Applied Analysis of Variance and Experimental Design

401-0625-01L Autumn Semester 2019

Introduction

Typically, data is collected to discover a **cause-effect relationship** of a (complex) “process” or a “system”. Ideally, we want to establish a causal relationship, i.e. we want to find out the effect on the **response** if we make an intervention on a **predictor**.

One distinguishes between **predictors** that

1. are of primary interest and that can be (ideally) varied according to our ”wishes”: the conditions we want to compare, or the ”treatments”.
2. are systematically recorded such that potential effects can be later eliminated in our calculations (”controlling for...”).
3. can be kept constant and whose effects can therefore be eliminated.
4. we can neither record nor keep constant.

2 to 4 are also called **nuisance variables**.

The **response** should be chosen such that it reflects useful information about the process under study. If not directly measurable, use a **surrogate** response.

Different types of **observational studies**:

- Cross-sectional study: ”Snapshot” of population at a given time-point.
- (Prospective) Cohort study: What will happen if...? For example: determining the risk (e.g. lung cancer) of exposed (smokers) vs. non-exposed (non- smokers) subjects (people).
- (Retrospective) Case-control study: Why did it develop this way? For example: comparison of habits of healthy vs. non-healthy persons.

It might very well be the case that some (hidden) predictors influence both the treatment “assignment” and the response, i.e. we have **confounders**.

An **experimental study** consists of

- Different treatments (the interventions you perform on the system)
- Experimental units, the “things” (“subjects”, “objects”) to which we apply the treatments by randomisation
- Method that assigns treatments to experimental units: randomisation, restricted randomisation (blocking)
- Response(s)

Experimental unit is what (the ”things”) we apply the treatments by randomisation. Rule: An experimental unit should be able to receive any treatment (independently of the others). **Measurement unit** is an actual “object” on which the response is measured. The measurement made on the whole experimental unit (potentially a sum or an average of the measurements on the measurement units) will be the basis of the analysis of the experiment.

Randomisation ensures that the only systematic difference between the groups is the treatment. It protects against confounding. This is why a (properly) randomised experiment allows us to make a statement about a causal effect. A **block** is a subset of the experimental units that is more homogenous than the entire set. Blocking increases precision of an experiment, because we use subsets of homogeneous units. General rule is: Block what you can; randomise what you cannot.

Experiments must be designed such that we have an estimate of this so called **experimental error**. This is achieved by using replicates, i.e. applying the same treatment to multiple experimental units.

Blinding is when evaluators don’t know which treatment is given to which experimental unit. With humans (patients): **double-blinding** such that neither the evaluators nor the patient know the assignment. It is an insurance against (unintentional) bias (e.g., due to expectations).

Control treatment is the “standard” treatment used as a baseline for comparison with other treatments. Sometimes “null” treatment (no treatment at all). **Placebo** is a null treatment in case that simply the act of applying a treatment (whatever) has an effect. Often used with humans, but can also be useful in other settings.

Completely Randomised Design (CRD) One-Way ANOVA

Standard errors for the parameters (using the sum of weighted treatment effects constraint)

Parameter	Estimator	Standard Error
μ	$\bar{y}_{\cdot\cdot}$	$\frac{\sigma}{\sqrt{N}}$
μ_i	$\bar{y}_{i\cdot}$	$\frac{\sigma}{\sqrt{n_i}}$
α_i	$\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$	$\sigma\sqrt{\frac{1}{n_i} - \frac{1}{N}}$
$\alpha_i - \alpha_j$	$\bar{y}_{i\cdot} - \bar{y}_{j\cdot}$	$\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

A 95% confidence interval for α_i is given by

$$\hat{\alpha}_i \pm t_{N-g}^{0.975} \cdot \hat{\sigma}\sqrt{\frac{1}{n_i} - \frac{1}{N}}.$$

Under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$, $F = \frac{MS_{\text{Trt}}}{MS_{\text{E}}} \approx 1$, because $E(MS_{\text{Trt}}) = \sigma^2 + \sum_{i=1}^{g} n_i \alpha_i^2/(g-1)$.

