



Data Analysis with Excel[®]

An Introduction
for Physical Scientists

LES KIRKUP

CAMBRIDGE

Contents

Preface xv

- 1 Introduction to scientific data analysis 1**
- 1.1 Introduction 1
- 1.2 Scientific experimentation 2
 - 1.2.1 Aim of an experiment 3
 - 1.2.2 Experimental design 4
- 1.3 Units and standards 6
 - 1.3.1 Units 7
 - 1.3.2 Standards 10
 - 1.3.3 Prefixes and scientific notation 10
 - 1.3.4 Significant figures 12
- 1.4 Picturing experimental data 14
 - 1.4.1 Histograms 14
 - 1.4.2 Relationships and the x - y graph 17
 - 1.4.3 Logarithmic scales 20
- 1.5 Key numbers summarise experimental data 21
 - 1.5.1 The mean and the median 22
 - 1.5.2 Variance and standard deviation 24
- 1.6 Population and sample 26
 - 1.6.1 Population parameters 27
 - 1.6.2 True value and population mean 27
 - 1.6.3 Sample statistics 28
 - 1.6.4 Which standard deviation do we use? 30
 - 1.6.5 Approximating s 31
- 1.7 Experimental error 32
 - 1.7.1 Random error 33

- 1.7.2 Systematic error 33
- 1.7.3 Repeatability and reproducibility 34
- 1.8 Modern tools of data analysis – the computer based spreadsheet 34
- 1.9 Review 35
- Problems 36

- 2 Excel® and data analysis 39**
- 2.1 Introduction 39
- 2.2 What is a spreadsheet? 40
- 2.3 Introduction to Excel® 41
 - 2.3.1 Starting with Excel® 42
 - 2.3.2 Worksheets and Workbooks 43
 - 2.3.3 Entering and saving data 44
 - 2.3.4 Rounding, range and the display of numbers 45
 - 2.3.5 Entering formulae 47
 - 2.3.6 Cell references and naming cells 50
 - 2.3.7 Operator precedence and spreadsheet readability 53
 - 2.3.8 Verification and troubleshooting 55
 - 2.3.9 Auditing tools 58
- 2.4 Built in mathematical functions 59
 - 2.4.1 Trigonometrical functions 61
- 2.5 Built in statistical functions 62
 - 2.5.1 SUM(), MAX() and MIN() 62
 - 2.5.2 AVERAGE(), MEDIAN() and MODE() 64
 - 2.5.3 Other useful functions 65
- 2.6 Presentation options 66
- 2.7 Charts in Excel® 68
 - 2.7.1 The x - y graph 68
 - 2.7.2 Plotting multiple sets of data on an x - y graph 73
- 2.8 Data analysis tools 75
 - 2.8.1 Histograms 76
 - 2.8.2 Descriptive statistics 79
- 2.9 Review 80
- Problems 81

- 3 Data distributions I 85**
- 3.1 Introduction 85
- 3.2 Probability 86
 - 3.2.1 Rules of probability 87
- 3.3 Probability distributions 89
 - 3.3.1 Limits in probability calculations 93

- 3.4 Distributions of real data 94
- 3.5 The normal distribution 97
 - 3.5.1 Excel®'s NORMDIST() function 99
 - 3.5.2 The standard normal distribution 101
 - 3.5.3 Excel®'s NORMSDIST() function 104
 - 3.5.4 \bar{x} and s as approximations to μ and σ 106
- 3.6 Confidence intervals and confidence limits 107
 - 3.6.1 The 68% and 95% confidence intervals 109
 - 3.6.2 Excel®'s NORMINV() function 112
 - 3.6.3 Excel®'s NORMSINV() function 113
- 3.7 Distribution of sample means 114
- 3.8 The central limit theorem 116
 - 3.8.1 Standard error of the sample mean 117
 - 3.8.1.1 Approximating σ_x 120
 - 3.8.2 Excel®'s CONFIDENCE() function 121
- 3.9 The t distribution 122
 - 3.9.1 Excel®'s TDIST() and TINV() functions 125
- 3.10 The lognormal distribution 126
- 3.11 Assessing the normality of data 128
 - 3.11.1 The normal quantile plot 128
- 3.12 Population mean and continuous distributions 131
- 3.13 Population mean and expectation value 132
- 3.14 Review 133
 - Problems 133

