate
  - Different value from MLE due to different VC estimators
- Proc MIXED uses REML as default

Restricted maximum likelihood (REML):

- For the balanced one-way random effects model the estimators are
- The **REML estimators** are

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

$$\hat{\sigma}_G^2 = \max\{0, [MS_B - MS_W]/n_0\}$$

$$\hat{\sigma}_E^2 = \begin{cases} MS_W & \text{if} \quad \hat{\sigma}_G^2 > 0 \\ \frac{SS_B + SS_W}{mn_0 - 1} & \text{if} \quad \hat{\sigma}_G^2 = 0 \end{cases}$$

- Comparisons with ANOVA estimators
  - Equal when $\hat{\sigma}_G^2 > 0$
  - Different when $\hat{\sigma}_G^2 = 0$, since the ANOVA estimator for σ_E^2 remains $\hat{\sigma}_E^2 = MS_W$

One-way ANOVA

Likelihood-based estimation

Example case study: (`diff_A`)

- Full SAS code:

```
PROC MIXED DATA=name
      METHOD=REML CL ASYCOV;
      CLASS CLASS;
      MODEL DIFF_A = /SOLUTION
      CL DDFM=SAT;
      RANDOM CLASS;
RUN;
```

- BLUPs for class averages can be added for all estimation methods

Comparing estimates:

- Fixed effects different due to other estimates for VC's
- VC's different due to unbalanced data
- DF's different since based on VC's
- ANOVA CI for σ_G^2 calculated differently

Parameter	ANOVA	ML	REML
μ	7.4608 [7.065; 7.857]	7.4601 [7.062; 7.858]	7.4598 [7.061; 7.859]
$df\ \mathrm{SE}(\hat{\mu})$	200	195	194
σ_G^2	6.6320 [5.087; 8.177]	6.7259 [5.384; 8.643]	6.7672 [5.415; 8.700]
σ_E^2	20.2131 [19.286; 21.209]	20.2159 [19.289; 21.212]	20.2158 [19.289; 21.212]

One-way ANOVA

Exercise A6

Theoretical questions REML:

- Derive the likelihood function
- Derive the likelihood equations
- Determine the REML solutions when the data is balanced
- Determine the REML estimators when the data is balanced

Data analytical questions REML:

- Fit one-way random effects model with REML (see exercise A3)
  - Estimate the model parameters
  - Obtain the confidence intervals
  - Calculate the intraclass correlation coefficient and a 95% confidence interval
  - Calculate the total variability and a 95% confidence interval

One-way ANOVA

Choice of estimation

<u>Opinion in literature:</u>

- There is a strong preference for REML
  - Solutions are unbiased (takes into account the fixed effect)
  - Does not result in negative VC estimates
  - Unique estimation for balanced and unbalanced data
  - Results into ANOVA estimators (when all VC's are positive) for balanced data
- When sample sizes are large, REML has hardly no advantage over ML
  - Then ML is preferred over REML when it comes to model selection (discussed later)

<u>Critical considerations:</u>

- ANOVA estimators do not make use of normality and are more general
- For the balanced data:
  - ANOVA is optimal (minimum variance)
  - We recommend truncation of negative VC's
  - For negative VC's REML and ANOVA are different, and ANOVA is better
- For unbalanced data:
  - ANOVA still have value when degrees of freedom is small for one VC
  - For large sample sizes ML is easiest and REML has hardly any advantage

Two-way ANOVA models

Nested mixed effects model

<u>**Case study:**</u> (outcome `IQV`)

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

- With y_{ijk} the outcome of child $k \in \{1,2,\dots,K_{ij}\}$ in class $j \in \{1,2,\dots,J_i\}$ within type of class $i \in \{1,2,\dots,I\}$
- With μ_i the mean outcome for class type i (thus assumed a fixed effect)
- With $a_{j(i)}$ the random effect of class j within type of class i, $a_{j(i)} \sim N(0, \sigma^2_{C(T)})$
- With e_{ijk} the residual (or error) for child j in class i, $e_{ijk} \sim N(0, \sigma^2_R)$
- With all random terms independent

- Fixed effect can be reparametrized
  - $\mu_i = \mu + \alpha_i$, with $\alpha_I = 0$
  - One parameter α_i must be zero for **identifiability** purposes
  - Software packages often chooses last one
  - Assumption of $\alpha_1 + \alpha_2 + \dots + \alpha_I = 0$ is less reasonable in unbalanced data sets
- Model is the combination of I different one-way random effects models
  - With restriction that between and within class variability is constant across class type
  - This **homoscedasticity** assumptions can be violated and tested for correctness

Two-way ANOVA models

Nested mixed effects model

Case study: (outcome IQV)

```
PROC MIXED DATA=ANALYSIS
 METHOD=TYPE3 CL COVTEST;

 CLASS CLASS COMBI;

 MODEL IQV = COMBI
     / SOLUTION CL
       DDFM=SAT;

 RANDOM CLASS(COMBI)

RUN;
```

- Variable COMBI has two levels
  - 0: single grade classes
  - 1: multi grade classes
- Class is now nested within class type

Source	df	SS	MS	$\mathbb{E}MS$
Type	1	50.397	50.397	$\sigma_R^2 + 13.354\sigma_{C(T)}^2 + Q(\boldsymbol{\mu})$
Class(Type)	204	2874.7	14.092	$\sigma_R^2 + 18.978\sigma_{C(T)}^2$
Residual	3704	14096	3.806	σ_R^2

Fixed effects estimates:

- μ: 11.64 [11.43; 11.85]
- α_1: 0.218 [-0.040; 0.475]

VC estimates:

- $\sigma_{C(T)}^2$: 0.542 [0.397; 0.687]
- σ_R^2: 3.806 [3.638; 3.985]

Additional estimates:

- Total: 4.348 [4.137; 4.575]
- ICC: 0.125 [0.098; 0.158]

