Design and Analysis of Experiments

Lecture notes

Part IV

Christoph Kopp

Bern University of Applied Sciences
School of Agricultural, Forest and Food Sciences HAFL

December 18, 2018

Contents

12 Extensions

The aim of this chapter is to give you the main idea (not a full analysis) of some important designs, and a pointer to alternative analysis methods.

12.1 Row-column designs

We start with Bailey 2008, Ex. 4.11, see also her Ch. 6 as well as Dean, Voss, and Draguljić 2017, Ch. 12 for more information.

Example 12.1. *Eight judges each taste and evaluate four wines. The order of the four tastings is randomized within each judge. A plot is one tasting by one judge, so there are 32 plots. A first analysis might only treat the judge as a block, but perhaps the tasters become more or less generous with increasing number of tastings, so perhaps the tasting order should also be used as a blocking variable.*

Designs with two systems of blocks are called *row-column* designs, one system of blocks is called "rows" and the other, "columns". The row and column factors are crossed (each combination of the two is observed). The most simple row-column designs satisfy the following conditions:

1. Each row meets each column in a single plot,
2. all treatments occur equally often in each row,
3. all treatments occur equally often in each column, and
4. there are m rows and n columns.

If these conditions are satisfied, there must be mn plots, the number of treatments t must be a divisor of m and n and each treatment is replicated mn/t times. Of course, it is also possible to have several plots for each combination of row and column blocking level.
For the statistical analysis, the blocks may be treated as having fixed or random effects depending on the context. If both blocking variables are modeled with random effects, we have *crossed random effects*.

12.1.1 Latin squares

The conditions given above are satisfied when $m = n = t$; such designs are called *Latin squares*. Latin squares (of order t) are an arrangement of t symbols in a $t \times t$ square such that each symbol occurs once in each row and once in each column. (Sudokus are order 9 Latin squares with additional local conditions for the small 3×3 squares.)

Example 12.2. *To compare the yield of four crops (A, B, C, D), four fields (I, II, III, IV) are available. Due to agronomic reasons, not every sequence of crops is sensible (crop rotation); the best rotation is A, B, C, D. Then, the following Latin square design could be used:*

	Field			
Year	*I*	*II*	*III*	*IV*
2018	*A*	*D*	*C*	*B*
2019	*B*	*A*	*D*	*C*
2020	*C*	*B*	*A*	*D*
2021	*D*	*C*	*B*	*A*

This Latin square satisfies the added crop rotation condition in each field over time.

A Latin square may be interpreted as a two-dimensional version of the randomized complete block design, since each blocking factor on its own defines an RCBD (ignoring the other factor). A number of designs is related to Latin squares.

Example 12.3. *In agronomy, field plots may be reused over several years. If in year one, a Latin square design was used, and the same treatments will be used on the same plots again in year two, then year two should be using an* orthogonal *Latin square to the first one. A pair of two Latin squares is called orthogonal if each letter of the first square occurs in the same position as each letter of the second square exactly once. (Not each Latin square has a Latin square orthogonal to it.) Two orthogonal Latin squares are often called* Graeco-Latin *squares because of the letters used:*

A	B	C
C	A	B
B	C	A

α	β	γ
β	γ	α
γ	α	β

Refer to Bailey 2008, Ch. 9.3 for further discussion.

Example 12.4. *Consider the following design with four subjects tasting and rating four types of chocolate and three runs:*

	Run		
Subject	*1*	*2*	*3*
Ari	*A*	*B*	*C*
Bernhard	*B*	*C*	*D*
Chris	*C*	*D*	*A*
Denise	*D*	*A*	*B*

Each subject tries only three brands. With respect to subjects, we have a balanced incomplete block design. In each run, each salmon will be tried by exactly one person, so that with respect to runs, a complete block design is used. Such a design is called a Youden *square (although it is not a square.) See Dean, Voss, and Draguljić 2017, Ch. 12 for further discussion.*

12.2 Confounded factorial experiments

We only give an example here to illustrate the main idea, see Dean, Voss, and Draguljić 2017, Ch. 13, 14.