`QQ-Plot` plots empirical quantiles of residuals vs. theoretical quantiles (of standard normal distribution). Points should lie more or less on a straight line if residuals are normally distributed. In `R`, use `plot(fit, which = 2)`. Outliers can show up as isolated points in the “corners”.

Tukey-Anscombe Plot (TA-Plot) displays residuals vs. fitted values. Checks homogeneity of variance and systematic bias (here not relevant). In `R`, use `plot(fit, which = 1)`.

Index Plot plots residuals against time index to check for potential serial correlation (i.e., dependence with respect to time). Similarly for potential spatial dependence.

Fixing problems:

- Transformation of response (square root, logarithm, ...) to improve QQ-plot and constant variance assumption.
- Carefully inspect potential outliers.
- Deviation from normality less problematic for large sample sizes (reason: central limit theorem).
- Extend model (e.g., allow for some dependency structure, different variances, etc.)

Contrasts and Multiple Testing

A **contrast** $c \in \mathbb{R}^g$ is a vector that encodes the null hypothesis $H_0 : c^\top \mu = \sum_{i=1}^{g} c_i \mu_i = 0$, subject to the constaint $c^\top 1 = \sum_{i=1}^{g} c_i = 0$.

The side constraint ensures that the contrast is about differences between group means and not about the overall level of our response. Mathematically speaking, c is orthogonal to $(1, 1, \ldots, 1)$ or $1/g, 1/g, \ldots, 1/g)$, which is the overall mean.

Associated sum of squares with contrast c

$$SS_c = \frac{(\sum_{i=1}^{g} c_i \bar{y}_{i\cdot})^2}{\sum_{i=1}^{g} c_i^2/n_i}$$

has one degree of freedom. A $(1-\alpha)$ confidence interval for $\sum_{i=1}^{g} c_i \mu_i$ is given by

$$\sum_{i=1}^{g} c_i \bar{y}_{i\cdot} \pm t_{N-g}^{1-\alpha/2} \cdot \hat{\sigma}\sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}$$

Two contrasts c and c^* are **orthogonal** if $\sum_{i=1}^{n} c_i c_i^*/n_i = 0$. Orthogonal contrasts contain independent information. For orthogonal contrasts $c^{(1)}, c^{(2)}, \ldots, c^{(g-1)}$, it holds that

$$SS_{c^{(1)}} + SS_{c^{(2)}} + SS_{c^{(g-1)}} = SS_{\text{Trt}}.$$

Bonferroni-Holm adjustment for multiple testing: Sort p-values from small to large: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. For $j = 1, 2, \ldots, m$: reject null hypothesis if $p_{(j)} \leq \frac{\alpha}{m-j+1}$. Stop when reaching the *first* non-significant p-value. In `R`, use `p.adjust`, etc. This is a so called step-down procedure (“stepping-down the sequence of hypotheses”).

Scheffé procedure for multiple contrasts: It works because

$$\max_c \frac{SS_c/(g-1)}{MS_E} \leq \frac{MS_{\text{Trt}}}{MS_E} \sim F_{g-1,N-g}.$$

Calculate F-ratio as if ordinary contrast and use the distribution $(g-1) \cdot F_{g-1,N-g}$ instead of $F_{1,N-g}$ to calculate p-values or critical values. The price for searching for any possible contrast is low power.

Tukey Honest Significant Difference (HSD) uses the distribution of $\sqrt{n}(\max_i \bar{y}_{i\cdot} - \min_j \bar{y}_{j\cdot})/\sqrt{MS_{\mathrm{E}}}$ for critical values on all pair-wise t-tests. p-values are exact if design is balanced. Tukey HSD is more powerful than Bonferroni if all pairwise comparisons are of interest. If only a subset: Re-consider Bonferroni. In `R`, use `TukeyHSD` or package `multcomp`.