Two-way ANOVA models

Nested mixed effects model

<u>Estimation balanced data:</u>

- Mean squares:
  - Type of class ($df_T = I - 1$):
    $MS_T = \sum_{i=1}^{I} JK(\bar{y}_{i..} - \bar{y}_{...})^2 / df_T$
  - Class within type ($df_{C(T)} = I(J-1)$):
    $MS_{C(T)} = \sum_{i=1}^{I} \sum_{j=1}^{J} K\left(\bar{y}_{ij.} - \bar{y}_{i..}\right)^2 / df_{C(T)}$
  - Residual ($df_R = IJ(K-1)$):
    $MS_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(y_{ijk} - \bar{y}_{ij.}\right)^2 / df_R$
- Expected mean squares
  - Using the one-way ANOVA
    $\mathbb{E}[MS_R] = \sigma_R^2$
    $\mathbb{E}\left[MS_{C(T)}\right] = \sigma_R^2 + K\sigma_{C(T)}^2$
- MS_T requires some algebra
  $$\mathbb{E}[MS_T] = \frac{1}{2} JK\alpha_1^2 + K\sigma_{C(T)}^2 + \sigma_R^2$$
  - $(I-1)MS_T/[K\sigma_{C(T)}^2 + \sigma_R^2]$ is non-central chi-square distributed with
- Fixed effects (assuming $\alpha_2 = 0$)
  - $\hat{\mu} = \bar{y}_{2..} \sim N(\mu, [K\sigma_{C(T)}^2 + \sigma_R^2]/[JK])$
  - $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{2..}$
    $\sim N(\alpha_1, 2[K\sigma_{C(T)}^2 + \sigma_R^2]/[JK])$
- Estimation SE fixed effects
  - $\widehat{\text{VAR}}(\hat{\mu}) = MS_{C(T)}/[JK]$
  - $\widehat{\text{VAR}}(\hat{\alpha}_1) = 2MS_{C(T)}/[JK]$

Two-way ANOVA models

Nested mixed effects model

Estimation unbalanced data:

- Sums of squares for factors/terms
  - Are not uniquely defined anymore in higher-order models
  - There exists many choices for partitioning the total sums of squares
- One general approach for ANOVA models is Henderson's third method
  - It can still be executed in several ways by building up the model in different ways
  - Two common approaches are the TYPE 1 and TYPE 3 options
  - For balanced data TYPE 1 = TYPE 3

- **TYPE 1 estimation** [`METHOD=TYPE1`]
  - ANOVA model is built up sequentially by including one factor/term at the time
  - The partial contribution of each term to the sums of squares is calculated
  - Used when the order of terms is known
- **TYPE 3 estimation** [`METHOD=TYPE3`]
  - Factors/terms are included in ANOVA model after all other terms are included
  - Referred to as the contribution of each term **corrected** for all other terms
  - Is the recommended method, but this is less obvious in nested models

Two-way ANOVA models

Nested mixed effects model

ANOVA table TYPE 1 estimation:

Type 1 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
COMBI	1	12.870352	12.870352	Var(Residual) + 16.629 Var(CLASS(COMBI)) + Q(COMBI)	0.8762 MS(CLASS(COMBI)) + 0.1238 MS(Residual)	219.85	1.00	0.3174
CLASS(COMBI)	204	2874.695337	14.091644	Var(Residual) + 18.978 Var(CLASS(COMBI))	MS(Residual)	3704	3.70	<.0001
Residual	3704	14096	3.805602	Var(Residual)	.	.	.	.

ANOVA table TYPE 3 estimation:

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
COMBI	1	50.387226	50.387226	Var(Residual) + 13.354 Var(CLASS(COMBI)) + Q(COMBI)	0.7036 MS(CLASS(COMBI)) + 0.2964 MS(Residual)	252.86	4.56	0.0336
CLASS(COMBI)	204	2874.695337	14.091644	Var(Residual) + 18.978 Var(CLASS(COMBI))	MS(Residual)	3704	3.70	<.0001
Residual	3704	14096	3.805602	Var(Residual)	.	.	.	.

- For fully nested models:
  - **TYPE 1/TYPE 3 estimation of model parameters is identical**
  - Only the ANOVA tables differ and thus testing for class type ($H_0\colon \alpha_1 = 0$) is different
  - TYPE 1 estimation is more reliable ($\alpha_1 = 0.218\ [-0.040;\ 0.475]$) for nested models

Two-way ANOVA models

Nested mixed effects model

Variance heteroscedasticity:

- Is the variability between classes for each type of class the same?
- Null hypothesis:

$$H_0: \sigma^2_{C(T)}(0) = \sigma^2_{C(T)}(1)$$

  - With $\sigma^2_{C(T)}(i)$ the variance for class difference within class type i
- This is easiest tested with a **likelihood ratio test**
  - Uses maximum likelihood estimation
  - Compares the likelihood values for two hierarchical models

Likelihood ratio test (LRT):

- $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ is the log likelihood function
  - With $\boldsymbol{\theta}$ the vector of all model parameters
  - With $\boldsymbol{y}$ the vector of all observations
- Under H_0 parameters $\boldsymbol{\theta}$ reduces to $\boldsymbol{\theta}_0$
- Test statistic:

$$LRT = -2[\ell(\widehat{\boldsymbol{\theta}}_0|\boldsymbol{y}) - \ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})]$$

  - With $\widehat{\boldsymbol{\theta}}$ the ML estimator
  - With $\widehat{\boldsymbol{\theta}}_0$ the ML estimator under H_0
- LRT is approximately χ^2 distributed
  - With df the difference in number of parameters between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$
  - When parameters are not on boundaries

Two-way ANOVA models

Nested mixed effects model

Case study: (outcome IQV)

```
PROC MIXED DATA=ANALYSIS
     METHOD=ML CL;
     CLASS CLASS COMBI;
     MODEL IQV = COMBI
        / SOLUTION CL DDFM=SAT;
     RANDOM CLASS(COMBI)
        / GROUP=COMBI;
RUN;
```

- The option GROUP=COMBI estimates for each level of COMBI a variance component of each random effect

