

Statistical Science.

Part IV. The General Linear Model. Multiple Explanatory Variables.

Chapter 13.4 Fixed x Random Effects (Randomized Block)

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11)
ReCap Multiple Regression (Ch 12)
13.1 Fixed Effects ANOVA (no interactive effects)
13.2 Fixed Effects ANOVA (interactive effects)
13.3 Fixed*Random Effects (Paired t-test)
13.4 Fixed*Random Effects (Randomized Block)
13.5 Fixed*Random Effects (Repeated Measures)
13.6 Nested Random Effects (Hierarchical ANOVA)
13.7 Random within Fixed (Hierarchical ANOVA)
13.8 More Than Two Factors (to be written)

Tribolium growth data
Sokal and Rohlf Box 11.4
Ch13.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning is based on models, including statistical analysis based on models.

ReCap Part II (Chapters 5,6,7)

Hypothesis testing uses the logic of the null hypothesis to declare a decision.

Estimation is concerned with the specific value of an unknown population parameter.

ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.

ReCap (Ch 12) GLM with more than one regression variable (multiple regression)

ReCap (Ch 13) GLM with more than one categorical variable (ANOVA).

Two fixed factors (Ch 13.1, Ch13.2)

One fixed and one random factor (Paired t-test)

Today: Special case of Two way ANOVA: Randomized Blocks.
One factor fixed by design, the other factor is random.

Wrap-up.

The randomized block design is analyzed with a general linear model consisting of two explanatory variables on a nominal scale.

One of these is fixed (two or more classes),

the other is random (two or more classes).

We are interested in the fixed effects controlled for the random effects.

Introduction.

Research context.

The flour beetle *Tribolium castaneum* was a model lab organism for establishing the basic facts and principles of quantitative genetics, including inbreeding and response to selection. Once the whole genome was sequenced, *Tribolium* was used in immunohistochemistry, in situ hybridization, gene sequencing for characterization of microRNAs, and gene editing.

Economic context.

Tribolium is a major pest of stored grain. Economic losses consist of reduced weight and product quality, difficulties in baking, reduced marketability of infested products and an accompanying unpleasant smell. Numbers are reduced by sieving or by adding inert dusts that cause death by desiccation.

Statistical context.

This study illustrates a randomized block design, which has a fixed and a random factor. The randomized block is an example of statistical control, in which the effects of one variable (the random factor) are removed in order to arrive at a better analysis of the fixed factor. This analysis is a more sensitive test because it removes some of the noise in the data before testing. The paired t-test is a special case of the randomized block, in which the fixed factor has just two categories.

Statistical control is used when manipulative control is not possible.

Epidemiology. (manipulative control is unethical)

Many field situations.

Manipulative control is impossible at large scales.

Manipulative control can be expensive, even at small scales.

Manipulative control can generate artefacts

So a study with well designed statistical controls can be more informative.

1. Construct model

Data are from Box 11.4 in Sokal and Rohlf 1995, p 350.

Dry weights (mg) of 3 genotypes of *Tribolium castaneum* in 4 experiments.

Verbal model.

Does weight of flour beetle *Tribolium* vary among genotype, after controlling for differences among experiments ?

Each experiment is a block.

The analysis removes the effects of blocks. This produces a more sensitive test of whether weight varies among genotypes.

Blocks	Gtype	Wt
1	1	0.958
2	1	0.971
3	1	0.927
4	1	0.971
1	2	0.986
2	2	1.051
3	2	0.891
4	2	1.010
1	3	0.925
2	3	0.952
3	3	0.829
4	3	0.955

1. Construct model

Graphical model.

Response variable is beetle mass
 The first explanatory variables is genotype, a fixed factor under the control of the experimenter.
 The second explanatory variable is block, a random factor beyond the control of the investigator.

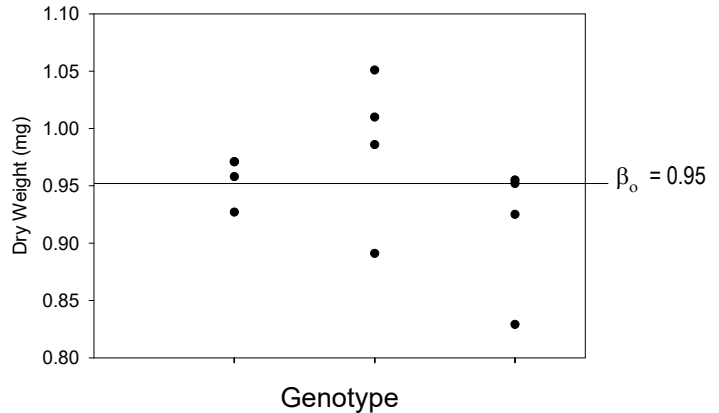


Table of variables

Symbol	Name	Fixed or Random?	Units	Type of measurement scale
M	beetle mass		mg	ratio
G	genotype (I II III)	Fixed	bb. +b, ++	categorical (nominal)
B	block	Random	Exp 1,2,3,4	categorical (nominal)

Formal model.

