

2816-9091

Modern Statistics for the Life Sciences

Alan Grafen

Rosie Hails

University of Oxford

OXFORD
UNIVERSITY PRESS

Contents

<i>Why use this book</i>	xi
How to use this book	xii
How to teach this text	xiv
1 An introduction to analysis of variance	1
1.1 Model formulae and geometrical pictures	1
1.2 General Linear Models	1
1.3 The basic principles of ANOVA	2
What happens when we calculate a variance?	3
Partitioning the variability	4
Partitioning the degrees of freedom	8
F-ratios	9
1.4 An example of ANOVA	10
Presenting the results	14
1.5 The geometrical approach for an ANOVA	16
1.6 Summary	19
1.7 Exercises	20
Melons	20
Dioecious trees	21
2 Regression	22
2.1 What kind of data are suitable for regression?	22
2.2 How is the best fit line chosen?	23
2.3 The geometrical view of regression	26
2.4 Regression—an example	28
2.5 Confidence and prediction intervals	33
Confidence intervals	33
Prediction intervals	33
2.6 Conclusions from a regression analysis	35
A strong relationship with little scatter	35
A weak relationship with lots of noise	36
Small datasets and pet theories	38
Significant relationships—but that is not the whole story	39
2.7 Unusual observations	40
Large residuals	40
Influential points	41
2.8 The role of X and Y—does it matter which is which?	42
2.9 Summary	45

2.10	Exercises	45
	Does weight mean fat?	45
	Dioecious trees	46
3	Models, parameters and GLMs	47
3.1	Populations and parameters	47
3.2	Expressing all models as linear equations	48
3.3	Turning the tables and creating datasets	52
	Influence of sample size on the accuracy of parameter estimates	54
3.4	Summary	55
3.5	Exercises	55
	How variability in the population will influence our analysis	55
4	Using more than one explanatory variable	56
4.1	Why use more than one explanatory variable?	56
	Leaping to the wrong conclusion	56
	Missing a significant relationship	57
4.2	Elimination by considering residuals	59
4.3	Two types of sum of squares	61
	Eliminating a third variable makes the second less informative	62
	Eliminating a third variable makes the second more informative	64
4.4	Urban Foxes—an example of statistical elimination	65
4.5	Statistical elimination by geometrical analogy	68
	Partitioning and more partitioning	68
	Picturing sequential and adjusted sums of squares	71
4.6	Summary	72
4.7	Exercises	73
	The cost of reproduction	73
	Investigating obesity	75
5	Designing experiments—keeping it simple	76
5.1	Three fundamental principles of experimental design	76
	Replication	76
	Randomisation	78
	Blocking	80
5.2	The geometrical analogy for blocking	85
	Partitioning two categorical variables	85
	Calculating the fitted model for two categorical variables	86
5.3	The concept of orthogonality	88
	The perfect design	88
	Three pictures of orthogonality	91
5.4	Summary	92
5.5	Exercises	93
	Growing carnations	93
	The dorsal crest of the male smooth newt	95
6	Combining continuous and categorical variables	96
6.1	Reprise of models fitted so far	96
6.2	Combining continuous and categorical variables	97
	Looking for a treatment for leprosy	97
	Sex differences in the weight—fat relationship	99

6.3	Orthogonality in the context of continuous and categorical variables	102
6.4	Treating variables as continuous or categorical	104
6.5	The general nature of General Linear Models	106
6.6	Summary	107
6.7	Exercises	108
	Conservation and its influence on biomass	108
	Determinants of the Grade Point Average	109
7	Interactions—getting more complex	110
7.1	The factorial principle	110
7.2	Analysis of factorial experiments	112
7.3	What do we mean by an interaction?	115
7.4	Presenting the results	117
	Factorial experiments with insignificant interactions	117
	Factorial experiments with significant interactions	120
	Error bars	123
7.5	Extending the concept of interactions to continuous variables	127
	Mixing continuous and categorical variables	127
	Adjusted Means (or least square means in models with continuous variables)	129
	Confidence intervals for interactions	130
	Interactions between continuous variables	131
7.6	Uses of interactions	132
	Is the story simple or complicated?	133
	Is the best model additive?	133
7.7	Summary	134
7.8	Exercises	134
	Antidotes	134
	Weight, fat and sex	135
8	Checking the models I: independence	136
8.1	Heterogeneous data	137
	Same conclusion within and between subsets	140
	Creating relationships where there are none	140
	Concluding the opposite	141
8.2	Repeated measures	142
	Single summary approach	142
	The multivariate approach	145
8.3	Nested data	147
8.4	Detecting non-independence	148
	Germination of tomato seeds	149
8.5	Summary	151
8.6	Exercises	151
	How non-independence can inflate sample size enormously	151
	Combining data from different experiments	152
9	Checking the models II: the other three assumptions	153
9.1	Homogeneity of variance	153
9.2	Normality of error	155
9.3	Linearity/additivity	157