Covariance Parameter Estimates					
Cov Parm	Group	Estimate	Alpha	Lower	Upper
CLASS(COMBI)	COMBI 0	0.3768	0.05	0.2696	0.5637
CLASS(COMBI)	COMBI 1	1.1167	0.05	0.7652	1.7822
Residual		3.8105	0.05	3.6426	3.9903

Comparing log likelihoods:

- Full model: $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = 16591.5$
- Reduced model: $-2\ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}) = 16606.3$
- $LRT = 16606.3 - 16591.5 = 14.8$
- $df = 5 - 4 = 1$
- $P < 0.001$ [1-PROBCHI(LRT,DF)]
- Homoscedasticity of the between class variability for type of class is rejected

Two-way ANOVA models

Exercise A7

<u>**Data analytical questions:**</u>

- Fit the two-way nested mixed effects model for the difference in language score using the ANOVA estimator
  - Investigate the differences between TYPE 1 and TYPE 3 estimation
  - Report the parameter estimates and their 95% confidence intervals
  - Report the total variability and intraclass correlation coefficient with their 95% confidence intervals

- Fit the two-way nested mixed effects model for the difference in language score using the ML estimator
  - Investigate heteroscedasticity for between class variability for the two class types.
  - Report the relevant information

Two-way ANOVA models

Nested random effects model

<u>**Case study:**</u> (outcome `IQV`)

$$y_{ijk} = \mu + a_i + b_{j(i)} + e_{ijk}$$

- With y_{ijk} the outcome of child $k \in \{1,2,\dots,K_{ij}\}$ in class $j \in \{1,2,\dots,J_i\}$ within type of class $i \in \{1,2,\dots,I\}$
- With μ the mean outcome
- With a_i the random effect of type of class i, $a_i \sim N(0, \sigma_T^2)$
- With $b_{j(i)}$ the random effect of class within type of class i, $b_{j(i)} \sim N(0, \sigma_{C(T)}^2)$
- With e_{ijk} the residual (or error) for child k in class j of class type i, $e_{ijk} \sim N(0, \sigma_R^2)$
- With all random terms independent

- The levels of class type are assumed a random draw of all possible levels
  - May be not very realistic in this setting
- Model inferences:
  - Variance: $\mathrm{VAR}(y_{ijk}) = \sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2$
  - Covariances:

$$\mathrm{CORR}(y_{ijk_1}, y_{ijk_2}) = \frac{\sigma_T^2 + \sigma_{C(T)}^2}{\sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2}$$

$$\mathrm{CORR}(y_{ij_1k_1}, y_{ij_2k_2}) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2}$$

$$\mathrm{CORR}(y_{i_1j_1k_1}, y_{i_2j_2k_2}) = 0$$

- Correlation structure more difficult than two-way nested mixed model

Two-way ANOVA models

Nested random effects model

General form ICCs:

- The correlations can all be written in

  $ICC = \frac{\sigma_G^2}{\sigma_G^2+\sigma_E^2}$

  - Where both variance components σ_G^2 and σ_E^2 can be single or sums of other variance components
  - For $\mathrm{CORR}\left(y_{ijk_1}, y_{ijk_2}\right)$ we have

    $\sigma_G^2 = \sigma_T^2 + \sigma_{C(T)}^2$ and $\sigma_E^2 = \sigma_R^2$
  - For $\mathrm{CORR}\left(y_{ij_1k_1}, y_{ij_2k_2}\right)$ we have

    $\sigma_G^2 = \sigma_T^2$ and $\sigma_E^2 = \sigma_{C(T)}^2 + \sigma_R^2$
- Calculation of confidence intervals can be formulated generally

General form ICCs:

- Let $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2, \ldots, \sigma_{p+q}^2$ be VC's in an ANOVA model with

  $\sigma_G^2 = \sum_{r=1}^{p} \sigma_r^2$

  $\sigma_E^2 = \sum_{r=p+1}^{p+q} \sigma_r^2$
- Let $\hat{\sigma}_r^2$ be the REML estimator of σ_r^2
  - Then $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ are the REML estimator for σ_G^2 and σ_E^2 by substituting $\hat{\sigma}_r^2$
- Now define $\tau_{rs} = \mathrm{COV}(\hat{\sigma}_r^2, \hat{\sigma}_s^2)$ and let $\hat{\tau}_{rs}$ be the estimated covariance
  - These estimates can be obtained from procedure MIXED (option `ASYCOV`)

Two-way ANOVA models

Nested random effects model

<u>F-approach confidence interval ICCs:</u>

- Standard errors of $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$

$$\hat{\tau}_G^2 \equiv \widehat{\text{VAR}}(\hat{\sigma}_G^2) = \sum_{r=1}^{p} \sum_{s=1}^{p} \hat{\tau}_{rs}$$

$$\hat{\tau}_E^2 \equiv \widehat{\text{VAR}}(\hat{\sigma}_E^2) = \sum_{r=p+1}^{p+q} \sum_{s=p+1}^{p+q} \hat{\tau}_{rs}$$

- Using Satterthwaite approach:

$$df_G = 2\hat{\sigma}_G^4 / \hat{\tau}_G^2$$

$$df_E = 2\hat{\sigma}_E^4 / \hat{\tau}_E^2$$

- We can use the F-approach when the following is approximately true
  - $df_G \hat{\sigma}_G^2 / \sigma_G^2 \sim \chi_{df_G}^2$
  - $df_E \hat{\sigma}_E^2 / \sigma_E^2 \sim \chi_{df_E}^2$
  - $\hat{\sigma}_G^2$ and $\hat{\sigma}_E^2$ independent

<u>F-approach confidence interval ICCs:</u>

- The $100\%(1-\alpha)$ approximate confidence limits are:

$$LCL = \frac{\hat{\sigma}_G^2 F_L}{\hat{\sigma}_G^2 F_L + \hat{\sigma}_E^2} \text{ and } UCL = \frac{\hat{\sigma}_G^2 F_U}{\hat{\sigma}_G^2 F_U + \hat{\sigma}_E^2}$$