Write GLM: $M = \beta_0 + \beta_B \cdot B + \beta_G \cdot G + \beta_{G \times B} \cdot G \cdot B + \text{residual}$
 S&R81 $Y_{ijk} = \mu + B_j + \alpha_i + (\alpha \cdot B)_{ij} + \epsilon_{ij}$

Common practice is to use roman letters to distinguish random from fixed effects.
 Genotype is treatment effect, so Greek letter is used.
 Block is random effect so Roman letter is used.

Write full model on board,
 cross out the interaction term.

Full GLM: $M = \beta_0 + \beta_G \cdot G + \beta_B \cdot B + \beta_{G \times B} \cdot G \cdot B + \text{residual}$
 Revised GLM $M = \beta_0 + \beta_G \cdot G + \beta_B \cdot B + \text{residual}$

2. Execute analysis.

Place data in model format:

Column labelled M , with response variable mass

Column labelled X_B with explanatory variable, $X_B = 1,2,3$, or 4

These are labels (categories), not numbers on ratio scale.

Column labelled X_G with explanatory variable, $X_G = bb, +b, ++$

Code model statement in statistical package according to the GLM

$$M = \beta_0 + \beta_B \cdot X_B + \beta_G \cdot X_G + \varepsilon$$

```
MTB> ANOVA `M' = `XB' `XG'
MTB> GLM `M' = `XB' `XG'
SUBC> fits c4;
SUBC> res c5
```

The grand mean.

$$\hat{\beta}_0 = 12^{-1} \Sigma M = 12^{-1} \cdot 11.426 = 0.95217 \text{ mg}$$

	grand mean	Gtype Effect	Block Effect	Fits	Res
Wt					
0.958	0.952	0.00458	0.004	0.9609	-0.0029
0.971	0.952	0.00458	0.039	0.9959	-0.0249
0.927	0.952	0.00458	-0.070	0.8869	0.0401
0.971	0.952	0.00458	0.027	0.9833	-0.0123
0.986	0.952	0.03233	0.004	0.9887	-0.0027
1.051	0.952	0.03233	0.039	1.0237	0.0273
0.891	0.952	0.03233	-0.070	0.9147	-0.0237
1.010	0.952	0.03233	0.027	1.0110	-0.0010
0.925	0.952	-0.03692	0.004	0.9194	0.0056
0.952	0.952	-0.03692	0.039	0.9544	-0.0024
0.829	0.952	-0.03692	-0.070	0.8454	-0.0164
0.955	0.952	-0.03692	0.027	0.9418	0.0133

The fitted values are computed from the genotype and block means.

$$\begin{aligned} \text{mean}(M_{G++}) &= 0.95675 \text{ mg} \\ \hat{\beta}_{G++} &= (0.95675 - 0.95217) \\ &= 0.00458 \text{ mg} \\ &\text{etc.} \end{aligned}$$

The table shows calculations in spreadsheet format.

3. Evaluate the model.

Plot residuals versus fits.

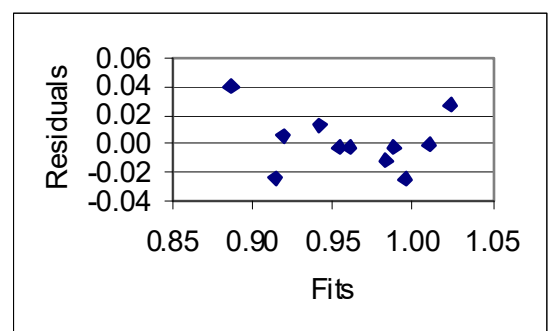
Structural model.

No line fitted in model, so skip this evaluation.

Errors homogeneous?

Residuals flare at high and low fitted values.

