

Statistics in Engineering

With Examples in MATLAB[®] and R

Second Edition

Andrew Metcalfe
David Green
Tony Greenfield
Mahayaudin Mansor
Andrew Smith
Jonathan Tuke



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Contents

Preface	xvii
1 Why understand statistics?	1
1.1 Introduction	1
1.2 Using the book	2
1.3 Software	2
2 Probability and making decisions	3
2.1 Introduction	3
2.2 Random digits	4
2.2.1 Concepts and uses	4
2.2.2 Generating random digits	5
2.2.3 Pseudo random digits	6
2.3 Defining probabilities	7
2.3.1 Defining probabilities – Equally likely outcomes	8
2.3.2 Defining probabilities – Relative frequencies	11
2.3.3 Defining probabilities – Subjective probability and expected monetary value	13
2.4 Axioms of probability	15
2.5 The addition rule of probability	15
2.5.1 Complement	16
2.6 Conditional probability	18
2.6.1 Conditioning on information	18
2.6.2 Conditional probability and the multiplicative rule	18
2.6.3 Independence	20
2.6.4 Tree diagrams	23
2.7 Bayes' theorem	25
2.7.1 Law of total probability	26
2.7.2 Bayes' theorem for two events	27
2.7.3 Bayes' theorem for any number of events	28
2.8 Decision trees	29
2.9 Permutations and combinations	31
2.10 Simple random sample	33
2.11 Summary	35
2.11.1 Notation	35
2.11.2 Summary of main results	36
2.11.3 MATLAB [®] and R commands	36
2.12 Exercises	37

3	Graphical displays of data and descriptive statistics	55
3.1	Types of variables	55
3.2	Samples and populations	58
3.3	Displaying data	61
3.3.1	Stem-and-leaf plot	61
3.3.2	Time series plot	62
3.3.3	Pictogram	65
3.3.4	Pie chart	68
3.3.5	Bar chart	68
3.3.6	Rose plot	70
3.3.7	Line chart for discrete variables	70
3.3.8	Histogram and cumulative frequency polygon for continuous variables	73
3.3.9	Pareto chart	77
3.4	Numerical summaries of data	79
3.4.1	Population and sample	79
3.4.2	Measures of location	81
3.4.3	Measures of spread	90
3.5	Box-plots	95
3.6	Outlying values and robust statistics	97
3.6.1	Outlying values	97
3.6.2	Robust statistics	98
3.7	Grouped data	99
3.7.1	Calculation of the mean and standard deviation for discrete data .	99
3.7.2	Grouped continuous data [Mean and standard deviation for grouped continuous data]	100
3.7.3	Mean as center of gravity	101
3.7.4	Case study of wave stress on offshore structure.	103
3.8	Shape of distributions	103
3.8.1	Skewness	103
3.8.2	Kurtosis	104
3.8.3	Some contrasting histograms	105
3.9	Multivariate data	108
3.9.1	Scatter plot	108
3.9.2	Histogram for bivariate data	110
3.9.3	Parallel coordinates plot	111
3.10	Descriptive time series	113
3.10.1	Definition of time series	113
3.10.2	Missing values in time series	114
3.10.3	Decomposition of time series	114
3.10.3.1	Trend - Centered moving average	114
3.10.3.2	Seasonal component - Additive monthly model	115
3.10.3.3	Seasonal component - Multiplicative monthly model	115
3.10.3.4	Seasonal adjustment	116
3.10.3.5	Forecasting	116
3.10.4	Index numbers	119
3.11	Summary	121
3.11.1	Notation	121
3.11.2	Summary of main results	121
3.11.3	MATLAB and R commands	122
3.12	Exercises	123