Dunnett procedure constructs simultaneous confidence intervals for the differences $\mu_i - \mu_g$, $i = 1, 2, \ldots, g-1$, assuming group g is the control group. In `R`, use package `multcomp`.

Can I only do pairwise comparisons etc. if the omnibus F-test is significant? No, although many textbooks recommend this. Conditioning on a significant F-test makes them over-conservative. Moreover, the conditional error or coverage rates can be (very) bad.

- Planned contrasts: Bonferroni
- All pairwise comparisons: Tukey HSD
- Comparison with a control: Dunnett
- Unplanned contrasts: Scheffé

Factorial Treatment Structure

Factorials in CRD when we assume that we have n replicates per combination of treatment effect A and treatment effect B.

Source	df
A	$a-1$
B	$b-1$
AB	$(a-1)(b-1)$
Error	$(n-1)ab$
Total	$abn-1$

For **individual analysis**, one can improve tests by "re-using" MS_{E} corresponding to the full model, which is more powerful since we have more df. In general, we can "plug in" the global σ^2 estimate with the correct df wherever appropriate.

For an additive model with **single replicates**, if there is an underlying interaction term, the error estimate is biased upwards, and tests will be conservative.

Transformations of the response help getting rid of interactions. For example, a multiplicative model can be log-transformed into a main effect model.

Tukey One-Degree of Freedom Interaction for the two-factor model use only one additional parameter for the interaction term

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + \epsilon_{ij},$$

where we might be interested in $H_0 : \lambda = 0$.

In the case of **unbalanced data**, parameters have to be estimated simultaneously using the principle of least squares (no problem for the computer). Similarly, sum of squares cannot be partitioned into different sources anymore. To deal with this problem, we can use model comparison approach in place of decomposition of sum of squares. We calculate the reduction of residual sum of squares when adding a new effect. The error sum of squares SS_{E} is typically taken from the full model.

- **Type I: Sequential** sum of squares. Sequentially build up model. Depends on ordering of factors.
- **Type II: Hierarchical** / **partially sequential** approach. Control for the influence of the largest hierarchical model not including the term of interest.
- **Type III: Fully adjusted** / **marginal** approach. Control for all other terms.

Type I	Type II	Type III
$SS(A\|1)$	$SS(A\|1,B)$	$SS(A\|1,B,AB)$
$SS(B\|1,A)$	$SS(B\|1,A)$	$SS(B\|1,A,AB)$
$SS(AB\|1,A,B)$	$SS(AB\|1,A,B)$	$SS(AB\|1,A,B)$
`aov`	`Anova` in `car`	`drop1`

For `drop1`, make sure that you'll use `options(contrasts = c("contr.sum", "contr.sum")`.

With balanced data, we always get the same result, no matter what type we use. For main effects only models, Type II and Type III coincide. If there is a significant interaction, tests of the corresponding main effects are typically difficult to interpret (better use individual models).

Complete Block Designs

A blocking design uses a **restricted randomisation** scheme. Each block gets its "own" randomisation.

We call a blocking design **complete** if every treatment is used in every block (every block contains all treatments). Therefore, we have no replicates (every combination of treatment and block is only observed once).

Factorials in CBD when we assume that we have one replicate per combination of block, treatment effect A and treatment effect B.