Suppose we are in situation where each plot is very resource-intensive to run. We want to run a factorial experiment with three factors A, B, C, each of which has two levels, called 0 and 1. There are $2^3 = 8$ treatments (combinations of the three factors), which we write as ABC, substituting the values of the factors. For example, 010 means $A = 0, B = 1, C = 0$. Suppose we have only one plot per treatment (a single-replicate experiment), and the experiment is to be run in two blocks.

A short reminder about contrasts

Suppose we estimate any linear model of the kind

$$Y_{ji} = \mu + \beta_j + \varepsilon_{ji}\,.$$

Then any linear combination of the kind $\sum c_j\beta_j$ with $\sum c_j = 0$ is called a *contrast*. We have used contrasts e. g. in Section 3.1 for comparing the treatments with a control. To compare group 1 with group 2, we can use the contrast with $c_1 = -1, c_2 = 1$ and all other $c_i = 0$.

With 8 observations, we only have 7 degrees of freedom. Estimation of the main effects and interactions of the three factors costs 7 degrees of freedom, so no degrees of freedom are available for the error variance. Blocking makes the problem even worse, since estimating b block effects costs $b - 1$ degrees of freedom. As a result, $b - 1$ of the treatment contrasts can no longer be distinguished from block contrasts, they are said to be *confounded* with blocks.

The only way this can work is if we do not need to fit the full model. Suppose it is known that the factor A does not interact with any of the other factors. This means that we do not need the AB, AC, and the ABC interaction terms in the model, since we may assume that they are zero.

How would we have to combine the means of the eight treatments to estimate the contrasts? For example, for the A contrast, we have to subtract the means of the four treatments where $A = 0$ (these are 000, 001, 010, 011, the first four rows in the table below) from the mean of the four treatments where $A = 1$. Similarly for the main effects of B and C.

For the interaction effects, we may simply take the product of the corresponding columns. So, for AB, we multiply the values in the A column with the values in the B column. Similarly for the other two-way interaction effects (you can fill them in), including the three-way interaction.

	Contrast						
Treatment	A	B	C	AB	AC	BC	ABC
000	-1	-1	-1	1			-1
001	-1	-1	1	1			1
010	-1	1	-1	-1			1
011	-1	1	1	-1			-1
100	1	-1	-1	-1			1
101	1	-1	1	-1			-1
110	1	1	-1	1			-1
111	1	1	1	1			1

For the allocation of treatments to blocks, we have to choose one contrast which is confounded with the blocks. Let us choose ABC for this, for example. As you see in the table above, there are four treatments with ABC contrast equal to 1, and four with ABC contrast equal to -1. Using this to define blocks yields the following design:

Block 1	000	011	101	110
Block 2	001	010	100	111

As a result, the block difference is confounded with the ABC contrast. In other words, we had to sacrifice one contrast anyway and did it such that we now can not distinguish between the block difference and the ABC contrast. Since we are willing to believe that ABC is zero anyway, this is not a problem. Since we also assume that AB and AC are zero, we even have two degrees of freedom to estimate σ^2. Similar techniques are applicable in more complex settings.

12.3 Fractional factorial designs

We again give only the main idea and refer to Dean, Voss, and Draguljić 2017, Ch. 15 (whose treatment we follow) or to Bailey 2008, Ch. 15 for more. Especially in industry, for example in early stages of product development, there is often a large list of potentially important factors. It would be useful to find out which factors are the most important, for example to design further experiments. One approach to do this are *screening experiments*.

Let us focus here on the simplest case to explain the principle. Suppose you have p potentially important factors, each with exactly two levels. A factorial experiment thus involves 2^p treatments. In the most simple *fractional factorial experiment*, one instead only observes a fraction of one half of the treatments, i. e. 2^{p-1} treatments. This requires only half as many experimental units and can thus substantially reduce time, costs and workload required. But how to choose which treatments are omitted, and what is the price of this strategy?