  - With F_L and F_U the lower and upper $\alpha/2$ quantile of the F-distribution with df_G and df_E degrees of freedom, respectively
  - It does NOT result in the F-approach for one-way random effects models
- The approach works best when many variance components are involved
- Must be programmed manually in SAS

TU/e

Two-way ANOVA models

Nested random effects model

Confidence interval sum of VCs:

- The ICC made use of the assumption
  - $df_G \hat{\sigma}_G^2 / \sigma_G^2 \sim \chi_{df_G}^2$
- Thus a $100\%(1-\alpha)$ approximate confidence intervals on σ_G^2 is now

$$LCL = df_G \hat{\sigma}_G^2 / \chi_{df_G}^{-2}(1-\alpha/2)$$

$$UCL = df_G \hat{\sigma}_G^2 / \chi_{df_G}^{-2}(\alpha/2)$$

  - With $\chi_d^{-2}(q)$ the qth quantile of the chi-square distribution with d degrees of freedom
- Can be applied to any sum of VC's
- Must be programmed manually in SAS

Case study: (outcome `IQV`)

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

- `CLASS(COMBI)` and `COMBI` are random
- VC estimates are collected in data set `VCS`
- Covariances are collected in data set `COVS`

Two-way ANOVA models

Nested random effects model

Case study: (outcome `IQV`)

- Parameter estimates (REML)
  - $\hat{\mu} = 11.76\ [10.37; 13.15]$
  - $\hat{\sigma}_T^2 = 0.0152\ [0.0020; 658861]$
  - $\hat{\sigma}_{C(T)}^2 = 0.5995\ [0.4616; 0.8103]$
  - $\hat{\sigma}_R^2 = 3.8166\ [3.6481; 3.9970]$
  - Variance for class type is unreliable

Asymptotic Covariance Matrix of Estimates				
Row	Cov Parm	CovP1	CovP2	CovP3
1	COMBI	0.001179	0.000017	-0.00002
2	CLASS(COMBI)	0.000017	0.007330	-0.00062
3	Residual	-0.00002	-0.00062	0.007904

- Total variance: $\sigma_{\mathrm{TOT}}^2 = \sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2$

  $\hat{\sigma}_{\mathrm{TOT}}^2 = 4.4313\ [4.1995; 4.6829]$

  $\hat{\tau}_{\mathrm{TOT}}^2 = 0.01516$

  $df_{\mathrm{TOT}} = 2590.5$
- $ICC_1 = \left[\sigma_T^2 + \sigma_{C(T)}^2\right] / [\sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2]$

  $I\hat{C}C_1 = 0.139\ [0.105; 0.175]$

  $\hat{\tau}_G^2 = 0.0085$

  $df_G = 88.5$

  $\hat{\tau}_E^2 = 0.0079$

  $df_E = 3685.9$
- $ICC_2 = \sigma_T^2 / [\sigma_T^2 + \sigma_{C(T)}^2 + \sigma_R^2]$

  $I\hat{C}C_2 = 0.003\ [0.000; 0.026]$

TU/e

Two-way ANOVA models

Crossed mixed effects model

Statistical model:

$$y_{hij} = \mu + \alpha_h + a_i + (\alpha a)_{hi} + e_{hij}$$

- With y_{hij} the outcome of child $j \in \{1,2,\dots,J_{hi}\}$ in class $i \in \{1,2,\dots,I_h\}$ with sex $h \in \{1,2,\dots,H\}$
- With μ the overall mean outcome
- With α_h the effect of sex ($\alpha_H = 0$ for model identifiability)
- With $a_i \sim N(0,\sigma_C^2)$ the effect of class
- With $(\alpha a)_{hi} \sim N(0,\sigma_{SC}^2)$ the interaction effect of sex and class
- With $e_{hij} \sim N(0,\sigma_R^2)$ the residual
- With all random terms independent

Interaction effects:

- An interaction between a fixed and random effects factor is random
  - The difference between a boy and a girls within class varies across classes
- Differences in outcomes within class:
  - For same sex: $y_{hij_1} - y_{hij_2} = e_{hij_1} - e_{hij_2}$ is normal $N(0, 2\sigma_R^2)$ with no class effect
  - For opposite sex: $y_{2ij_1} - y_{1ij_2} = \alpha_2 - \alpha_1 + (\alpha a)_{2i} - (\alpha a)_{1i} + e_{hij_1} - e_{hij_2}$ is normal $N(\alpha_2 - \alpha_1, 2[\sigma_{SC}^2 + \sigma_R^2])$ with a class effect
  - The VC σ_{SC}^2 represents this variability

TU/e

Two-way ANOVA models

Crossed mixed effects model

Characterization statistical model:

- The distribution of y_{hij} is normal

  $y_{hij} \sim N(\mu + \alpha_h, \sigma^2_{\mathrm{TOT}})$

  - With $\sigma^2_{\mathrm{TOT}} = \sigma^2_C + \sigma^2_{SC} + \sigma^2_R$
- Correlation coefficients
  - Within class and sex

    $$ICC = \frac{\sigma^2_C + \sigma^2_{SC}}{\sigma^2_C + \sigma^2_{SC} + \sigma^2_R}$$
  - Within class for different sex

    $$ICC = \frac{\sigma^2_C}{\sigma^2_C + \sigma^2_{SC} + \sigma^2_R}$$
- Confidence intervals on total variability and ICC are as before

Estimation balanced data:

- Sums of squares:
  - Sex ($df_S = H - 1$):

    $SS_S = \sum_{h=1}^{H} IJ(\bar{y}_{h..} - \bar{y}_{...})^2$
  - Class ($df_C = I - 1$):

    $SS_C = \sum_{i=1}^{I} HJ(\bar{y}_{.i.} - \bar{y}_{...})^2$
  - Sex*Class ($df_{SC} = (H-1)(I-1)$)

    $$SS_{SC} = \sum_{h=1}^{H} \sum_{i=1}^{I} J(\bar{y}_{hi.} - \bar{y}_{h..} - \bar{y}_{.i.} + \bar{y}_{...})^2$$
  - Residual ($df_R = HI(J-1)$):