Source	df
Block	$r-1$
A	$a-1$
B	$b-1$
AB	$(a-1)(b-1)$
Error	$(ab-1)(r-1)$
Total	$rab-1$

If we want to have the same precision for treatment means, we have to ensure that $\frac{\sigma^2_{\mathrm{RCB}}}{r} = \frac{\sigma^2_{\mathrm{CRD}}}{n}$. **Relative efficiency** $RE = \hat{\sigma}^2_{\mathrm{CRD}}/\hat{\sigma}^2_{\mathrm{RCB}}$ gives us the ratio n/r. $\hat{\sigma}^2_{\mathrm{CRD}}$ can be estimated using a properly weighed average of MS_{E} and MS_{Block}: $\sigma^2_{\mathrm{CRD}} = w \cdot MS_{\mathrm{Block}} + (1-w) \cdot MS_{\mathrm{E}}$, where

$$w = \frac{r-1}{(r-1)+(g-1)+(r-1)(g-1)}.$$

Quick check: $MS_{\mathrm{Block}} > MS_{\mathrm{E}}$ is equivalent to $RE > 1$.

A **Latin Square** has two block factors. Each treatment (the Latin letters) appears exactly once in each row and exactly once in each column. It needs to have g treatments, two block factors having g levels each, and a total of g^2 experimental units. Fisher-Yates algorithm for picking a random Latin Square.

A **Graeco-Latin Squares** has three block factors. Both the Latin letters and the Greek letters occur once in each row and column. In addition, each Latin letter occurs exactly once with each Greek letter.

Typically, a block effect is assumed to be additive (i.e., main effects only). Block factors are not tested but they can be examined with respect to efficiency gain. Because we have used restricted randomisation, the block factor must be included in the model ("the analysis must follow the randomisation used in the experiment").

Random Effects

The more random effects two observations share, the larger the correlation. It is given by

$$\frac{\text{sum of \textbf{shared} variance components}}{\text{sum of \textbf{all} variance components}}.$$

Term	Fixed effect model	Random effect model
α_i	fixed, unknown constant	$\alpha_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\alpha)$
Side constraint on α_i	needed	not needed
$E(Y_{ij})$	$\mu + \alpha_i$	μ, but $E(Y_{ij}\|\alpha_i) = \mu + \alpha_i$
$\mathrm{Var}(Y_{ij})$	σ^2	$\sigma^2 + \sigma^2_\alpha$
$\mathrm{Corr}(Y_{ij}, Y_{kl})$	$0\ (j \neq l)$	$\begin{cases} 0, & i \neq k \\ \sigma^2_\alpha/(\sigma^2 + \sigma^2_\alpha), & i = k, j \neq l \\ 1, & i = k, j = l \end{cases}$

A note on the sampling mechanism. Fixed: Draw new random errors only, everything else is kept constant. Random: Draw new "treatment effects" and new random errors.

Hierarchy is typically less problematic in random effects models.

Could ask question about main effects even when interaction is present.

Restricted maximum-likelihood estimator (REML) is a modification of the maximum likelihood principle that removes bias in the estimation of the variance components. `lme4` and `lmerTest` allow to fit so called mixed effects models.

Variances are "difficult" to estimate in the sense that you'll need a lot of observations to have some reasonable accuracy. Approximate confidence intervals (or tests) can be obtained by calling the function `confint`. Exact tests (simulation based) for variance components can be found in the package `RLRsim`.

Nesting and Mixed Effects

Always ask yourself whether factor level "1" really corresponds to the same "object" across all levels of the other factor.

Errors are always nested.

Typically we use a nested structure due to practical / logistical constraints. For example: Patients are nested in hospitals as we don't want to send patients to all clinics across the country. Samples are nested in batches (in quality control).

We call a design **fully nested** if every factor is nested in its predecessor. **Fully nested design (balanced design)** leads to the decomposition

$$SS_{\text{Total}} = SS_A + SS_{B(A)} + SS_{C(AB)} + SS_{D(ABC)} + SS_{\text{E}}$$

with ANOVA table for random effects model

Source	df	$E(MS)$
A	$a-1$	$\sigma^2 + n\sigma_\delta^2 + nd\sigma_\gamma^2 + ncd\sigma_\beta^2 + nbcd\sigma_\alpha^2$
B	$a(b-1)$	$\sigma^2 + n\sigma_\delta^2 + nd\sigma_\gamma^2 + ncd\sigma_\beta^2$
C	$ab(c-1)$	$\sigma^2 + n\sigma_\delta^2 + nd\sigma_\gamma^2$
D	$abc(d-1)$	$\sigma^2 + n\sigma_\delta^2$
Error	$abcd(n-1)$	σ^2

F-tests are constructed by taking the ratio of "neighboring" mean squares as they just differ by the variance component of interest. For example, use $F = MS_A / MS_{B(A)}$ to test $H_0 : \sigma_\alpha^2 = 0$.