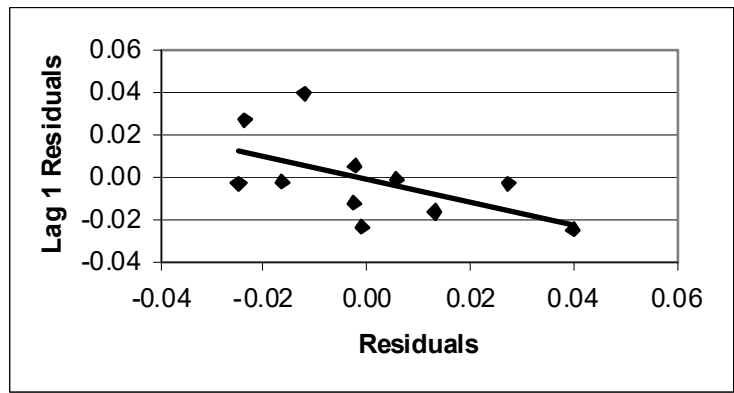
Symmetrical flaring is expected to have little effect on parameter estimates and p-values. We will check this with randomization.



3. Evaluate the model.

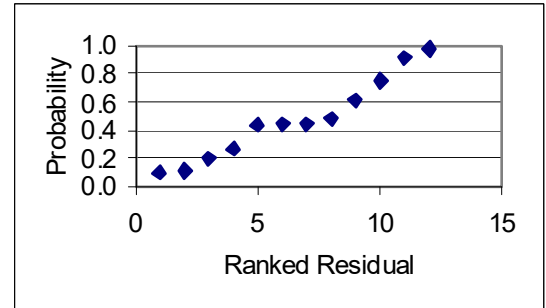
Errors independent?

The graph suggests a downward trend for residuals listed in order of blocks within genotype. There is a suspicious negative trend.



Errors normal ?

The residuals deviate somewhat from normal, as judged relative to the trend line in the normal probability plot. There is some indication of clustering of values around the median value of the residuals, The plot runs horizontally from the 5th ranked to the 8th ranked residual.



Conclusion: Errors deviate somewhat from normal,

4. Partition df and SS. Calculate LR

GLM	$M - \beta_o$	=	$\beta_B \cdot X_B$	+	$\beta_G \cdot X_G$	+	ϵ
Source	Total	=	Block	+	Genotype	+	Resid
df	12-1	=	4-1	+	3-1	+	12-3-2
SS	0.0353	=	0.021391	+	0.009717	+	0.004184

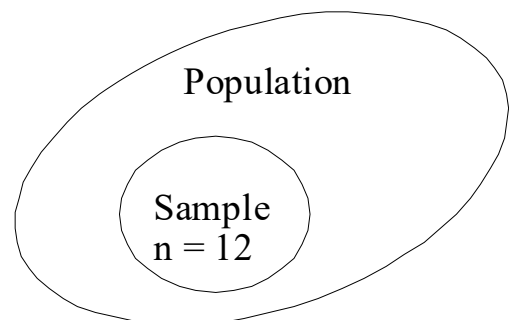
$$LR = (0.353/0.004184)1^{-12/2} = 3.6 \times 10^5$$

5. Choose mode of inference.

This text example was presented as an example of a randomized complete block design, which consists of all treatments in each block. Manipulative experiments are conducted with attention to control of variability. Control of Type I error is implicit in the analysis of such experiments.

5. State population and whether sample is representative.

The sample was taken from a population of beetles maintained by the investigator (R.R. Sokal). Inference is to the population of beetles in this lab and presumably to all beetles of this species, which share the same genetics. Inference is to the same chance set-up, measurement of weight of beetles.



6. State H_A H_0 pairs, test statistic, distribution, tolerance for Type I error.

Interaction term.

There is no H_A for the interaction term. The term is the product of a random and fixed term. This results in a mixed term, which is treated as random. To eliminate bias from this term, genotypes were weighed in random order (Sokal and Rohlf 1995, p351). Note that the interaction and error term cannot both be estimated.

Dropping the interaction

$df_{total} = 11 \quad df_G = 2 \quad df_B = 3 \quad df_{G \times B} = 2 * 3 = 6$ $df_{res} = 11 - 2 - 3 - 6 = 0$

term from the model flushes this component of variance to the residual, which now consists of a mixed term (which can be estimated) in addition to a residual term that cannot be estimated in this example.

Block term (experiment)

There is no H_A about this term. We are not interested in this effect. Instead, we want to estimate the variance component due to blocks and remove this variance from the error term, to produce a more sensitive test with a better chance of detecting main effects.

Experimental term

This is a fixed effect, the means are of interest.

