

# Biostatistics: Data and Models

John D. Reeve  
Southern Illinois University Carbondale  
Carbondale, IL 62901

© 2016 John D. Reeve  
All Rights Reserved

**Acknowledgments**

I would like to thank my parents, Ann M. and John H. Reeve, and Kim A. Cole, for their enduring support. I would also like to thank James T. Cronin, Luis Miguel Flores-Campaña, Jamie M. Kneitel, and Fernando T. Maestre for providing data sets used in this book.

# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
1.1	Why this textbook? . . . . .	13
1.2	Types of data . . . . .	16
1.3	Data and models . . . . .	16
1.4	Sequence of topics . . . . .	17
1.5	References . . . . .	19
<b>2</b>	<b>Review of Mathematics</b>	<b>21</b>
2.1	Exponents . . . . .	21
2.2	Inequalities . . . . .	23
2.3	Functions . . . . .	24
	2.3.1 Functions in Statistics . . . . .	25
	2.3.2 Plotting functions using SAS - SAS demo . . . . .	26
2.4	Solving linear equations . . . . .	35
2.5	Roots of equations . . . . .	36
2.6	Calculus . . . . .	37
	2.6.1 Derivatives . . . . .	37
	2.6.2 Function plot with derivative - SAS demo . . . . .	39
	2.6.3 Integrals . . . . .	43
2.7	References . . . . .	45
2.8	Problems . . . . .	46
<b>3</b>	<b>Populations and Statistics</b>	<b>47</b>
3.1	Statistical populations . . . . .	47
3.2	Descriptive statistics and frequency . . . . .	48
	3.2.1 Sample mean . . . . .	49
	3.2.2 Median . . . . .	50
	3.2.3 Sample variance . . . . .	51

3.2.4	Standard deviation . . . . .	51
3.2.5	Coefficient of variation . . . . .	51
3.2.6	Range . . . . .	52
3.2.7	Frequency distributions - SAS demo . . . . .	52
3.2.8	Mode . . . . .	59
3.2.9	Skewness . . . . .	59
3.2.10	Kurtosis . . . . .	62
3.2.11	Development time - SAS demo . . . . .	65
3.2.12	Frequency distributions for categorical data - SAS demo	71
3.3	References . . . . .	75
3.4	Problems . . . . .	76
<b>4</b>	<b>Probability Theory</b>	<b>77</b>
4.1	Probability theory . . . . .	77
4.1.1	Events . . . . .	77
4.1.2	Union, intersection, and complement of events . . . . .	78
4.1.3	Probability distributions . . . . .	82
4.1.4	Probability spaces . . . . .	84
4.1.5	Independence of events . . . . .	84
4.1.6	Conditional probability . . . . .	85
4.1.7	A biological probability distribution . . . . .	86
4.1.8	Bayes theorem . . . . .	89
4.2	References . . . . .	93
4.3	Problems . . . . .	94
<b>5</b>	<b>Discrete Random Variables</b>	<b>95</b>
5.1	Binomial distribution . . . . .	96
5.1.1	Binomial distribution - SAS demo . . . . .	98
5.2	Poisson distribution . . . . .	102
5.2.1	Poisson distribution - SAS demo . . . . .	103
5.3	Negative binomial distribution . . . . .	106
5.3.1	Negative binomial distribution - SAS demo . . . . .	106
5.4	Expected values for discrete distributions . . . . .	110
5.4.1	Variance for discrete distributions . . . . .	112
5.5	Discrete random variables and samples . . . . .	114
5.5.1	Parasitic wasps - SAS demo . . . . .	114
5.5.2	Corn borers - SAS demo . . . . .	123
5.6	Classifying spatial or temporal patterns . . . . .	136