9.4	Model criticism and solutions	157
	Histogram of residuals	158
	Normal probability plots	160
	Plotting the residuals against the fitted values	163
	Transformations affect homogeneity and normality simultaneously	166
	Plotting the residuals against each continuous explanatory variable	167
	Solutions for nonlinearity	168
	Hints for looking at residual plots	172
9.5	Predicting the volume of merchantable wood: an example of model criticism	173
9.6	Selecting a transformation	178
9.7	Summary	180
9.8	Exercises	181
	Stabilising the variance	181
	Stabilising the variance in a blocked experiment	181
	Lizard skulls	183
	Checking the 'perfect' model	184
10	Model selection I: principles of model choice and designed experiments	186
10.1	The problem of model choice	186
10.2	Three principles of model choice	189
	Economy of variables	189
	Multiplicity of p -values	191
	Considerations of marginality	192
	Model choice in the polynomial problem	193
10.3	Four different types of model choice problem	195
10.4	Orthogonal and near orthogonal designed experiments	196
	Model choice with orthogonal experiments	196
	Model choice with loss of orthogonality	198
10.5	Looking for trends across levels of a categorical variable	201
10.6	Summary	205
10.7	Exercises	206
	Testing polynomials requires sequential sums of squares	206
	Partitioning a sum of squares into polynomial components	207
11	Model selection II: datasets with several explanatory variables	209
11.1	Economy of variables in the context of multiple regression	210
	R -squared and adjusted R -squared	210
	Prediction Intervals	213
11.2	Multiplicity of p-values in the context of multiple regression	217
	The enormity of the problem	217
	Possible solutions	217
11.3	Automated model selection procedures	220
	How stepwise regression works	220
	The stepwise regression solution to the whale watching problem	221
11.4	Whale Watching: using the GLM approach	225
11.5	Summary	228
11.6	Exercises	229
	Finding the best treatment for cat fleas	229
	Multiplicity of p -values	231

12	Random effects	232
12.1	What are random effects?	232
	Distinguishing between fixed and random factors	232
	Why does it matter?	234
12.2	Four new concepts to deal with random effects	234
	Components of variance	234
	Expected mean square	235
	Nesting	236
	Appropriate Denominators	237
12.3	A one-way ANOVA with a random factor	238
12.4	A two-level nested ANOVA	241
	Nesting	241
12.5	Mixing random and fixed effects	244
12.6	Using mock analyses to plan an experiment	247
12.7	Summary	252
12.8	Exercises	253
	Examining microbial communities on leaf surfaces	253
	How a nested analysis can solve problems of non-independence	254
13	Categorical data	255
13.1	Categorical data: the basics	255
	Contingency table analysis	255
	When are data truly categorical?	257
13.2	The Poisson distribution	258
	Two properties of a Poisson process	258
	The mathematical description of a Poisson distribution	259
	The dispersion test	261
13.3	The chi-squared test in contingency tables	265
	Derivation of the chi-squared formula	265
	Inspecting the residuals	267
13.4	General linear models and categorical data	269
	Using contingency tables to illustrate orthogonality	269
	Analysing by contingency table and GLMs	271
	Omitting important variables	276
	Analysing uniformity	277
13.5	Summary	278
13.6	Exercises	279
	Soya beans revisited	279
	Fig trees in Costa Rica	280
14	What lies beyond?	281
14.1	Generalised Linear Models	281
14.2	Multiple y variables, repeated measures and within-subject factors	283
14.3	Conclusions	284
15	Answers to exercises	285
	Chapter 1	285
	Chapter 2	287
	Chapter 3	288

Chapter 4	289
Chapter 5	292
Chapter 6	294
Chapter 7	295
Chapter 8	298
Chapter 9	299
Chapter 10	308
Chapter 11	310
Chapter 12	313
Chapter 13	314
Revision section: The basics	317
R1.1 Populations and samples	317
R1.2 Three types of variability: of the sample, the population and the estimate	318
Variability of the sample	318
Variability of the population	319
Variability of the estimate	319
R1.3 Confidence intervals: a way of precisely representing uncertainty	322
R1.4 The null hypothesis—taking the conservative approach	324
R1.5 Comparing two means	327
Two sample <i>t</i> -test	327
Alternative tests	328
One and two tailed tests	329
R1.6 Conclusion	331
Appendix 1: The meaning of <i>p</i>-values and confidence intervals	332
What is a <i>p</i> -value?	332
What is a confidence interval?	334
Appendix 2: Analytical results about variances of sample means	335
Introducing the basic notation	335
Using the notation to define the variance of a sample	335
Using the notation to define the mean of a sample	336
Defining the variance of the sample mean	336
To illustrate why the sample variance must be calculated with $n - 1$ in its denominator (rather than n) to be an unbiased estimate of the population variance	337
Appendix 3: Probability distributions	339
Some gentle theory	339
Confirming simulations	341
Bibliography	343
Index	345