$E(M_{bb})$ is the expected value (true mean) of the weight of genotype bb

$E(M_{+b})$ is the expected value (true mean) of the weight of genotype +b

$E(M_{++})$ is the expected value (true mean) of the weight of genotype ++

$H_0: E(M_{bb}) = E(M_{+b}) = E(M_{++})$

$H_A: \text{the means differ}$

The hypothesis pair above is equivalent to the following pair concerning variance.

$H_A: \text{Var}(\beta_{Genotype}) > 0$ There is variance due to experimental factor.

$H_0: \text{Var}(\beta_{Genotype}) = 0$

Are there more specific hypotheses about parameters? No

State test statistic	F-ratio
Distribution of test statistic	F-distribution
Tolerance for Type I error	5% (conventional level)

7. ANOVA - Calculate then partition df and SS according to model.

Model at top of board on left.
ANOVA table at top of board on right.

GLM	$M - \beta_0 = \beta_B \cdot X_B + \beta_G \cdot X_G + \varepsilon$
Source	Total = Block + Genotype + Resid
df	12-1 = 4-1 + 3-1 + 12-3-2
SS	0.0353 = 0.021391 + 0.009717 + 0.004184

$SS_{tot} = \text{Var}(M) \cdot df_{total} = 11 * \text{Var}(M) = 11 * 0.0032084 = 0.0353$

$SS_{tot} = \Sigma Y^2 - n^{-1}(\Sigma Y)^2 = 10.914748 - 12^{-1} \cdot 11.426^2 = 0.0353$

SS_{total} computed by Minitab.

```
MTB> let k1 = ssq('weight')
MTB> print k1
```

GLM commands in other packages perform in similar ways, to partition the variance.

GLM commands compute MS and variance ratio F. MS block was not computed, there is no interest in testing whether this term is significant. The interest is in estimating it.

Source	df	SS	MS	F---->	p
blocks	3	0.021391			
gtype	2	0.009717	0.004858	6.97	0.027
residual	6	0.004184	0.000697		
total	11	0.0353			

Calculate Type I error.

$p = 0.027$ calculated from F-distribution with $df = 2, 6$

Statistical control

Compare this partitioning to that when the Block term was not included in model

```
MTB > anova 'weights' = 'gtype'
Factor      Type Levels Values
gtype      fixed      3      1      2      3
Analysis of Variance for weights

Source      DF      SS      MS      F      P
gtype       2      0.009717  0.004859  1.71  0.235
Error       9      0.025575  0.002842
Total      11      0.035292  0.003208
```

- Compare SS SS_{error} shrinks from 0.0256 to 0.0042
- Compare MS MS_{error} shrinks from 0.02842 to 0.000697
- Compare F-ratio F-ratio increases from 1.71 to 6.97
- Compare Type I errors p-value shrinks from 0.235 to 0.027

7. ANOVA - Statistical control

Because the block effects are estimated and removed, the residual SS is much smaller. This allows smaller genotypic differences to be detected. Sokal and Rohlf 1995 (p 350), provide a calculation of the increased efficiency of the randomized block design. Reducing the error variance via statistical control is one of the key concepts of experimental design.

8. Decide whether to recompute p-value.

Residuals were symmetrically heterogeneous, perhaps not independent, and slightly deviant from a normal distribution.

Sample size n was small, but $p = 0.027$ and hence factor 2-fold increase would change our assessment of Type I error in the gray zone from 0.1 to 0.01.

We recompute the p-value.

The p-value via randomization in this case is $128/5000 = 0.0256$

The p-value changed by a factor of $0.027 / 0.0256 = 1.05$.

The deviations from perfectly homogeneous, normal, and independent did not distort the estimate of Type I error (p-value).

Having computed the p-value based on randomization, we report it because it is free of assumption concerning the residuals.

9. Report statistical conclusion.

Only one term, the fixed factor, is tested.

We reject the null hypothesis, H_0 of no difference.

There was significant variation in mean dry weight among genotypes.

$$F_{2,6} = 6.97 \quad p = 0.027$$

No parameters are reported for the block term because it is a random factor and so the means are of no interest.

The differences in weight among genotypes are small and not detectable with these standard errors.

mean(M_{bb}) = 0.957 mg	st.err = 0.0104 mg
mean(M_{+b}) = 0.985 mg	st.err = 0.0339 mg
mean(M_{++}) = 0.915 mg	st. err = 0.0295 mg

These differences become detectable (as in the ANOVA table) when variation among experiments (blocks) is removed from error term.

10. Report science conclusion.

The differences among means were small. At the same time they were greater than those from chance, once we control for among block variance. Because of the substantial variation among experiments, statistical control was necessary to detect differences in dry weight among genotypes.