5.7	References . . . . .	139
5.8	Problems . . . . .	140
<b>6</b>	<b>Continuous Random Variables</b>	<b>143</b>
6.1	Uniform distribution . . . . .	144
6.1.1	Random sampling coordinates - SAS demo . . . . .	147
6.2	Normal distribution . . . . .	151
6.2.1	Normal distribution - SAS demo . . . . .	152
6.2.2	Sample calculations - standard normal distribution . . . . .	155
6.2.3	Sample calculations - other normal distributions . . . . .	160
6.3	Expected values and variance for continuous distributions . . . . .	163
6.4	Continuous random variables and samples . . . . .	164
6.4.1	Elytra lengths - SAS demo . . . . .	165
6.4.2	Development time - SAS demo . . . . .	172
6.5	References . . . . .	179
6.6	Problems . . . . .	180
<b>7</b>	<b>Expected Value, Variance, and Samples</b>	<b>181</b>
7.1	Expected value and variance . . . . .	181
7.2	Linear functions and sums - expected value and variance . . . . .	183
7.3	Sample mean - expected value and variance . . . . .	184
7.4	Sample variance - expected value . . . . .	185
7.5	Sample calculations and simulation - SAS demo . . . . .	186
7.6	Central limit theorem . . . . .	195
7.6.1	Central limit theorem - SAS demo . . . . .	195
7.7	Applications of the central limit theorem . . . . .	202
7.8	References . . . . .	203
7.9	Problems . . . . .	204
<b>8</b>	<b>Sampling and Estimation</b>	<b>205</b>
8.1	Random samples . . . . .	205
8.2	Parameter estimation . . . . .	206
8.2.1	Maximum likelihood for Poisson data . . . . .	207
8.2.2	Poisson likelihood function - SAS demo . . . . .	209
8.2.3	Maximum likelihood for normal data . . . . .	214
8.2.4	Normal likelihood function - SAS demo . . . . .	215
8.3	Optimality of maximum likelihood estimates . . . . .	220
8.4	References . . . . .	220

8.5	Problems . . . . .	221
<b>9</b>	<b>Confidence Intervals</b>	<b>223</b>
9.1	Preliminaries to confidence intervals . . . . .	223
9.1.1	Parameters and estimates . . . . .	223
9.1.2	Sampling distributions . . . . .	224
9.2	Confidence intervals . . . . .	229
9.2.1	Confidence intervals for $\mu$ when $\sigma^2$ is known . . . . .	230
9.2.2	Confidence intervals for $\mu$ when $\sigma^2$ is estimated . . . . .	232
9.2.3	Confidence intervals for $\sigma^2$ and $\sigma$ . . . . .	234
9.2.4	Confidence intervals - SAS demo . . . . .	236
9.2.5	Confidence interval size . . . . .	237
9.3	References . . . . .	242
9.4	Problems . . . . .	243
<b>10</b>	<b>Hypothesis Testing</b>	<b>245</b>
10.1	The null and alternative hypotheses . . . . .	245
10.2	Test statistics . . . . .	246
10.3	Acceptance and rejection regions – Type I error . . . . .	247
10.3.1	One-sample $Z$ test - sample calculation . . . . .	250
10.4	$P$ values . . . . .	250
10.5	Type II error and power . . . . .	253
10.6	Summary table . . . . .	258
10.7	One-sample $t$ test . . . . .	259
10.7.1	One-sample $t$ test - sample calculation . . . . .	260
10.7.2	Hypothesis testing - SAS demo . . . . .	261
10.7.3	Power analysis for one-sample $t$ tests - SAS demo . . . . .	264
10.8	One-tailed $t$ test . . . . .	268
10.8.1	One-tailed $t$ test - sample calculation . . . . .	270
10.8.2	One-tailed $t$ test - SAS demo . . . . .	270
10.8.3	One-tailed tests - a warning . . . . .	271
10.9	Confidence intervals and hypothesis testing . . . . .	272
10.10	Likelihood ratio tests . . . . .	273
10.10.1	Example of a likelihood ratio test . . . . .	273
10.11	References . . . . .	277
10.12	Problems . . . . .	278

