Contents

Preface ix

Glossary and Notation xi

1 Summary diagrams and tables 1

1.1 The purpose of Statistics 1
1.2 Variables and observations 2
1.3 Types of data 2
1.4 Tally charts and frequency distributions 3
1.5 Stem-and-leaf diagrams 3
1.6 Bar charts 6
1.7 Multiple bar charts 8
1.8 Compound bars for proportions 8
1.9 Population pyramids 9
1.10 Pictograms 10
1.11 Pie charts 10
1.12 Triangles 13
1.13 Grouped frequency tables 15
1.14 Difficulties with grouped frequencies 16
1.15 Histograms 17
1.16 Frequency polygons 22
1.17 Cumulative frequency diagrams 23
Step diagrams 24
1.18 Cyclic and circular data 26
1.19 Time series 27
1.20 Train timetables 27
1.21 Moscow or bust! 28
1.22 Scatter diagrams 28
1.23 Contingency tables 30
1.24 Cartograms 31
1.25 Chernoff faces 32
Chapter summary 34

2 General summary statistics 36

2.1 The purpose of summary statistics 36
2.2 The mode 36
Modal class 37
2.3 The median 38
2.4 The mean 39
2.5 Advantages and disadvantages of the mode, mean, and median 41
Advantages 41
Disadvantages 41
2.6 Sigma (Σ) notation 42
Applications of sigma notation 42
2.7 The mean of a frequency distribution 44

2.8 The mean of grouped data 44
2.9 Using coded values to simplify calculations 46
2.10 The median of grouped data 50
2.11 Quartiles, deciles and percentiles 51
Grouped data 51
Ungrouped data 52
2.12 Range and interquartile range 55
2.13 Box whisker diagrams 55
Refined boxplots 56
2.14 Deviations from the mean 58
2.15 The mean deviation 59
2.16 The variance 59
Using the divisor n 59
Using the divisor $(n-1)$ 59
2.17 Calculating the variance 60
2.18 The sample standard deviation 61
Approximate properties of the standard deviation 63
2.19 Variance and standard deviation for frequency distributions 65
2.20 Variance calculations using coded values 66
2.21 Symmetric and skewed data 69
2.22 Standardising to a prescribed mean and standard deviation 71
2.23 Calculating the combined mean and variance of several samples 72
2.24 Combining proportions 73
Chapter summary 75

3 Special summary statistics 84
3.1 Standardised birth and death rates 84
3.2 The weighted mean and index numbers 86
3.3 Price indices 88
3.4 Record prices and record earnings! 89
3.5 Time series 91
An underlying trend 91
Periodicity 92
Cycles 92
Random variation 93
Moving averages 93
4-point and 12-point moving averages 95
Estimating periodic effects 97
Predicting the future 98
Chapter summary 101

4 Data sources **106**

4.1 Data collection by observation 106
4.2 Methods of data collection by questionnaire (or survey) 106
The face-to-face interview 107
The 'postal' questionnaire 107
The telephone interview 108
4.3 Questionnaire design 108
Some poor questions 108
Some good questions 109
The order of questions 109
Question order and bias 110
Filtered questions 110
Open and closed questions 110
The order of answers for closed questions 110
The pilot study 111
4.4 National surveys 111
National censuses 111
Government surveys 111
Other national surveys in Britain 112
4.5 Government publications 112
4.6 The Data Archive 113

5 Probability **114**

5.1 Relative frequency 114
5.2 Preliminary definitions 114
5.3 The probability scale 115
5.4 Probability with equally likely outcomes 115
5.5 The complementary event, E' 117
5.6 Venn diagrams 118
5.7 Unions and intersections of events 118
5.8 Mutually exclusive events 121
The addition rule 121
5.9 Exhaustive events 122
5.10 Probability trees 123
5.11 Sample proportions and probability 125
5.12 Unequally likely possibilities 127
5.13 Physical independence 127
The multiplication rule 127
5.14 Orderings 130
Orderings of similar objects 132
5.15 Permutations and combinations 135
5.16 Sampling with replacement 138
5.17 Sampling without replacement 138
Chapter summary 143

6 Conditional probability **146**

6.1 Notation
The generalised multiplication rule 146

6.2 Statistical independence 148
6.3 Mutual and pairwise independence 152
6.4 The total probability theorem 153
6.5 Bayes' theorem 157
Chapter summary 163

7 Probability distributions and expectations 167
7.1 Notation 167
7.2 Probability distributions 168
The probability function 169
Illustrating probability distributions 169
Estimating probability distributions 171
The cumulative distribution function 172
7.3 Some special discrete probability distributions 173
The discrete uniform distribution 173
The Bernoulli distribution 174
7.4 The geometric distribution 174
Cumulative probabilities 175
A paradox! 176
7.5 Expectations 177
Expected value or expected number 179
Expectations of functions of random variables 181
7.6 The variance 182
7.7 The standard deviation 187
7.8 Greek notation 189
Chapter summary 190