- 4 Data distributions II 138**
- 4.1 Introduction 138
- 4.2 The binomial distribution 138
 - 4.2.1 Calculation of probabilities using the binomial distribution 140
 - 4.2.2 Probability of a success, p 142
 - 4.2.3 Excel®'s BINOMDIST() function 143
 - 4.2.4 Mean and standard deviation of binomially distributed data 144
 - 4.2.5 Normal distribution as an approximation to the binomial distribution 145
- 4.3 The Poisson distribution 149
 - 4.3.1 Applications of the Poisson distribution 151
 - 4.3.2 Standard deviation of the Poisson distribution 152
 - 4.3.3 Excel®'s POISSON() function 154
 - 4.3.4 Normal distribution as an approximation to the Poisson distribution 156

- 4.4 Review 157
- Problems 158

- 5 Measurement, error and uncertainty 161**
- 5.1 Introduction 161
- 5.2 The process of measurement 162
- 5.3 True value, error and uncertainty 165
 - 5.3.1 Calculation of uncertainty, u 167
- 5.4 Precision and accuracy 170
- 5.5 Random and systematic errors 171
- 5.6 Random errors 172
 - 5.6.1 Common sources of error 172
- 5.7 Absolute, fractional and percentage uncertainties 174
 - 5.7.1 Combining uncertainties caused by random errors 176
 - 5.7.2 Equations containing a single variable 176
 - 5.7.3 Equations containing more than one variable 178
 - 5.7.4 Most probable uncertainty 180
 - 5.7.5 Review of combining uncertainties 183
- 5.8 Coping with extremes in data variability 183
 - 5.8.1 Outliers 183
 - 5.8.2 Chauvenet's criterion 184
 - 5.8.3 Dealing with values that show no variability 187
- 5.9 Uncertainty due to systematic errors 190
 - 5.9.1 Calibration errors and specifications 191
 - 5.9.2 Offset and gain errors 192
 - 5.9.3 Loading errors 195
 - 5.9.4 Dynamic effects 197
 - 5.9.5 Zero order system 198
 - 5.9.6 First order system 198
- 5.10 Combining uncertainties caused by systematic errors 201
- 5.11 Combining uncertainties due to random and systematic errors 202
 - 5.11.1 Type A and Type B categorisation of uncertainties 205
- 5.12 Weighted mean 205
 - 5.12.1 Standard error in the weighted mean 207
 - 5.12.2 Should means be combined? 208
- 5.13 Review 208
- Problems 209

- 6 Least squares I 213**
- 6.1 Introduction 213
- 6.2 The equation of a straight line 214
 - 6.2.1 The 'best' straight line through x - y data 215

- 6.2.2 Unweighted least squares 217
- 6.2.3 Trendline in Excel® 222
- 6.2.4 Uncertainty in a and b 223
- 6.2.5 Least squares, intermediate calculations and significant figures 226
- 6.2.6 Confidence intervals for a and b 226
- 6.3 Excel®'s LINEST() function 228
- 6.4 Using the line of best fit 230
 - 6.4.1 Comparing a 'physical' equation to $y = a + bx$ 231
 - 6.4.1.1 Uncertainties in parameters which are functions of a and b 232
 - 6.4.2 Estimating y for a given x 233
 - 6.4.2.1 Uncertainty in prediction of y at a particular value of x 236
 - 6.4.3 Estimating x for a given y 236
- 6.5 Fitting a straight line to data when random errors are confined to the x quantity 239
- 6.6 Linear correlation coefficient, r 242
 - 6.6.1 Calculating r using Excel® 246
 - 6.6.2 Is the value of r significant? 247
- 6.7 Residuals 249
 - 6.7.1 Standardised residuals 252
- 6.8 Data rejection 254
- 6.9 Transforming data for least squares analysis 257
 - 6.9.1 Consequences of data transformation 262
- 6.10 Weighted least squares 264
 - 6.10.1 Weighted uncertainty in a and b 267
 - 6.10.2 Weighted standard deviation, σ_w 268
 - 6.10.3 Weighted least squares and Excel® 272
- 6.11 Review 272
- Problems 273