<b>11 Analysis of Variance (One-Way)</b>	<b>281</b>
11.1 ANOVA models . . . . .	285
11.1.1 Fixed and random effects . . . . .	285
11.1.2 Fixed effects model . . . . .	286
11.1.3 Random effects model . . . . .	289
11.2 Hypothesis testing for ANOVA . . . . .	291
11.2.1 Sums of squares and mean squares . . . . .	291
11.2.2 $F$ statistic and distribution . . . . .	295
11.2.3 ANOVA tables . . . . .	297
11.2.4 One-way ANOVA for Example 1 - SAS demo . . . . .	299
11.2.5 One-way ANOVA for Example 2 - sample calculation . . . . .	305
11.2.6 One-way ANOVA for Example 2 - SAS demo . . . . .	307
11.3 Maximum likelihood estimates . . . . .	315
11.4 $F$ test as a likelihood ratio test . . . . .	317
11.5 One-way ANOVA and two-sample $t$ tests . . . . .	318
11.5.1 Two-sample $t$ test for Example 1 - SAS demo . . . . .	318
11.6 References . . . . .	322
11.7 Problems . . . . .	323
<b>12 Power Analysis for One-Way ANOVA</b>	<b>325</b>
12.1 Power analysis for one-way ANOVA . . . . .	326
12.2 Power analysis - SAS Demo . . . . .	330
12.3 Power analysis continued - SAS demo . . . . .	334
12.4 Power analysis continued - SAS demo . . . . .	338
12.5 References . . . . .	340
12.6 Problems . . . . .	341
<b>13 Multiple Comparisons</b>	<b>343</b>
13.1 Models for multiple comparisons . . . . .	343
13.2 Error rates in multiple comparisons . . . . .	344
13.3 All pairwise comparisons . . . . .	346
13.3.1 Least significant difference . . . . .	347
13.3.2 Least significant difference - SAS demo . . . . .	349
13.3.3 The Tukey procedure . . . . .	357
13.3.4 Tukey procedure - SAS demo . . . . .	358
13.3.5 Multiple range tests - REGW . . . . .	361
13.3.6 REGW procedure - SAS demo . . . . .	363
13.4 Comparisons with a control - Dunnett procedure . . . . .	365

13.4.1	Dunnett's procedure - SAS demo . . . . .	365
13.5	Bonferroni and Sidak corrections . . . . .	367
13.6	Vascular plant cover - SAS demo . . . . .	369
13.7	False discovery rate method . . . . .	377
13.7.1	False discovery rate - SAS demo . . . . .	379
13.8	References . . . . .	382
13.9	Problems . . . . .	383
<b>14</b>	<b>Analysis of Variance (Two-Way)</b>	<b>387</b>
14.1	Random assignment of treatments . . . . .	393
14.1.1	Random assignment of treatments - SAS Demo . . . . .	394
14.2	Two-way fixed effects model . . . . .	396
14.2.1	Factor A effect . . . . .	397
14.2.2	Factor B effect . . . . .	397
14.2.3	Factor A and B effect . . . . .	397
14.2.4	Interaction effect . . . . .	397
14.3	Hypothesis testing for two-way ANOVA . . . . .	401
14.3.1	Sum of squares and mean squares . . . . .	401
14.3.2	ANOVA tables and tests . . . . .	405
14.3.3	Two-way ANOVA for Example 1 - SAS demo . . . . .	408
14.3.4	Two-way ANOVA for Example 2 - SAS demo . . . . .	416
14.3.5	Tests for main effects with interaction . . . . .	424
14.4	Unbalanced designs and two-way ANOVA . . . . .	428
14.5	Two-way ANOVA without replication . . . . .	431
14.5.1	Hypothesis testing . . . . .	431
14.5.2	Two-way ANOVA no replication - SAS demo . . . . .	438
14.6	Randomized block designs . . . . .	447
14.6.1	Randomized block models . . . . .	449
14.6.2	Hypothesis testing and variance components . . . . .	449
14.6.3	Randomized block design - SAS demo . . . . .	450
14.6.4	Likelihood ratio test for the block effect . . . . .	459
14.7	References . . . . .	466
14.8	Problems . . . . .	467
<b>15</b>	<b>Assumptions and Transformations</b>	<b>469</b>
15.1	ANOVA assumptions . . . . .	469
15.1.1	Independence of observations . . . . .	469
15.1.2	Homogeneity of variances . . . . .	470