8 Expectation algebra 192
8.1 $E(X + a)$ and $Var(X + a)$ 193
8.2 $E(aX)$ and $Var(aX)$ 195
8.3 $E(aX + b)$, $Var(aX + b)$ and $E[g(X) + h(X)]$ 196
Alternative expressions for $Var(X)$ 197
8.4 Expectations of functions of more than one variable 197
$Var(X + Y)$ 198
$E(X_1 + X_2)$ and $Var(X_1 + X_2)$ 200
The difference between $2X$ and $X_1 + X_2$ 201
8.5 The expectation and variance of the sample mean 203
8.6 The unbiased estimate of the population variance 205
8.7 $E(t^X)$ – the probability generating function 206
Pgf for the sum of independent random variables 210
Chapter summary 213

9 The binomial distribution **216**
9.1 Derivation 216
9.2 Notation 223
9.3 'Successes' and 'failures' 223
9.4 The shape of the distribution 223
9.5 Calculating binomial probabilities 225
9.6 Tables of binomial distributions 227
9.7 The mean and variance of a binomial distribution 229
9.8 The probability generating function 231
9.9 The negative binomial distribution 232
Chapter summary 233

10 The Poisson distribution **236**
10.1 The Poisson process 236
10.2 The form of the distribution 238
10.3 The properties of a Poisson distribution 241
The mean 241
The variance 241
The shape of a Poisson distribution 242
10.4 Calculation of Poisson probabilities 243
10.5 Tables for Poisson distributions 245
10.6 The Poisson approximation to the binomial 247
Derivation of the result 247
10.7 Sums of independent Poisson random variables 250
10.8 The National Lottery 251
10.9 The probability generating function 252
Chapter summary 253

11 Continuous random variables **258**
11.1 Histograms and sample size 258
11.2 The probability density function, f 260
Properties of the pdf 261
11.3 The cumulative distribution function, F 265
The median, m 268
11.4 Expectation and variance 274
11.5 Obtaining f from F 281
11.6 The uniform (rectangular) distribution 282
11.7 The exponential distribution 287
Shape of the exponential distribution 287
Expectation and variance of the exponential distribution 289

Connection with a Poisson process 290
11.8 The Cauchy distribution 294
11.9 Moments 295
11.10 $E(e^{tX})$ – the moment generating function 295
Mgf of the sum of independent random variables 297
Chapter summary 300

12 The normal distribution 303

12.1 The standard normal distribution 303
12.2 Tables of $\Phi(z)$ 304
12.3 Probabilities for other normal distributions 308
12.4 Finer detail in the tables of $\Phi(z)$ 310
12.5 Tables of percentage points 313
12.6 Using calculators 317
12.7 Applications of the normal distribution 318
12.8 General properties 319
12.9 Linear combinations of independent normal random variables 320
Extension to more than two variables 323
12.10 The Central Limit Theorem 327
The distribution of the sample mean, $\bar{X}$ 329
12.11 The normal approximation to a binomial distribution 336
Inequalities 337
Choosing between the normal and Poisson approximations to a binomial distribution 340
12.12 Normal approximation to a Poisson distribution 343
12.13 Normal probability paper 347
12.14 Mgf of a normal distribution 352
Chapter summary 354

13 Sampling and simulation 361

13.1 The purpose of sampling 361
13.2 Methods for sampling a population 361
The simple random sample 361
Cluster sampling 362
Stratified sampling 362
Systematic sampling 363
Quota sampling 363
Self-selection 363
A national survey 363

13.3 Random numbers 364
Pseudo-random numbers 365
Tables of random numbers 365
13.4 Simulation 365
Simulating non-numerical occurrences 367
Simulating a continuous random variable 368
Chapter summary 371

14 Point and interval estimation 374
14.1 Point estimates 374
14.2 Confidence intervals 374
14.3 Confidence interval for a population mean 374
Normal distribution with known variance 375
Unknown population distribution, known population variance, large sample 377
Unknown population distribution, unknown population variance, large sample 377
Poisson distribution, large mean 379
14.4 Confidence interval for a population proportion 381
14.5 One-sided confidence intervals 385
14.6 The *t*-distribution 387
Tables of the *t*-distribution 388
14.7 Confidence interval for a population mean using the *t*-distribution 390
14.8 Desirable properties of an estimator 394
Chapter summary 398