4	Discrete probability distributions	137
4.1	Discrete random variables	137
4.1.1	Definition of a discrete probability distribution	138
4.1.2	Expected value	139
4.2	Bernoulli trial	140
4.2.1	Introduction	140
4.2.2	Defining the Bernoulli distribution	141
4.2.3	Mean and variance of the Bernoulli distribution	141
4.3	Binomial distribution	142
4.3.1	Introduction	142
4.3.2	Defining the Binomial distribution	142
4.3.3	A model for conductivity	147
4.3.4	Mean and variance of the binomial distribution	148
4.3.5	Random deviates from binomial distribution	149
4.3.6	Fitting a binomial distribution	149
4.4	Hypergeometric distribution	150
4.4.1	Defining the hypergeometric distribution	151
4.4.2	Random deviates from the hypergeometric distribution	152
4.4.3	Fitting the hypergeometric distribution	152
4.5	Negative binomial distribution	153
4.5.1	The geometric distribution	153
4.5.2	Defining the negative binomial distribution	154
4.5.3	Applications of negative binomial distribution	155
4.5.4	Fitting a negative binomial distribution	157
4.5.5	Random numbers from a negative binomial distribution	157
4.6	Poisson process	158
4.6.1	Defining a Poisson process in time	158
4.6.2	Superimposing Poisson processes	158
4.6.3	Spatial Poisson process	158
4.6.4	Modifications to Poisson processes	159
4.6.5	Poisson distribution	159
4.6.6	Fitting a Poisson distribution	160
4.6.7	Times between events	161
4.7	Summary	162
4.7.1	Notation	162
4.7.2	Summary of main results	162
4.7.3	MATLAB and R commands	163
4.8	Exercises	164
5	Continuous probability distributions	175
5.1	Continuous random variables	175
5.1.1	Definition of a continuous random variable	175
5.1.2	Definition of a continuous probability distribution	176
5.1.3	Moments of a continuous probability distribution	177
5.1.4	Median and mode of a continuous probability distribution	181
5.1.5	Parameters of probability distributions	181
5.2	Uniform distribution	181
5.2.1	Definition of a uniform distribution	182
5.2.2	Applications of the uniform distribution	183
5.2.3	Random deviates from a uniform distribution	183
5.2.4	Distribution of $F(X)$ is uniform	183

5.2.5	Fitting a uniform distribution	184
5.3	Exponential distribution	184
5.3.1	Definition of an exponential distribution	184
5.3.2	Markov property	186
5.3.2.1	Poisson process	186
5.3.2.2	Lifetime distribution	186
5.3.3	Applications of the exponential distribution	187
5.3.4	Random deviates from an exponential distribution	189
5.3.5	Fitting an exponential distribution	190
5.4	Normal (Gaussian) distribution	194
5.4.1	Definition of a normal distribution	194
5.4.2	The standard normal distribution $Z \sim N(0, 1)$	195
5.4.3	Applications of the normal distribution	199
5.4.4	Random numbers from a normal distribution	203
5.4.5	Fitting a normal distribution	203
5.5	Probability plots	203
5.5.1	Quantile-quantile plots	204
5.5.2	Probability plot	204
5.6	Lognormal distribution	205
5.6.1	Definition of a lognormal distribution	205
5.6.2	Applications of the lognormal distribution	208
5.6.3	Random numbers from lognormal distribution	209
5.6.4	Fitting a lognormal distribution	209
5.7	Gamma distribution	209
5.7.1	Definition of a gamma distribution	210
5.7.2	Applications of the gamma distribution	212
5.7.3	Random deviates from gamma distribution	212
5.7.4	Fitting a gamma distribution	212
5.8	Gumbel distribution	213
5.8.1	Definition of a Gumbel distribution	213
5.8.2	Applications of the Gumbel distribution	215
5.8.3	Random deviates from a Gumbel distribution	215
5.8.4	Fitting a Gumbel distribution	216
5.9	Summary	218
5.9.1	Notation	218
5.9.2	Summary of main results	218
5.9.3	MATLAB and R commands	219
5.10	Exercises	220
6	Correlation and functions of random variables	233
6.1	Introduction	233
6.2	Sample covariance and correlation coefficient	236
6.2.1	Defining sample covariance	236
6.3	Bivariate distributions, population covariance and correlation coefficient	244
6.3.1	Population covariance and correlation coefficient	245
6.3.2	Bivariate distributions - Discrete case	246
6.3.3	Bivariate distributions - Continuous case	248
6.3.3.1	Marginal distributions	248
6.3.3.2	Bivariate histogram	249
6.3.3.3	Covariate and correlation	250
6.3.3.4	Bivariate probability distributions	251