- 7 Least squares II 280**
 - 7.1 Introduction 280
 - 7.2 Extending linear least squares 281
 - 7.3 Formulating equations to solve for parameter estimates 283
 - 7.4 Matrices and Excel® 285
 - 7.4.1 The MINVERSE() function 285
 - 7.4.2 The MMULT() function 286
 - 7.4.3 Fitting the polynomial $y = a + bx + cx^2$ to data 288
 - 7.5 Multiple least squares 290
 - 7.6 Standard errors in parameter estimates 293
 - 7.6.1 Confidence intervals for parameters 296
 - 7.7 Weighting the fit 297

- 7.8 Coefficients of multiple correlation and multiple determination 299
- 7.9 The LINEST() function for multiple least squares 300
- 7.10 Choosing equations to fit to data 302
 - 7.10.1 Comparing equations fitted to data 303
- 7.11 Non-linear least squares 306
- 7.12 Review 309
 - Problems 309

- 8 Tests of significance 315**
- 8.1 Introduction 315
- 8.2 Confidence levels and significance testing 316
- 8.3 Hypothesis testing 320
 - 8.3.1 Distribution of the test statistic, z 322
 - 8.3.2 Using Excel® to compare sample mean and hypothesised population mean 325
 - 8.3.3 One tailed and two tailed tests of significance 327
 - 8.3.4 Type I and type II errors 328
- 8.4 Comparing \bar{x} with μ_0 when sample sizes are small 329
- 8.5 Significance testing for least squares parameters 331
- 8.6 Comparison of the means of two samples 334
 - 8.6.1 Excel®'s TTEST() 337
- 8.7 t test for paired samples 339
 - 8.7.1 Excel®'s TTEST() for paired samples 341
- 8.8 Comparing variances using the F test 342
 - 8.8.1 The F distribution 342
 - 8.8.2 The F test 344
 - 8.8.3 Excel®'s FINV() function 346
 - 8.8.4 Robustness of the F test 346
- 8.9 Comparing expected and observed frequencies using the χ^2 test 347
 - 8.9.1 The χ^2 distribution 347
 - 8.9.2 The χ^2 test 349
 - 8.9.3 Is the fit too good? 350
 - 8.9.4 Degrees of freedom in χ^2 test 351
 - 8.9.5 Excel®'s CHINV() function 354
- 8.10 Analysis of variance 354
 - 8.10.1 Principle of ANOVA 355
 - 8.10.2 Example of ANOVA calculation 357
- 8.11 Review 359
 - Problems 360

9	Data Analysis tools in Excel® and the Analysis ToolPak	364
9.1	Introduction	364
9.2	Activating the Data Analysis tools	365
9.2.1	General features	366
9.3	Anova: Single Factor	367
9.4	Correlation	368
9.5	<i>F</i> test Two-Sample for Variances	369
9.6	Random Number Generation	371
9.7	Regression	373
9.7.1	Advanced linear least squares using Excel®'s Regression tool	375
9.8	<i>t</i> tests	376
9.9	Other tools	378
9.9.1	Anova: Two-Factor With Replication and Anova: Two-Factor Without Replication	378
9.9.2	Covariance	378
9.9.3	Exponential Smoothing	379
9.9.4	Fourier analysis	379
9.9.5	Moving average	379
9.9.6	Rank and percentile	380
9.9.7	Sampling	380
9.10	Review	380
Appendix 1	Statistical tables	381
Appendix 2	Propagation of uncertainties	390
Appendix 3	Least squares and the principle of maximum likelihood	392
A3.1	Mean and weighted mean	392
A3.2	Best estimates of slope and intercept	394
A3.3	The line of best fit passes through \bar{x}, \bar{y}	396
A3.4	Weighting the fit	397
Appendix 4	Standard errors in mean, intercept and slope	398
A4.1	Standard error in the mean and weighted mean	398
A4.2	Standard error in intercept <i>a</i> and slope <i>b</i> for a straight line	399
Appendix 5	Introduction to matrices for least squares analysis	403
Appendix 6	Useful formulae	409
	Answers to exercises and problems	413
	References	438
	Index	441