15.1.3	Normality . . . . .	471
15.1.4	Absence of outliers . . . . .	471
15.1.5	Additivity . . . . .	472
15.2	Variance-stabilizing transformations . . . . .	473
15.3	Residual analysis . . . . .	474
15.3.1	Models, estimates, and predictors . . . . .	475
15.3.2	Predicted and residual values . . . . .	475
15.3.3	Evaluating ANOVA assumptions . . . . .	477
15.3.4	Residual analysis and transformations - SAS demo . . . . .	478
15.3.5	$\arcsin(\sqrt{Y})$ transformation - SAS demo . . . . .	485
15.3.6	Transformations when data are limited . . . . .	492
15.4	References . . . . .	493
<b>16</b>	<b>Nonparametric Tests</b>	<b>495</b>
16.1	Wilcoxon two-sample test . . . . .	499
16.1.1	Wilcoxon test for Example 1 - SAS demo . . . . .	501
16.2	Kruskal-Wallis test . . . . .	507
16.2.1	Kruskal-Wallis test for Example 1 - SAS demo . . . . .	508
16.2.2	Kruskal-Wallis test for Example 2 - SAS demo . . . . .	509
16.3	Kolmogorov-Smirnov test . . . . .	512
16.3.1	Kolmogorov-Smirnov test for Example 1 - SAS demo . . . . .	513
16.4	Randomization tests . . . . .	516
16.4.1	Randomization test for Example 3 - SAS demo . . . . .	519
16.5	Limitations of nonparametric tests . . . . .	525
16.6	Problems . . . . .	527
16.7	References . . . . .	528
<b>17</b>	<b>Linear Regression</b>	<b>529</b>
17.1	Linear regression model . . . . .	533
17.2	Linear regression and likelihood . . . . .	533
17.2.1	Sample calculation - $\hat{\beta}$ , $\hat{\alpha}$ , and $F$ test . . . . .	539
17.3	Confidence and prediction intervals . . . . .	542
17.3.1	Sample calculation - confidence and prediction intervals . . . . .	544
17.4	$R^2$ values . . . . .	546
17.5	Linear regression for Example 1 - SAS demo . . . . .	547
17.6	Assumptions and transformations . . . . .	558
17.6.1	Species-area data - SAS demo . . . . .	559
17.6.2	Population growth rates - SAS demo . . . . .	569

17.7 Problems . . . . .	578
17.8 References . . . . .	580
<b>18 Correlation</b>	<b>581</b>
18.1 Correlation model . . . . .	584
18.2 Correlation and maximum likelihood . . . . .	589
18.2.1 Correlation for Example 1 - SAS demo . . . . .	591
18.2.2 Testing $H_0 : \rho = \rho_0$ - SAS demo . . . . .	595
18.2.3 Correlation for <i>I. setosa</i> , all data - SAS demo . . . . .	596
18.3 Correlation assumptions . . . . .	600
18.4 Nonparametric correlation . . . . .	602
18.4.1 Spearman rank correlation for Example 1 - SAS demo . . . . .	604
18.5 Problems . . . . .	605
18.6 References . . . . .	607
<b>19 More Complex ANOVA Designs</b>	<b>609</b>
19.1 Three-way ANOVA . . . . .	609
19.1.1 Three-way fixed effects model . . . . .	612
19.1.2 Three-way ANOVA for Example 1 - SAS demo . . . . .	613
19.1.3 Tests for main effects with interaction . . . . .	623
19.1.4 Other three-way designs . . . . .	628
19.2 One-way nested ANOVA . . . . .	629
19.2.1 Nested ANOVA models . . . . .	631
19.2.2 Nested ANOVA for Example 2 - SAS demo . . . . .	632
19.3 Analysis of covariance . . . . .	641
19.3.1 ANCOVA model . . . . .	643
19.3.2 ANCOVA for Example 3 - SAS demo . . . . .	643
19.4 References . . . . .	651
19.5 Problems . . . . .	652
<b>20 Methods for Categorical Data</b>	<b>655</b>
20.1 Goodness-of-fit tests . . . . .	657
20.1.1 Goodness-of-fit tests for $a$ categories . . . . .	663
20.1.2 Goodness-of-fit tests with estimated parameters . . . . .	668
20.2 Tests of independence . . . . .	674
20.2.1 Sample calculation . . . . .	676
20.2.2 Test of independence - SAS demo . . . . .	677
20.2.3 Test of independence - SAS demo 2 . . . . .	683

<i>CONTENTS</i>	11
20.3 Problems . . . . .	690
20.4 References . . . . .	692
<b>21 Data Sets</b>	<b>693</b>
21.1 Elytra Length . . . . .	694
21.2 Development Time . . . . .	698
21.3 Plant Biomass . . . . .	701
21.4 <i>Anagrus</i> fecundity . . . . .	703
21.5 Fitness of <i>T. dubius</i> . . . . .	711
21.6 <i>Iris</i> flower measurements . . . . .	713
<b>22 Statistical Tables</b>	<b>717</b>
22.1 Table Z: Probabilities for the standard normal distribution. . .	718
22.2 Table T: Quantiles of the $t$ distribution . . . . .	720
22.3 Table C: Quantiles of the $\chi^2$ distribution . . . . .	723
22.4 Table F: Quantiles of the $F$ distribution . . . . .	726