Again, this is possible in case we are willing to make some compromises regarding the contrasts that we estimate. Suppose we run only Block 2 in the experiment in Section 12.2. Then the former contrasts look as follows:

Treatment	A	B	C	AB	AC	BC	ABC
001	-1	-1	1	1	-1	-1	1
010	-1	1	-1	-1	1	-1	1
100	1	-1	-1	-1	-1	1	1
111	1	1	1	1	1	1	1

Comparing the columns shows that $A = BC$, $B = AC$, and $C = AB$. The ABC column does not define a contrast ($\sum c_j \neq 0$), instead the weights corresponds to the mean. So ABC is confounded with the mean. This design is called a 2^{3-1} fractional factorial design and one often writes $I = ABC$ and calls ABC the *defining contrast* for the design; this contrast can not be estimated. The contrasts which are equal to each other (e. g. B and AC) are said to be *aliased*; they cannot be separated from each other by design. So one has to accept either that the corresponding contrast is a mixture of two effects or to assume that for example the interaction terms are equal to zero.

12.4 Alternative analysis approaches

Three main problems are common with regard to the assumptions underlying linear (mixed) models: outliers, heteroskedasticity, or non-normality of the residuals.
Because sample sizes are often small in the context of designed experiments, the problems above can have severe consequences and seriously distort the conclusions. We showed some countermeasures for simple designs in Chapter 4.
For more complex designs, it is sometimes difficult to find good alternative analysis methods in case the model assumptions are not fulfilled.
For independent observations (no plot structure), the *Brunner-Munzel* approach offers a *nonparametric* two- and higher-way factorial ANOVA for which works under much weaker conditions than classical ANOVA. To my knowledge, it is not yet implemented in an `R` package (but in SAS) but one can do this by application of other modeling functions in `R` as well.
The *Brunner-Langer* models (see Brunner, Domhof, and Langer 2002) are nonparametric techniques for *longitudinal data* and offer a way of dealing with temporally correlated data that does not rely on normality assumptions. There are a family of models for different situations. These models are implemented in the `nparLD` package.
We further refer to Wilcox 2012 for robust methods (in principle, any robust linear (mixed) model estimation technique can be used) to deal with heteroskedasticity or outliers.
Finally, Basso et al. 2009 discuss a permutation approach for two-way ANOVA. The permutation approach requires very little assumptions (which are often satisfied due to the randomization in experiments), but it only provides p values, not effect sizes.

13 Power and sample size calculations

Suppose we are planning an experiment and we know already how many treatments we use, we know about plot structure and also about the statistical model/test that we will use.

A short reminder about hypothesis tests

In the classical hypothesis test setting, we have a null hypothesis H_0, an alternative hypothesis H_1 and a hypothesis test to choose between the two hypotheses. In the setting of experiments, the null hypothesis is usually that some treatment has no effect (e. g. the new treatment is not better than placebo), and the alternative is that there is a treatment effect. The hypothesis test can make two types of errors: Type I denotes the probability of rejecting H_0 although it is true (false alarm), and Type II denotes the probability of not rejecting H_0 although it is false (false non-alarm). We focus on level α tests here, which have the property that they reject a true H_0 with a probability of at most α (the significance level). We let β denote the probability of a Type II error. The power $1 - \beta$ of the test is a lower bound for the probability of rejecting a false null hypothesis. Refer to the statistical inference notes for more details.

We fix the significance level α as well. Two related questions are then often asked:

1. How big is the power $1 - \beta$ of the test if we use n observations?
2. How many observations are required to achieve a given power with the test?

The first question is about the power, given the sample size, whereas the second question is about the required sample size to achieve a given power. If we can solve the first question for each n, we can then find the minimal n such that the power reaches some desired level.