    $SS_R = \sum_{h=1}^{H} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(y_{hij} - \bar{y}_{hi.}\right)^2$
- Mean squares are calculated as usual:

  $MS = SS/df$

Two-way ANOVA models

Crossed mixed effects model

<u>Estimation balanced data:</u>

- Expected mean squares:

$$\mathbb{E}[MS_R] = \sigma_R^2$$
$$\mathbb{E}[MS_{SC}] = \sigma_R^2 + J\sigma_{SC}^2$$
$$\mathbb{E}[MS_C] = \sigma_R^2 + J\sigma_{SC}^2 + HJ\sigma_C^2$$
$$\mathbb{E}[MS_S] = \sigma_R^2 + J\sigma_{SC}^2 + \frac{1}{H-1}\sum_{h=1}^{H} IJ\alpha_h^2$$

- Variance component estimators

$$\hat{\sigma}_R^2 = MS_R$$
$$\hat{\sigma}_{SC}^2 = [MS_{SC} - MS_R]/J$$
$$\hat{\sigma}_C^2 = [MS_C - MS_{SC}]/[HJ]$$

- Fixed effects (assuming $\alpha_2 = 0$)
  - $\hat{\mu} = \bar{y}_{2..} \sim N(\mu, [\sigma_C^2 + \sigma_{SC}^2]/I + \sigma_R^2/[IJ])$
  - $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{2..} \sim N(\alpha_1, [\sigma_R^2 + J\sigma_{SC}^2]/[IJ])$

- Estimation SE's of fixed effects:
  - $\widehat{\mathrm{VAR}}(\hat{\mu}) = [MS_C + (H-1)MS_{SC}]/[HIJ]$ Requires Satterthwaite degrees of freedom
  - $\widehat{\mathrm{VAR}}(\hat{\alpha}_1) = MS_{SC}/[IJ]$
- Fixed effects (assuming $\alpha_1 + \alpha_2 = 0$)
  - $\hat{\mu} = \bar{y}_{...} \sim N(\mu, [\sigma_R^2 + J\sigma_{SC}^2 + HJ\sigma_C^2]/[HIJ])$
  - $\hat{\alpha}_h = \bar{y}_{h..} - \bar{y}_{...}$ $\sim N(\alpha_h, (H-1)[\sigma_R^2 + J\sigma_{SC}^2]/[HIJ])$
- Estimation SE's of fixed effects:
  - $\widehat{\mathrm{VAR}}(\hat{\mu}) = MS_C/[HIJ]$ Requires no Satterthwaite degrees of freedom
  - $\widehat{\mathrm{VAR}}(\hat{\alpha}_h) = (H-1)MS_{SC}/[HIJ]$
  - Thus $\alpha_1 + \alpha_2 = 0$ makes more sense

Two-way ANOVA models

Crossed mixed effects model

Hypothesis testing balanced data:

- Hypothesis testing:
  - $H_0: \sigma_{SC}^2 = 0 \Longrightarrow F = MS_{SC}/MS_R$
  - $H_0: \sigma_C^2 = 0 \Longrightarrow F = MS_C/MS_{SC}$
  - $H_0: \alpha_h = 0 \Longrightarrow F = MS_S/MS_{SC}$
- Under the null hypothesis, these test statistics are all F-distributed with the appropriate degrees of freedom
- Null hypothesis $\boldsymbol{H_0: \sigma_R^2 = 0}$ **is never tested** (there is always a residual)
  - If you do get $\sigma_R^2 = 0$ in an analysis, you have included too many terms in model

Estimation unbalanced data:

- All four estimation techniques (TYPE 1, TYPE 3, ML, REML) would make sense
  - **TYPE 1:** Need to know the order of factors: sex before class or class before sex?
  - **TYPE 3:** When order of factors is unknown
  - **ML:** When degrees of freedom for all terms in the model are large
  - **REML:** most generally accepted estimation technique, except when one or more VC's may become negative with TYPE 1/TYPE 3
- For balanced data, TYPE 1, TYPE 3, and REML are all identical (when VC's >0)

Two-way ANOVA models

Crossed mixed effects model

<u>**Case Study:**</u> (outcome `IQV`)

```
PROC MIXED DATA=name METHOD=TYPE1 CL;
    CLASS CLASS GIRL;
    MODEL IQV = GIRL / SOLUTION CL DDFM=SAT;
    RANDOM CLASS GIRL*CLASS;
RUN;
```

- `GIRL` in model statement for fixed effect
- `GIRL*CLASS` in random statement for the random interaction effects
- Other options can be added:
  - `ASYCOV` for the covariance parameters of the VC estimators
  - `RESIDUAL` , for the visualization of the residuals
  - `OUTP=` and `OUTPM=` for the predictions per class and sex

Two-way ANOVA models

Crossed mixed effects model

Parameters	TYPE 1	TYPE 3	ML	REML
μ	11.77	11.77	11.76	11.76
	[11.63; 11.90]	[11.63; 11.90]	[11.63; 11.90]	[11.63; 11.90]
$df\ SE(\hat{\mu})$	335	326	290	289
α_1	0.094	0.094	0.094	0.094
	[-0.032; 0.221]	[-0.032; 0.221]	[-0.033; 0.222]	[-0.033; 0.222]
$df\ SE(\hat{\alpha}_1)$	205	205	198	197
σ_C^2	0.540	0.552	0.597	0.600
	[0.392; 0.688]	[0.400; 0.703]	[0.458; 0.811]	[0.460; 0.816]
σ_{SC}^2	0.022	0.022	0.027	0.029
	[-0.054; 0.098]	[-0.054; 0.098]	[0.005; 50.16]	[0.006; 22.57]
σ_R^2	3.800	3.800	3.807	3.807
	[3.631; 3.980]	[3.631; 3.980]	[3.639; 3.988]	[3.639; 3.988]

Two-way ANOVA models

Exercise A8

Theoretical questions:

- Consider a balanced two-way crossed mixed effects model with no interaction
  - Determine the mean squares
  - Determine the expected mean squares
  - Determine the estimators for all model parameters

Data analytical questions:

- Fit the two-way crossed ANOVA model for the difference in language score
  - Report the parameters and their 95% confidence intervals
  - Calculate the total variability and the intraclass correlation coefficients with their 95% confidence intervals
  - Investigate the conditional and marginal residuals
  - Run all analysis for the four estimation methods: TYPE1, TYPE3, ML, REML and discuss the differences

Three-way ANOVA models

Expected mean squares

Case study: (outcome `IQV`)

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

- With y_{hijk} the outcome
- With μ the overall mean
- With α_h the effect of sex ($\alpha_2 = 0$)
- With β_i the effect of class type ($\beta_2 = 0$)
- With $(\alpha\beta)_{hi}$ the interaction effect of sex and class type ($(\alpha\beta)_{11} \neq 0$)
- With $a_{j(i)} \sim N(0, \sigma^2_{C(T)})$ the effect of class
- With $(\alpha a)_{hj(i)} \sim N(0, \sigma^2_{SC(T)})$ the interaction effect of sex and class with type of class
- With $e_{hijk} \sim N(0, \sigma^2_R)$ the residual

Mean squares for balanced design:

- Means involved in mean squares:
  - Use the specific index for each term: H for sexes, I for class types, J for classes within type, K for children within class
  - Within brackets: use the index
  - Outside brackets: use index minus one
- Factors/terms involved in model
  - Sex $H-1$
  - Type $I-1$
  - Sex*Type $(H-1)(I-1)$
  - Class(Type) $(J-1)I$
  - Sex*Class(Type) $(H-1)(J-1)I$
  - Residual $(K-1)HIJ$

Three-way ANOVA models

Expected mean squares

Mean squares for balanced design:

- Sex $[H-1]$: $SS_S = \sum_{h=1}^{H} IJK(\bar{y}_{h...} - \bar{y}_{....})^2$
- Type $[I-1]$: $SS_S = \sum_{i=1}^{I} HJK(\bar{y}_{.i..} - \bar{y}_{....})^2$
- Sex*Type $[HI-H-I+1]$:

  $$SS_{ST} = \sum_{h=1}^{H}\sum_{i=1}^{I} JK(\bar{y}_{hi..} - \bar{y}_{h...} - \bar{y}_{.i..} + \bar{y}_{....})^2$$
- Class(Type) $[IJ-I]$:

  $$SS_{C(T)} = \sum_{i=1}^{I}\sum_{j=1}^{J} HK\left(\bar{y}_{.ij.} - \bar{y}_{.i..}\right)^2$$
- Sex*Class(Type) $[HIJ-HI-IJ+I]$:

  $$SS_{SC(T)} = \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J} K\left(\bar{y}_{hij.} - \bar{y}_{hi..} - \bar{y}_{.ij.} + \bar{y}_{.i..}\right)^2$$
- Residual $[HIJK-HIJ]$:

  $$SS_R = \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\bar{y}_{hijk} - \bar{y}_{hij.}\right)^2$$

Expected mean squares:

- Creating ANOVA table
- Assume first that all effects are random
  - Sources in rows
  - VC's in columns
- Cells in table are filled with sample sizes when the "letters" of the source fits within the "letters" of the VC's
- Cells in table are 0 when letters do not match
- Off diagonal filled cells are set to 0 for non-existing VC's

Three-way ANOVA models

Expected mean squares

Source	σ_S^2	σ_T^2	σ_{ST}^2	$\sigma_{C(T)}^2$	$\sigma_{SC(T)}^2$	σ_R^2
S	X		X		X	X
T		X	X	X	X	X
ST			X		X	X
$C(T)$				X	X	X
$SC(T)$					X	X
R						X

- All letters always appear in the residual
- Sample sizes are formed by the indices for the missing letters in the VC's
- VC's are replaced by quadratic functions of the fixed effects

Source	σ_S^2	σ_T^2	σ_{ST}^2	$\sigma_{C(T)}^2$	$\sigma_{SC(T)}^2$	σ_R^2
S	IJK		JK		K	1
T		HJK	JK	HK	K	1
ST			JK		K	1
$C(T)$				HK	K	1
$SC(T)$					K	1
R						1

Source	Q_S	Q_T	Q_{ST}	$\sigma_{C(T)}^2$	$\sigma_{SC(T)}^2$	σ_R^2
S	IJK				K	1
T		HJK		HK	K	1
ST			JK		K	1
$C(T)$				HK	K	1
$SC(T)$					K	1
R						1

Three-way ANOVA models

Moment-based estimators

Balanced data: S and T both fixed

- Expected mean squares are estimated with the mean squares
- Solving the equations
  - $\hat{\sigma}_R^2 = MS_R$
  - $\hat{\sigma}_{SC(T)}^2 = \left[MS_{SC(T)} - MS_R\right]/K$
  - $\hat{\sigma}_{C(T)}^2 = \left[MS_{C(T)} - MS_{SC(T)}\right]/[HK]$
- Fixed effects:
  - $\hat{\mu} = \bar{y}_{....}$
  - $\hat{\alpha}_h = \bar{y}_{h...} - \bar{y}_{....}$
  - $\hat{\beta}_i = \bar{y}_{.i..} - \bar{y}_{....}$

We assumed that

- $\alpha_1 + \cdots + \alpha_H = 0$
- $\beta_1 + \cdots + \beta_I = 0$

Balanced data: S and T both random

- Solving the equations
  - $\hat{\sigma}_R^2 = MS_R$
  - $\hat{\sigma}_{SC(T)}^2 = \frac{MS_{SC(T)} - MS_R}{K}$
  - $\hat{\sigma}_{C(T)}^2 = \frac{MS_{C(T)} - MS_{SC(T)}}{HK}$
  - $\hat{\sigma}_{ST}^2 = \frac{MS_{ST} - MS_{SC(T)}}{JK}$
  - $\hat{\sigma}_T^2 = \frac{MS_T - MS_{ST} - MS_{C(T)} + MS_{SC(T)}}{HJK}$
  - $\hat{\sigma}_S^2 = \frac{MS_S - MS_{ST}}{IJK}$
- Fixed effects: $\hat{\mu} = \bar{y}_{....}$

