

Contents

Preface	xi
1. Variation (or What Statistics Is All About)	1
1.1. Variation	1
1.2. Collecting Data	2
1.3. Summarizing Your Data	3
1.3.1 Learning to Use Excel	4
1.4. Reporting Your Results: the Classroom Data	7
1.4.1 Picturing Data	10
1.4.2 Displaying Multiple Variables	10
1.4.3 Percentiles of the Distribution	15
1.5. Types of Data	20
1.5.1 Depicting Categorical Data	21
1.5.2 From Observations to Questions	23
1.6. Measures of Location	23
1.6.1 Which Measure of Location?	25
1.6.2 The Bootstrap	27
1.7. Samples and Populations	30
1.7.1 Drawing a Random Sample	32
1.7.2 Ensuring the Sample is Representative	34
1.8. Variation—Within and Between	34
1.9. Summary and Review	36
2. Probability	39
2.1. Probability	39
2.1.1 Events and Outcomes	41
2.1.2 Venn Diagrams	41
2.2. Binomial	43
2.2.1 Permutations and Rearrangements	45
2.2.2 Back to the Binomial	47

2.2.3	The Problem Jury	47
2.2.4	Properties of the Binomial	48
2.2.5	Multinomial	52
2.3.	Conditional Probability	53
2.3.1	Market Basket Analysis	55
2.3.2	Negative Results	56
2.4.	Independence	57
2.5.	Applications to Genetics	59
2.6.	Summary and Review	60
3.	Distributions	63
3.1.	Distribution of Values	63
3.1.1	Cumulative Distribution Function	64
3.1.2	Empirical Distribution Function	66
3.2.	Discrete Distributions	66
3.3.	Poisson: Events Rare in Time and Space	68
3.3.1	Applying the Poisson	69
3.3.2	Comparing Empirical and Theoretical Poisson Distributions	70
3.4.	Continuous Distributions	71
3.4.1	The Exponential Distribution	71
3.4.2	The Normal Distribution	72
3.4.3	Mixtures of Normal Distributions	74
3.5.	Properties of Independent Observations	74
3.6.	Testing a Hypothesis	76
3.6.1	Analyzing the Experiment	77
3.6.2	Two Types of Errors	80
3.7.	Estimating Effect Size	81
3.7.1	Confidence Interval for Difference in Means	82
3.7.2	Are Two Variables Correlated?	84
3.7.3	Using Confidence Intervals to Test Hypotheses	86
3.8.	Summary and Review	87
4.	Testing Hypotheses	89
4.1.	One-Sample Problems	89
4.1.1	Percentile Bootstrap	89
4.1.2	Parametric Bootstrap	90
4.1.3	Student's t	91
4.2.	Comparing Two Samples	93
4.2.1	Comparing Two Poisson Distributions	93
4.2.2	What Should We Measure?	94

4.2.3	Permutation Monte Carlo	95
4.2.4	Two-Sample t -Test	97
4.3.	Which Test Should We Use?	97
4.3.1	p Values and Significance Levels	98
4.3.2	Test Assumptions	98
4.3.3	Robustness	99
4.3.4	Power of a Test Procedure	100
4.3.5	Testing for Correlation	101
4.4.	Summary and Review	104
5.	Designing an Experiment or Survey	105
5.1.	The Hawthorne Effect	106
5.1.1	Crafting an Experiment	106
5.2.	Designing an Experiment or Survey	108
5.2.1	Objectives	109
5.2.2	Sample from the Right Population	110
5.2.3	Coping with Variation	112
5.2.4	Matched Pairs	113
5.2.5	The Experimental Unit	114
5.2.6	Formulate Your Hypotheses	114
5.2.7	What Are You Going to Measure?	115
5.2.8	Random Representative Samples	116
5.2.9	Treatment Allocation	117
5.2.10	Choosing a Random Sample	118
5.2.11	Ensuring that Your Observations are Independent	119
5.3.	How Large a Sample?	120
5.3.1	Samples of Fixed Size	121
•	Known Distribution	122
•	Almost Normal Data	125
•	Bootstrap	127
5.3.2	Sequential Sampling	129
•	Stein's Two-Stage Sampling Procedure	129
•	Wald Sequential Sampling	129
•	Adaptive Sampling	133
5.4.	Meta-Analysis	134
5.5.	Summary and Review	135
6.	Analyzing Complex Experiments	137
6.1.	Changes Measured in Percentages	137
6.2.	Comparing More Than Two Samples	138

6.2.1	Programming the Multisample Comparison with Excel	139
6.2.2	What Is the Alternative?	141
6.2.3	Testing for a Dose Response or Other Ordered Alternative	141
6.3.	Equalizing Variances	145
6.4.	Stratified Samples	147
6.5.	Categorical Data	148
6.5.1	One-Sided Fisher's Exact Test	150
6.5.2	The Two-Sided Test	151
6.5.3	Multinomial Tables	152
6.5.4	Ordered Categories	153
6.6.	Summary and Review	154
7.	Developing Models	155
7.1.	Models	155
7.1.1	Why Build Models?	156
7.1.2	Caveats	158
7.2.	Regression	159
7.2.1	Linear Regression	160
7.3.	Fitting a Regression Equation	161
7.3.1	Ordinary Least Squares	162
	• Types of Data	166
7.3.2	Least Absolute Deviation Regression	168
7.3.3	Errors-in-Variables Regression	168
7.3.4	Assumptions	171
7.4.	Problems with Regression	172
7.4.1	Goodness of fit versus prediction	172
7.4.2	Which Model?	173
7.4.3	Measures of Predictive Success	174
7.4.4	Multivariable Regression	175
7.5.	Quantile Regression	182
7.6.	Validation	183
7.6.1	Independent Verification	183
7.6.2	Splitting the Sample	184
7.6.3	Cross-Validation with the Bootstrap	185
7.7.	Classification and Regression Trees	186
7.8.	Data Mining	190
7.9.	Summary and Review	193

8. Reporting Your Findings	195
8.1. What to Report	195
8.2. Text, Table, or Graph?	199
8.3. Summarizing Your Results	200
8.3.1 Center of the Distribution	201
8.3.2 Dispersion	203
8.4. Reporting Analysis Results	204
8.4.1 p Values? Or Confidence Intervals?	205
8.5. Exceptions Are the Real Story	206
8.5.1 Nonresponders	206
8.5.2 The Missing Holes	207
8.5.3 Missing Data	207
8.5.4 Recognize and Report Biases	208
8.6. Summary and Review	209
9. Problem Solving	211
9.1. The Problems	211
9.2. Solving Practical Problems	215
9.2.1 The Data's Provenance	215
9.2.2 Inspect the Data	216
9.2.3 Validate the Data Collection Methods	217
9.2.4 Formulate Hypotheses	217
9.2.5 Choosing a Statistical Methodology	218
9.2.6 Be Aware of What You Don't Know	218
9.2.7 Qualify Your Conclusions	218
Appendix: An Microsoft Office Excel Primer	221
Index to Excel and Excel Add-In Functions	227
Subject Index	229