Power or sample size computations are meant to be performed *before* the experiment.[1] They are extremely important for financial and ethical reasons, especially for research on living organisms. A too small sample will yield low power, so that an important effect may be missed. A too big sample uses unnecessarily many patients. (See also early stopping.)

As explained in the notes on statistical inference, the power of a particular test is a generally a function of the significance level, the sample size and the amount of disagreement between the null and the alternative hypothesis. Of course, it also depends

[1] Plugging in the actually observed values in your sample *after* conducting the study to try to find out how high the power was is called *post hoc* power computation and in general not a good idea. Seriously. Don't do this.

on the distribution of the data, and on any particular plot structure (e. g. strength of correlation inside a block).

> Before designing an experiment, it is advisable to consult a statistician to determine the statistical test, to compute the required sample size for the different groups such that an effect of a given size is found with a desired power, at a given significance level. Preliminary information on the group means, the common variance and any plot structure is required to carry out these computations.

To obtain this information, you could consult the literature or conduct a pilot study, or simply try plausible hypothetical scenarios.
Some scenarios can be looked up in the statistical literature. Some very basic settings are covered in the `pwr` package. For mixed models, some convenience packages exist. There also exists specialized commercial software for the required calculations for more complex designs. We show a very generally applicable approach here.

13.1 Monte Carlo approach

For the sake of having a concrete example, we focus on a randomized complete block design (not replicated) here. Suppose we have three treatments and we have an idea about the treatment means, the error variance and the block variance. If we fix the significance level, how many blocks should we observe to achieve a power of, say, 80%? Here is the idea:

1. Simulate a random sample under the alternative hypothesis.
2. Calculate the p value of the test.
3. Perform 1. and 2. a sufficient number of times count the proportion of significant results. Use this as Monte Carlo approximation to the power.

Below, we show implement this in a very simple way. As a little extra, we compare the power of the Friedman test with the power from the marginal F test of the linear mixed model with random block effect. The argument `runs` is the number of simulated samples.

```
> compare.sim <- function(n.block, n.treatments, means, sds, sd.block,
+                         runs, sig.level = 0.05) {
+     df <- cbind(block = paste0("B", rep(1:n.block, each = n.treatments)),
+                 treatment = paste0("T", rep(1:n.treatments)))
+     df <- data.frame(df)
+     p.val <- data.frame(matrix(NA, nrow = runs, ncol = 2))
+     names(p.val) <- c("Friedman.p.value", "Mixed.model.pvalue")
```

```
+     sig <- function(x) length(x[x < sig.level]) / length(x)
+     for (i in 1: dim(p.val)[1]) {
+         ## Simulate data
+         df$y <- rnorm(n = dim(df)[1],
+                       mean = rep(rnorm(n = n.block, mean = 0, sd = sd.block),
+                                  each = n.treatments) +
+                           rep(rep(means), times = n.block),
+                       sd = rep(rep(sds), times = n.block))
+         ## Calculate p values
+         p.val[i, ] <- c(friedman.test(y ~ treatment | block,
+                                       data = df)$p.value,
+                         anova(lme(y ~ treatment, random = ~ 1 | block,
+                                   data = df))$`p-value`[2])
+     }
+     return(apply(p.val, MARGIN = 2, sig))
+ }
```

To apply this now, we set the random number seed to get reproducible results and then pass the arguments. The treatment effects are reasonably strong. We start with a low number of runs here to keep runtime low. For more precise results, the number of runs should be increased.

```
> set.seed(17)
> compare.sim(n.block = 4, n.treatments = 3, means = c(1, 2.7, 3.2),
+             sds = c(1, 1, 1), sd.block = 1.5, runs = 100)

#   Friedman.p.value Mixed.model.pvalue
#               0.60                0.59
```