Three-way ANOVA models

Exercise A9

Theoretical questions:

- Consider a balanced three-way ANOVA model with factors S, T, and $C(T)$, with S fixed and T and $C(T)$ random
  - Determine the mean squares
  - Determine the expected mean squares
  - Determine the estimators for all model parameters
- Consider a balanced three-way ANOVA model with factors S, T, and $C(T)$, with T fixed and S and $C(T)$ random
  - Determine the mean squares
  - Determine the expected mean squares
  - Determine the estimators for all model parameters

Three-way ANOVA models

Hypothesis testing

Balanced data: S and T both random

- Null hypothesis $H_0: \sigma_T^2 = 0$
  - Under H_0, $\mathbb{E}[MS_T]$ can not be estimated with a single mean square other than MS_T
  - We need several mean squares
- Test statistic implemented in SAS:
  - $F_1 = MS_T/[MS_{ST} + MS_{C(T)} - MS_{SC(T)}]$
  - $(I-1)MS_T/\mathbb{E}[MS_T] \sim \chi^2_{I-1}$
  - Denominator $MS_{ST} + MS_{C(T)} - MS_{SC(T)}$ can be approximated with chi-square distribution using Satterthwaite
  - Means squares are independent, thus F_1 is approximately F-distributed under H_0
- Alternative test statistic
  - $F_2 = [MS_T + MS_{SC(T)}]/[MS_{ST} + MS_{C(T)}]$
  - $MS_T + MS_{SC(T)}$ can be approximated with chi-square distribution using Satterthwaite
  - $MS_{ST} + MS_{C(T)}$ can be approximated with chi-square distribution using Satterthwaite
  - Means squares are independent, thus F_2 is approximately F-distributed under H_0
- Advantages/disadvantages
  - F_2 both numerator and denominator are approximations, but good approximations
  - F_1 only denominator is approximated, but not a good approximation due to $-MS_{SC(T)}$

Three-way ANOVA models

Unbalanced data

Moment-based estimation:

- Mean squares remain independent, but are not chi-square distributed
- Both TYPE1 and TYPE3 are possible
  - TYPE3 less suitable for nested structures
- Fixed effects are based on generalized least squares estimates

Likelihood-based estimation:

- Both ML and REML are possible
  - ML estimates fixed effects jointly
  - REML is two step procedure
  - Fixed effects are similar to generalized least squares estimates

Case study: (outcome IQV)

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

Three-way ANOVA models

Unbalanced data

Fixed Effects:

Solution for Fixed Effects										
Effect	COMBI	GIRL	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept			11.5114	0.1322	406	87.07	<.0001	0.05	11.2515	11.7713
GIRL		0	0.2459	0.1344	446	1.83	0.0681	0.05	-0.01835	0.5101
GIRL		1	0	.	.	.	.	.	.	.
COMBI	0		0.3332	0.1580	342	2.11	0.0356	0.05	0.02250	0.6439
COMBI	1		0	.	.	.	.	.	.	.
COMBI*GIRL	0	0	-0.2114	0.1541	325	-1.37	0.1710	0.05	-0.5146	0.09176
COMBI*GIRL	0	1	0	.	.	.	.	.	.	.
COMBI*GIRL	1	0	0	.	.	.	.	.	.	.
COMBI*GIRL	1	1	0	.	.	.	.	.	.	.

- The average IQ for girls in multi-grade classes is 11.51 [11.25; 11.77]
- Boys have a 0.25 [-0.02; 0.51] larger IQ than girls in multi-grade classes
- Girls in single grade classes have a 0.33 [0.02; 0.64] larger IQ than girls in multi-grade classes
- Boys in single grade classes have a 0.21 [-0.09; 0.51] lower IQ than girls in multi-grade classes

Three-way ANOVA models

Unbalanced data

Random Effects:

Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
CLASS(COMBI)	0.5899	0.05	0.4492	0.8092
CLASS*GIRL(COMBI)	0.02350	0.05	0.003875	767.08
Residual	3.7998	0.05	3.6280	3.9841

Asymptotic Covariance Matrix of Estimates				
Row	Cov Parm	CovP1	CovP2	CovP3
1	CLASS(COMBI)	0.007758	-0.00079	-0.00027
2	CLASS*GIRL(COMBI)	-0.00079	0.001670	-0.00074
3	Residual	-0.00027	-0.00074	0.008234

- Between class variability seems clearly present
- The between class variability does not seem to be much different for boys and girls
- Total variability: $\sigma^2_{C(T)} + \sigma^2_{GC(T)} + \sigma^2_R = 4.41\ [4.19;\ 4.66]$
- Correlation children in one class and of the same sex:

  $$\frac{\sigma^2_{C(T)}+\sigma^2_{GC(T)}}{\sigma^2_{C(T)}+\sigma^2_{GC(T)}+\sigma^2_R} = 0.139\ [0.106;\ 0.174]$$

