Statistics Explained

An Introductory Guide for Life Scientists

SECOND EDITION

Steve McKillup
Central Queensland University,
Rockhampton

CAMBRIDGE
UNIVERSITY PRESS

Contents

	Preface	*page* xiii
1	**Introduction**	1
1.1	Why do life scientists need to know about experimental design and statistics?	1
1.2	What is this book designed to do?	5
2	**Doing science: hypotheses, experiments and disproof**	7
2.1	Introduction	7
2.2	Basic scientific method	7
2.3	Making a decision about an hypothesis	11
2.4	Why can't an hypothesis or theory ever be proven?	11
2.5	'Negative' outcomes	12
2.6	Null and alternate hypotheses	12
2.7	Conclusion	14
2.8	Questions	14
3	**Collecting and displaying data**	15
3.1	Introduction	15
3.2	Variables, experimental units and types of data	15
3.3	Displaying data	17
3.4	Displaying ordinal or nominal scale data	23
3.5	Bivariate data	25
3.6	Multivariate data	26
3.7	Summary and conclusion	28
4	**Introductory concepts of experimental design**	29
4.1	Introduction	29
4.2	Sampling – mensurative experiments	30
4.3	Manipulative experiments	34
4.4	Sometimes you can only do an unreplicated experiment	41

4.5	Realism	42
4.6	A bit of common sense	43
4.7	Designing a 'good' experiment	44
4.8	Reporting your results	45
4.9	Summary and conclusion	46
4.10	Questions	46
5	**Doing science responsibly and ethically**	48
5.1	Introduction	48
5.2	Dealing fairly with other people's work	48
5.3	Doing the experiment	50
5.4	Evaluating and reporting results	52
5.5	Quality control in science	53
5.6	Questions	54
6	**Probability helps you make a decision about your results**	56
6.1	Introduction	56
6.2	Statistical tests and significance levels	57
6.3	What has this got to do with making a decision about your results?	60
6.4	Making the wrong decision	60
6.5	Other probability levels	61
6.6	How are probability values reported?	62
6.7	All statistical tests do the same basic thing	63
6.8	A very simple example – the chi-square test for goodness of fit	64
6.9	What if you get a statistic with a probability of exactly 0.05?	66
6.10	Statistical significance and biological significance	67
6.11	Summary and conclusion	69
6.12	Questions	70
7	**Probability explained**	71
7.1	Introduction	71
7.2	Probability	71
7.3	The addition rule	71

7.4	The multiplication rule for independent events	72
7.5	Conditional probability	75
7.6	Applications of conditional probability	77
8	**Using the normal distribution to make statistical decisions**	87
8.1	Introduction	87
8.2	The normal curve	87
8.3	Two statistics describe a normal distribution	89
8.4	Samples and populations	93
8.5	The distribution of sample means is also normal	95
8.6	What do you do when you only have data from one sample?	99
8.7	Use of the 95% confidence interval in significance testing	102
8.8	Distributions that are not normal	102
8.9	Other distributions	103
8.10	Other statistics that describe a distribution	105
8.11	Summary and conclusion	106
8.12	Questions	106
9	**Comparing the means of one and two samples of normally distributed data**	108
9.1	Introduction	108
9.2	The 95% confidence interval and 95% confidence limits	108
9.3	Using the Z statistic to compare a sample mean and population mean when population statistics are known	108
9.4	Comparing a sample mean to an expected value when population statistics are not known	112
9.5	Comparing the means of two related samples	116
9.6	Comparing the means of two independent samples	118
9.7	One-tailed and two-tailed tests	121
9.8	Are your data appropriate for a t test?	124
9.9	Distinguishing between data that should be analysed by a paired sample test and a test for two independent samples	125
9.10	Reporting the results of t tests	126
9.11	Conclusion	127
9.12	Questions	128

10	**Type 1 error and Type 2 error, power and sample size**	130
10.1	Introduction	130
10.2	Type 1 error	130
10.3	Type 2 error	131
10.4	The power of a test	135
10.5	What sample size do you need to ensure the risk of Type 2 error is not too high?	135
10.6	Type 1 error, Type 2 error and the concept of biological risk	136
10.7	Conclusion	138
10.8	Questions	139
11	**Single-factor analysis of variance**	140
11.1	Introduction	140
11.2	The concept behind analysis of variance	141
11.3	More detail and an arithmetic example	147
11.4	Unequal sample sizes (unbalanced designs)	152
11.5	An ANOVA does not tell you which particular treatments appear to be from different populations	153
11.6	Fixed or random effects	153
11.7	Reporting the results of a single-factor ANOVA	154
11.8	Summary	154
11.9	Questions	155
12	**Multiple comparisons after ANOVA**	157
12.1	Introduction	157
12.2	Multiple comparison tests after a Model I ANOVA	157
12.3	An *a posteriori* Tukey comparison following a significant result for a single-factor Model I ANOVA	160
12.4	Other *a posteriori* multiple comparison tests	162
12.5	Planned comparisons	162
12.6	Reporting the results of *a posteriori* comparisons	164
12.7	Questions	166
13	**Two-factor analysis of variance**	168
13.1	Introduction	168
13.2	What does a two-factor ANOVA do?	170