This suggests that four blocks provide a power of roughly 60% regardless of the test. Let us now find the required number of blocks to achieve a power of 80% using the Friedman test. This is where expensive software can make a difference with clever search strategies. We simply search on a grid until we are near the solution and then increase the number of runs. This is computationally inefficient, but this does not matter as long as we do not have to run too many power calculations.

```
> for (bl in 4:10) {
+     cat(paste("blocks: ", bl, sep = ""))
+     print(compare.sim(n.block = bl, n.treatments = 3, means = c(1, 2.7, 3.2),
+             sds = c(1, 1, 1), sd.block = 1.5, runs = 100))
+ }
```

```
# blocks: 4  Friedman.p.value Mixed.model.pvalue
#               0.62                0.70
# blocks: 5  Friedman.p.value Mixed.model.pvalue
#               0.67                0.77
# blocks: 6  Friedman.p.value Mixed.model.pvalue
#               0.84                0.95
# blocks: 7  Friedman.p.value Mixed.model.pvalue
#               0.86                0.94
# blocks: 8  Friedman.p.value Mixed.model.pvalue
#               0.86                0.99
# blocks: 9  Friedman.p.value Mixed.model.pvalue
#               0.93                0.98
# blocks: 10  Friedman.p.value Mixed.model.pvalue
#               0.96                1.00
```

It seems that six blocks are required. But 100 runs are not nearly enough, so we should increase the number of runs now to validate the results:

```
> for (bl in 5:7) {
+     cat(paste("blocks: ", bl, sep = ""))
+     print(compare.sim(n.block = bl, n.treatments = 3, means = c(1, 2.7, 3.2),
+             sds = c(1, 1, 1), sd.block = 1.5, runs = 5000))
+ }
```

Increasing the number of runs shows that six blocks actually only provide a power of roughly 76%, while seven block yield a power of about 81%, so seven blocks should be used. In practice, often a set of plausible scenarios is run (for example, with different effect sizes) to get an idea of the robustness of the results.
For this particular design, analytical results are available in case the data come from a normal distribution. As soon as we want to allow more flexibility, the analytical approach will have difficulties, while the Monte Carlo approach may still be applied, and not only to this design, but to any design.

13.2 Analytical results

As long as the data come from a normal distribution, there is no need for power simulations, one can usually find a formula for the power and simply plug in the parameters (usually involving the non-central F distribution). It is even possible to calculate the required sample size such that the confidence interval for a given contrast (say, treatments versus control) has a specified length.

The solutions are specific to the design used and are found in the literature, for example in Dean, Voss, and Draguljić 2017. If you have a standard design, then it is a good idea to consult the literature for the required sample size. As the saying goes, coding for one week can save you half a day in the library ...

13.3 Efficiency

In some situations, several designs would be possible, so it is interesting to compare them in statistical terms, especially in terms of their *relative efficiency*. We do not give the details here, see e. g. Oehlert 2010, Ch. 13.2.3, Bailey 2008, Ch. 11.6 or Hinkelmann and Kempthorne 2008, Ch. 9.3. The relative efficiency of design D_1 to design D_2 roughly tells us how many times more observations we would need to get statistically similar results if we used design D_2 instead of D_1. Efficiency may be used to help plan future experiments.

Complete vs. incomplete block designs

We use Bailey 2008, Ex. 11.6 to illustrate.

Example 13.1. *Suppose you have 7 treatments and 21 plots are available. Now you can either run the experiment as a complete block design, with three blocks and error variance σ^2_{RCBD}, or as a balanced incomplete block design with seven blocks, three observations per block and variance σ^2_{BIBD} (such a BIBD exists, as we showed in Chapter 9.2.1).*
Suppose blocks are days, and if you hurry, you may run all seven treatments in the lab in one day. But you will have to hurry, and use the whole day, so that moisture and temperature will perhaps fluctuate more. On the other hand, you could run the experiment on seven days and only observe three treatments per day. Thus, you need to hurry less and the conditions stay more constant because you use less than the whole day for the experiment. Is it possible that the smaller variance within a block could compensate the smaller efficiency of the BIBD relative to the RCBD?
Let us ignore degrees of freedom adjustments. In the above BIBD, the variance of each estimator of a difference between two treatments is $\frac{6}{7}\sigma^2_{BIBD}$. In the above RCBD, the variance of each estimator of a difference between two treatments is $\frac{2}{3} \cdot \sigma^2_{RCBD}$. Thus, the incomplete block design is preferred over the complete block design if and only if

$$\sigma^2_{BIBD} < \frac{7}{9}\sigma^2_{RCBD}\,.$$

If making the blocks smaller sufficiently reduces the variability, an incomplete block design can outperform a complete block design with the same number of observations. The references cited above contain more general information for comparing other designs.