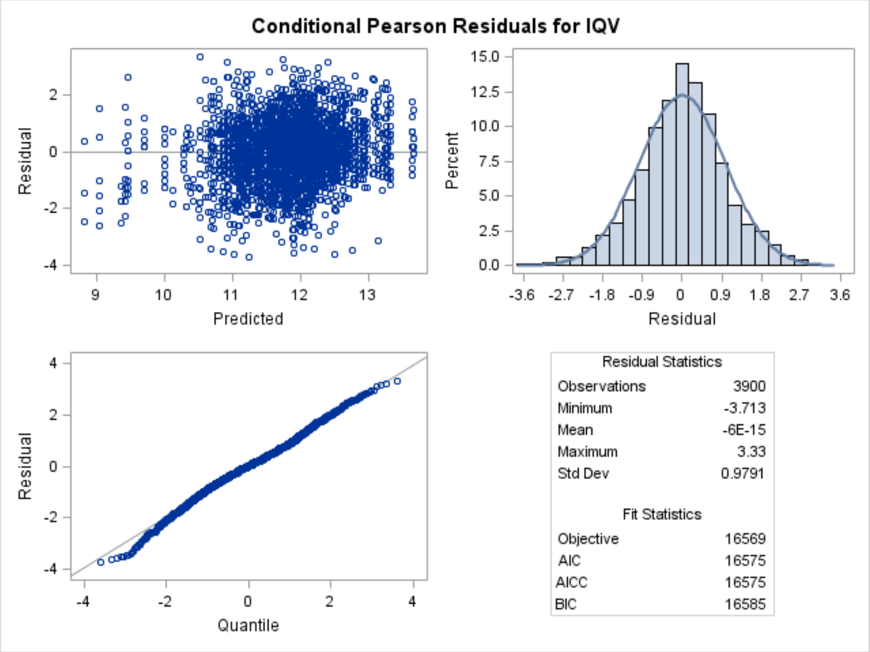
- Correlation children in one class and of different sex:

  $$\frac{\sigma^2_{C(T)}}{\sigma^2_{C(T)}+\sigma^2_{GC(T)}+\sigma^2_R} = 0.133\ [0.101;\ 0.169]$$

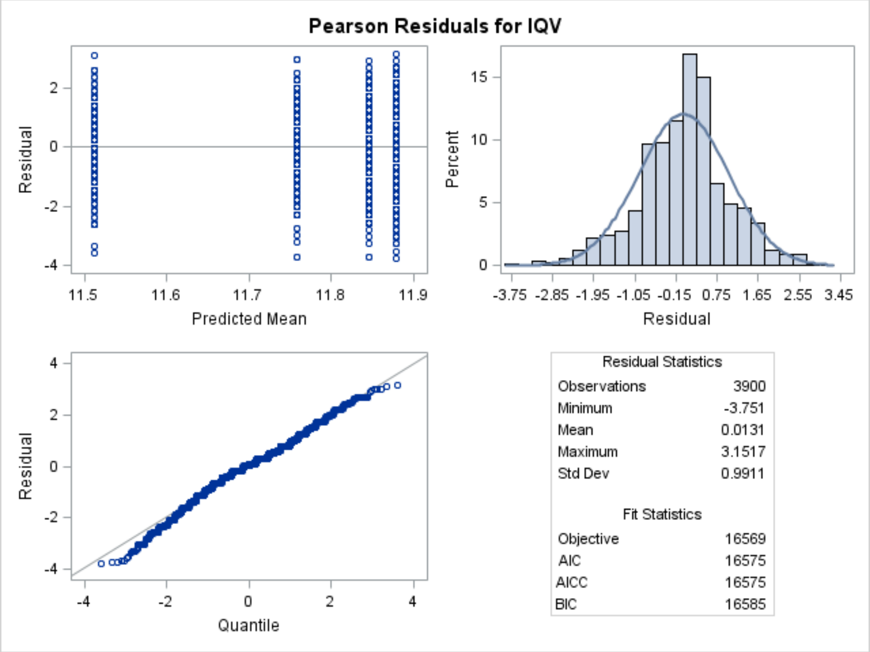
Three-way ANOVA models

Unbalanced data

Conditional residuals:



Marginal residuals



Three-way ANOVA models

Exercise A10

<u>Data analytical questions:</u>

- Fit a three-way ANOVA model for the difference in language score with sex, type of class, and class within type of class as factors
  - Report the parameters and their 95% confidence intervals
  - Calculate the total variability and the intraclass correlation coefficients with their 95% confidence intervals
  - Investigate the conditional and marginal residuals
  - Run all analysis for the four estimation methods: TYPE1, TYPE3, ML, REML and discuss the differences

Three-way ANOVA models

Exercise A11

<u>Theoretical questions:</u>

- Consider a balanced three-way ANOVA model with factors S (fixed), T (fixed), and $C(T)$ (random), but with no interactions
  - Determine the mean squares
  - Determine the expected mean squares
  - Determine the estimators for all model parameters

The delta method

- Let $\boldsymbol{X} = (X_1, X_2, \dots, X_m)^T$ be a vector of random variables
  - With $\mathbb{E}[X_k] = \mu_k$
  - With $\mathrm{COV}(X_k, X_l) = \eta_{kl}$
  - With $\mathrm{VAR}(X_k) = \eta_{kk} = \eta_k^2$
- Let $g: \mathbb{R}^m \to \mathbb{R}$ be a differentiable function
- Then what is $\mathrm{VAR}(g(\boldsymbol{X}))$?
  - Only under special cases (like linear functions of X_1, X_2, …, X_m) can we determine this variance exactly
  - Generally it requires an approximation
  - An multivariate form exists
- First order approximation

$$g(\boldsymbol{x}) \approx g(\boldsymbol{\mu}) + \sum_{k=1}^{m} g_k'(\boldsymbol{\mu})(x_k - \mu_k)$$

  - With $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m)^T$
  - With $\boldsymbol{x} = (x_1, x_2, \dots, x_m)^T$
  - With $g_k'(\boldsymbol{x}) = \partial g(\boldsymbol{x}) / \partial x_k$
- If the approximation is close we have

$$\mathbb{E}[g(\boldsymbol{X})] \approx g(\boldsymbol{\mu})$$

$$\mathbb{E}[g(\boldsymbol{X}) - g(\boldsymbol{\mu})]^2 \approx \sum_{k=1}^{m} \sum_{l=1}^{m} g_k'(\boldsymbol{\mu}) g_l'(\boldsymbol{\mu}) \eta_{kl} \equiv \eta^2$$

- The approximation is often warranted under asymptotics
  - If $\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \to N(0, \mathrm{H})$, then $\sqrt{n}(g(\boldsymbol{X}_n) - g(\boldsymbol{\mu})) \to N(0, \eta^2)$