13.3	A pictorial example	174
13.4	How does a two-factor ANOVA separate out the effects of each factor and interaction?	176
13.5	An example of a two-factor analysis of variance	180
13.6	Some essential cautions and important complications	181
13.7	Unbalanced designs	192
13.8	More complex designs	192
13.9	Reporting the results of a two-factor ANOVA	193
13.10	Questions	194
14	**Important assumptions of analysis of variance, transformations, and a test for equality of variances**	196
14.1	Introduction	196
14.2	Homogeneity of variances	196
14.3	Normally distributed data	197
14.4	Independence	201
14.5	Transformations	201
14.6	Are transformations legitimate?	203
14.7	Tests for heteroscedasticity	204
14.8	Reporting the results of transformations and the Levene test	205
14.9	Questions	207
15	**More complex ANOVA**	209
15.1	Introduction	209
15.2	Two-factor ANOVA without replication	209
15.3	*A posteriori* comparison of means after a two-factor ANOVA without replication	214
15.4	Randomised blocks	214
15.5	Repeated-measures ANOVA	216
15.6	Nested ANOVA as a special case of a single-factor ANOVA	222
15.7	A final comment on ANOVA – this book is only an introduction	229
15.8	Reporting the results of two-factor ANOVA without replication, randomised blocks design, repeated-measures ANOVA and nested ANOVA	229
15.9	Questions	230

16	**Relationships between variables: correlation and regression**	233
16.1	Introduction	233
16.2	Correlation contrasted with regression	234
16.3	Linear correlation	234
16.4	Calculation of the Pearson *r* statistic	235
16.5	Is the value of *r* statistically significant?	241
16.6	Assumptions of linear correlation	241
16.7	Summary and conclusion	242
16.8	Questions	242
17	**Regression**	244
17.1	Introduction	244
17.2	Simple linear regression	244
17.3	Calculation of the slope of the regression line	246
17.4	Calculation of the intercept with the *Y* axis	249
17.5	Testing the significance of the slope and the intercept	250
17.6	An example – mites that live in the hair follicles	258
17.7	Predicting a value of *Y* from a value of *X*	260
17.8	Predicting a value of *X* from a value of *Y*	260
17.9	The danger of extrapolation	262
17.10	Assumptions of linear regression analysis	263
17.11	Curvilinear regression	266
17.12	Multiple linear regression	273
17.13	Questions	281
18	**Analysis of covariance**	284
18.1	Introduction	284
18.2	Adjusting data to remove the effect of a confounding factor	285
18.3	An arithmetic example	288
18.4	Assumptions of ANCOVA and an extremely important caution about parallelism	289
18.5	Reporting the results of ANCOVA	295
18.6	More complex models	296
18.7	Questions	296

19	**Non-parametric statistics**	298
19.1	Introduction	298
19.2	The danger of assuming normality when a population is grossly non-normal	298
19.3	The advantage of making a preliminary inspection of the data	300
20	**Non-parametric tests for nominal scale data**	301
20.1	Introduction	301
20.2	Comparing observed and expected frequencies: the chi-square test for goodness of fit	302
20.3	Comparing proportions among two or more independent samples	305
20.4	Bias when there is one degree of freedom	308
20.5	Three-dimensional contingency tables	312
20.6	Inappropriate use of tests for goodness of fit and heterogeneity	312
20.7	Comparing proportions among two or more related samples of nominal scale data	314
20.8	Recommended tests for categorical data	316
20.9	Reporting the results of tests for categorical data	316
20.10	Questions	318
21	**Non-parametric tests for ratio, interval or ordinal scale data**	319
21.1	Introduction	319
21.2	A non-parametric comparison between one sample and an expected distribution	320
21.3	Non-parametric comparisons between two independent samples	325
21.4	Non-parametric comparisons among three or more independent samples	331
21.5	Non-parametric comparisons of two related samples	335
21.6	Non-parametric comparisons among three or more related samples	338

21.7	Analysing ratio, interval or ordinal data that show gross differences in variance among treatments and cannot be satisfactorily transformed	341
21.8	Non-parametric correlation analysis	342
21.9	Other non-parametric tests	344
21.10	Questions	344
22	**Introductory concepts of multivariate analysis**	346
22.1	Introduction	346
22.2	Simplifying and summarising multivariate data	347
22.3	An *R*-mode analysis: principal components analysis	348
22.4	*Q*-mode analyses: multidimensional scaling	361
22.5	*Q*-mode analyses: cluster analysis	368
22.6	Which multivariate analysis should you use?	372
22.7	Questions	374
23	**Choosing a test**	375
23.1	Introduction	375
	Appendix: Critical values of chi-square, *t* and *F*	388
	References	394
	Index	396