14 Planning experiments

This chapter is essentially a copy of Bailey 2008, Ch. 14.3 with some additions from Dean, Voss, and Draguljić 2017, Ch. 2.2. The aim is to highlight important aspects in planning an experiment. The classical reference is Fisher 1949, a modern classic Box, Hunter, and Hunter 2005. A wealth of `R` specific information may be found at `http://stat.ethz.ch/CRAN/web/views/ExperimentalDesign.html`.
A research plan consists of a series of experiments used to investigate different aspects of some random phenomenon. Often, this involves several research questions building on each other. Here, we focus on the case of a single experiment.

14.1 The protocol

The aim is now to give a protocol that describes the experiment. Such a protocol is usually written in collaboration between the scientist and the statistician. At least the following questions should be addressed. Many of its elements can later be reused for a scientific publication.

14.1.1 What is the purpose of the experiment?

It is important to be precise here. For example,

- Is it about *estimating* the effect of a new fertilizer on the yield, compared to the standard?
- Is it the aim to *test* if whether the temperature and the moisture effect on the consistence of chocolate interact?
- Is it the aim to *model* the effect of the level of an drug on the outcome, while controlling for gender and age?
- Is it a screening experiment?
- Is it a dose-response study?

14.1.2 What are the treatments?

Here, the treatments are described in all technical details. How much of what is given how and when? Recall that treatments may be structured, for example they could have a factorial structure. Or treatment levels that correspond to a numeric variable could be equispaced. Any control treatments (or placebo) should be mentioned here. How many treatments are there?

14.1.3 What are the methods?

This is a non-statistical part that describes exactly how the treatments are applied and all that happens until the measurements are taken. The aim is that other scientists can replicate the study with these details. Are there several persons that apply the treatments? How are they allocated to experimental units?

14.1.4 What are the experimental units?

An exact description of the experimental units. Is there any relevant history (previous year study on the same plots?) Are there any relationships between the experimental units, such as relevant spatial or temporal proximity; are there any blocking factors? How many experimental units are there?

14.1.5 What are the observational units?

Are observational units the same as the experimental units? Then it should be mentioned; if it is not the case, there will most likely be several observational units per experimental unit. How many are there, and how are they defined in detail? All information about plot structure belongs here.

14.1.6 What measurements are going to be recorded?

What is going to be recorded? When does it happen, and to what precision? What are the measurement units? It makes sense to prepare a data sheet for the field and for the computer already at this stage. Each observational unit has a row and each measurement has a column. Are there several people involved in the measurements? Who measures which observational units? Are covariates/nuisance factors measured? How exactly?

14.1.7 What is the design?

Here, the design (what experimental unit gets which treatment) is given; in case a standard design is used, it is sufficient to give its name and the role of any factors. Special designs need to be described explicitly.
The amount of replication should be justified here, this is often very poorly done. Any pilot studies, sample size simulations, related experiments or such should be mentioned here. Blocking should be mentioned again here; how were blocks and block sizes chosen? Any relevant practical restrictions on the application of the treatments should be mentioned here. In case special designs with confounding or aliasing are used, the precise choice should be explained.

14.1.8 What randomization was used?

What method of randomization was used? You should always keep a record of the seed used in drawing random numbers.

14.1.9 What is the plan of the experiment?

The design has abstract treatments $T_1, T_2, \ldots$, what actual treatments were randomized to what explicitly named experimental units? In a field experiment, what plot obtained which actual treatment? Do not forget the north arrow on a map.