15 Hypothesis tests 402
15.1 The null and alternative hypotheses 402
15.2 Critical regions and significance levels 403
15.3 The general test procedure 404
15.4 Test for mean, known variance, normal distribution or large sample 404
15.5 Identifying the two hypotheses 408
The null hypothesis 408
The alternative hypothesis 408
15.6 Test for mean, large sample, variance unknown 409
15.7 Test for large Poisson mean 411
15.8 Test for proportion, large sample size 413

15.9 Test for mean, small sample, variance unknown 415
15.10 The p-value approach 418
15.11 Hypothesis tests and confidence intervals 419

16 Hypothesis tests: error and problems 424
16.1 Type I and Type II errors 424
The general procedure 424
16.2 Comparing precise hypotheses 430
16.3 The power curve 433
16.4 Hypothesis tests for a proportion based on a small sample 436
16.5 Hypothesis tests for a Poisson mean based on a small sample 439
16.6 Type II errors in small samples 443
16.7 Quality control 443
Process control 443
Control charts in practice: (i) The mean chart 445
Control charts in practice: (ii) The range chart 445
Acceptance sampling 447
The OC curve 449
Chapter summary 452

17 Two samples and paired samples 453
17.1 Comparison of two means 453
17.2 Comparison of two means – known population variances 454
Confidence interval for the common mean 454
17.3 Comparison of two means – common unknown population variance 458
Large sample sizes 459
Small sample sizes 462
17.4 Confidence intervals for the difference between two means 464
Known variances 464
Unknown common variance 465
Large samples 465
17.5 Comparison of two proportions 468
17.6 Paired samples
Experimental design 472
Distinguishing between the paired-sample and two-sample cases 473
The paired-sample comparison of means 473
Chapter summary 477

18 Goodness of fit **479**

18.1 The chi-squared distribution 480
Properties of the chi-squared distribution 480
Tables of the chi-squared distribution 481
18.2 Goodness of fit to prescribed probabilities 481
18.3 Small expected frequencies 487
18.4 Goodness of fit to prescribed distribution type 490
18.5 Contingency tables 496
The Yates correction 499
18.6 The dispersion test 503
18.7 Comparing distribution functions 505
Chapter summary 505

19 Handling variances **509**

19.1 Confidence intervals for σ^2 509
19.2 Testing a hypothesis about σ^2 512
19.3 The F-distribution 514
Tables of the F-distribution 515
19.4 Comparison of two variances 516
19.5 Confidence interval for a variance ratio 518
Chapter summary 519

20 Regression and correlation **521**

20.1 The equation of a straight line 522
Determining the equation 522
20.2 The estimated regression line 523
20.3 Why 'regression'? 528
20.4 The method of least squares 529
Derivation of the estimates 529
20.5 Dependent random variable Y 531
Estimating a future y-value 532
20.6 Transformations, extrapolation and outliers 534
20.7 Confidence interval for the population regression coefficient β 536
Mean and variance of the estimator of β 537
Significance test for the regression coefficient 537
20.8 Distinguishing x and Y 540
20.9 Deducing x from a Y-value 540
20.10 Two regression lines 541
20.11 Correlation 545
20.12 The product–moment correlation coefficient 546
Correlation and regression 551

The population product–moment correlation coefficient, ρ 551
Testing the significance of r 552
Chapter summary 555

21 Experimental design and the analysis of variance (ANOVA) 560

21.1 The comparison of more than two means 552
21.2 One-way ANOVA 562
21.3 Models 565
21.4 Randomised blocks 566
21.5 Two-way ANOVA 567
Using coded values 569
21.6 Latin squares 571
21.7 Replication 574
Chapter summary 577

22 Nonparametric tests 580

22.1 The single-sample sign test 580
22.2 The Wilcoxon signed-rank test 583
22.3 The paired-sample sign test 586
22.4 The Wilcoxon matched-pairs signed-rank test 588
22.5 The Wilcoxon rank-sum test 589
22.6 Spearman's rank correlation coefficient, r_s 595
Testing the significance of r_s 597
Alternative table formats 598
22.7 Using r_s for non-linear relationships 600
22.8 r_s is the product–moment correlation coefficient for ranks 602
22.9 Kendall's τ 603
Testing the significance of τ 606
Chapter summary 610

Appendices

Cumulative probabilities for the binomial distribution 617
Cumulative probabilities for the Poisson distribution 618
The nominal distribution function 619
Upper-tail percentage points for the standard normal distribution 620
Percentage points for the t-distribution 621
Percentage points for the χ^2-distribution 622
Percentage points for the F-distribution 623
Critical values for the product–moment correlation coefficient, r 625

Critical values for Spearman's rank correlation coefficient, r_s 626
Critical values for Kendall's τ 627
Factors for control chart action lines 628
Random numbers 628

Answers 629

Index 652