Two alternative `R` formulae for nested structure: `y ~ (1 | nesting/nested)` or `y ~ (1 | nesting) + (1 | nesting:nested)`.

Mixed effect model with interaction

$$Y_{ijk} = \mu + \underset{\text{fixed}}{\alpha_i} + \underset{\text{random}}{\beta_j} + \underset{\text{random}}{(\alpha\beta)_{ij}} + \epsilon_{ijk}$$

Random effect allows for an individual response level. Random interaction (between one random effect and one fixed effect) allows for an individual deviation from the population average of the fixed effect.

If we only consider a "main effect" model (no random interaction $(\alpha\beta)_{ij}$), then treatment effects α_i and corresponding F-test statistics, p-values should be the same regardless of whether β_j is fixed or random. Reason: We don't model a subject specific treatment effect.

If reference level constraint is used, the intercept of fixed effect has different interpretations: for fixed β_j, it corresponds to reference treatment, reference subject; for random β_j, it corresponds to reference treatment, expected value over all subjects.

p-value is usually larger than when treating β_j's and $(\alpha\beta)_{ij}$'s as fixed effects. What we observe in our data is a "contaminated" version (because every worker has its own individual deviation due to the random interaction term). Hence we do not have access to all observations of treatment α_i's but a summarised version over the levels of β_j's.

Split-Plot Designs

A **split-plot design** is a special case of a design with factorial treatment structure. It is used when some factors are harder (or more expensive) to vary than others. Basically, the standard split-plot design consists of two experiments with different experimental units of different "sizes".

The main effect of the whole-plot factor is estimated less precisely and the test is less powerful (compared to the split-plot level).

In the situation of **multi-step randomisation**, for each "level" (whole plot / split plot) of the experiment we have to introduce a corresponding random effect (better terminology here: error) which acts as the experimental error on that level.

Incomplete Block Designs

We call a design **disconnected** if we can build two groups of treatments such that it never happens that we see members of both groups together in the same block. If the design is not disconnected, we call it **connected**.

We call an incomplete block design **balanced (BIBD)** if all treatments pairs occur together in the same block equally often (we denote this number by λ). The precision (variance) of the estimated treatment differences $\alpha_i - \alpha_j$ is the same no matter what combination of i and j we are considering.

g	number of treatments
b	number of blocks
k	number of units per block with $k < g$
r	number of replicates per treatment
N	total number of units

A necessary condition (not sufficient) for the existence of a BIBD in the case where $g > k$ and $b < C_g^k$ is that $\lambda = r(k-1)/(g-1)$ must be an integer.

Use Type III sum of squares to test treatment effects adjusted for block effects. In other words, analyse the treatment effects while controlling for the block effects. Make sure to use `drop1` to ensure that the correct sum of squares is being used.

Modelling based on fixed block factors is called **intrablock analysis** of the (B)IBD. If based on random block factors, it is called an **interblock analysis.** This makes it possible to recover some information by comparing different blocks.

In the case where no BIBD is possible, we could use a **partially balanced** incomplete block design, where some treatment pairs occur together more often than other pairs.

A **row-column incomplete block design** is a design where we block on rows and columns and one or both of them are incomplete blocks.

A **Youden Square** is rectangular (!) such that columns (rows) form a BIBD, and each treatment appears equally often in each row (column). In short, columns (rows) form a BIBD, rows (columns) an RCB.