14.1.10 What statistical analysis is planned?

This needs to be decided anyway in order to calculate the required sample size. The more strata the design has, the more carefully this should be planned. Are the important factors measured at the proper level (think about split plots)? Are there enough degrees of freedom at each level? Do blocking factors have random or fixed effects, and in case of several factors, are they nested or crossed?
It is a very simple and good strategy to simulate some data in the planning stage of the experiment (for example, when simulating the sample size), then a proof-of-concept data analysis of the simulated data should reveal any systematic problems.
Is it planned to transform the data before analysis (due to experience in similar experiments)? It is not a problem to simulate heteroskedastic data and use this from the beginning.
(Many things may happen during the experiment that necessitate a change in the final analysis. For example, missings may turn the design into an unbalanced design, outliers may affect the results, and so on.)

References

R. A. Bailey (2008). *Design of Comparative Experiments*. Cambridge University Press.

D. Basso, F. Pesarin, L. Salmaso, and A. Solari (2009). *Permutation Tests for Stochastic Ordering and ANOVA*. Springer.

G. Beall (1942). The Transformation of Data from Entomological Field Experiments so that the Analysis of Variance Becomes Applicable. *Biometrika* 32, 243–262.

G. E. P. Box, J. S. Hunter, and W. G. Hunter (2005). *Statistics for Experimenters. Design, Innovation and Discovery*. 2nd ed. Wiley.

E. Brunner, S. Domhof, and F. Langer (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. Wiley.

R. A. Cribbie, R. R. Wilcox, C. Bewell, and H. J. Keselman (2007). Tests for Treatment Group Equality When Data are Nonnormal and Heteroscedastic. *Journal of Modern Applied Statistical Methods* 6, 117–132.

A. Dean, D. Voss, and D. Draguljić (2017). *Design and Analysis of Experiments*. 2nd ed. Springer.

R. A. Fisher (1949). *The Design of Experiments*. 5th ed. Oliver and Boyd.

J. Fox and S. Weisberg (2011). *An R Companion to Applied Regression*. 2nd ed. Sage.

K. Hinkelmann and O. Kempthorne (2008). *Design and Analysis of Experiments*. 2nd ed. Vol. 1. Wiley.

M. Hollander, D. A. Wolfe, and E. Chicken (2014). *Nonparametric Statistical Methods*. 3rd ed. Wiley.

G. A. Milliken and D. E. Johnson (2009). *Analysis of Messy Data*. 2nd ed. CRC Press.

G. W. Oehlert (2010). *A first course in design and analysis of experiments*.

J. Pinheiro and D. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

D. J. Saville and G. R. Wood (1991). *Statistical Methods: The Geometric Approach*. Springer.

W. Venables (1998). "Exegeses on Linear Models". In: *S-Plus User's Conference, Washington DC*.

W. Venables and B. Ripley (2002). *Modern Applied Statistics with S*. 4th ed. Springer.

R. R. Wilcox (2010). *Fundamentals of Modern Statistical Methods*. 2nd ed. Springer.

– (2012). *Introduction to Robust Estimation & Hypothesis Testing*. 3rd ed. Elsevier.

A Data set overview

Data set	Place	Topic
`InsectSprays`	Ch. 2	completely randomized design
`InsectSprays`	Ch. 6	randomized complete block design
`immer`	Ch. 6	Friedman test
`turnip`	Ch. 7	factorial designs
`detergent`	Ch. 9	incomplete designs
`Oats`	Ch. 10	split plot desings
`gasel`	Ch. 11	using covariates
`morley`	Graded Set 1	One-way ANOVA
`metal`	Graded Set 2	randomized complete block design
`catalyst`	Graded Set 3	balanced incomplete block design
`soybean`	Local Session 3	factorial designs
`paper`	Local Session 4	split plot designs
`warpbreaks`	Case study	factorial designs, transformations, post-hoc