6.3.4	Copulas	256
6.4	Linear combination of random variables (propagation of error)	256
6.4.1	Mean and variance of a linear combination of random variables	257
6.4.1.1	Bounds for correlation coefficient	259
6.4.2	Linear combination of normal random variables	260
6.4.3	Central Limit Theorem and distribution of the sample mean	262
6.5	Non-linear functions of random variables (propagation of error)	265
6.6	Summary	267
6.6.1	Notation	267
6.6.2	Summary of main results	267
6.6.3	MATLAB and R commands	268
6.7	Exercises	268
7	Estimation and inference	279
7.1	Introduction	279
7.2	Statistics as estimators	279
7.2.1	Population parameters	280
7.2.2	Sample statistics and sampling distributions	280
7.2.3	Bias and MSE	282
7.3	Accuracy and precision	285
7.4	Precision of estimate of population mean	285
7.4.1	Confidence interval for population mean when σ known	285
7.4.2	Confidence interval for mean when σ unknown	288
7.4.2.1	Construction of confidence interval and rationale for the t -distribution	288
7.4.2.2	The t -distribution	289
7.4.3	Robustness	291
7.4.4	Bootstrap methods	292
7.4.4.1	Bootstrap resampling	292
7.4.4.2	Basic bootstrap confidence intervals	293
7.4.4.3	Percentile bootstrap confidence intervals	293
7.4.5	Parametric bootstrap	296
7.5	Hypothesis testing	299
7.5.1	Hypothesis test for population mean when σ known	300
7.5.2	Hypothesis test for population mean when σ unknown	302
7.5.3	Relation between a hypothesis test and the confidence interval	303
7.5.4	p -value	304
7.5.5	One-sided confidence intervals and one-sided tests	304
7.6	Sample size	305
7.7	Confidence interval for a population variance and standard deviation	307
7.8	Comparison of means	309
7.8.1	Independent samples	309
7.8.1.1	Population standard deviations differ	309
7.8.1.2	Population standard deviations assumed equal	312
7.8.2	Matched pairs	315
7.9	Comparing variances	317
7.10	Inference about proportions	318
7.10.1	Single sample	318
7.10.2	Comparing two proportions	320
7.10.3	McNemar's test	323
7.11	Prediction intervals and statistical tolerance intervals	325

7.11.1	Prediction interval	325
7.11.2	Statistical tolerance interval	326
7.12	Goodness of fit tests	327
7.12.1	Chi-square test	328
7.12.2	Empirical distribution function tests	330
7.13	Summary	332
7.13.1	Notation	332
7.13.2	Summary of main results	333
7.13.3	MATLAB and R commands	335
7.14	Exercises	335
8	Linear regression and linear relationships	357
8.1	Linear regression	357
8.1.1	Introduction	357
8.1.2	The model	359
8.1.3	Fitting the model	361
8.1.3.1	Fitting the regression line	361
8.1.3.2	Identical forms for the least squares estimate of the slope	363
8.1.3.3	Relation to correlation	363
8.1.3.4	Alternative form for the fitted regression line	364
8.1.3.5	Residuals	365
8.1.3.6	Identities satisfied by the residuals	366
8.1.3.7	Estimating the standard deviation of the errors	367
8.1.3.8	Checking assumptions A3, A4 and A5	368
8.1.4	Properties of the estimators	368
8.1.4.1	Estimator of the slope	369
8.1.4.2	Estimator of the intercept	371
8.1.5	Predictions	371
8.1.5.1	Confidence interval for mean value of Y given x	371
8.1.5.2	Limits of prediction	373
8.1.5.3	Plotting confidence intervals and prediction limits	374
8.1.6	Summarizing the algebra	375
8.1.7	Coefficient of determination R^2	376
8.2	Regression for a bivariate normal distribution	376
8.2.1	The bivariate normal distribution	377
8.3	Regression towards the mean	378
8.4	Relationship between correlation and regression	380
8.4.1	Values of x are assumed to be measured without error and can be preselected	381
8.4.2	The data pairs are assumed to be a random sample from a bivariate normal distribution	381
8.5	Fitting a linear relationship when both variables are measured with error	383
8.6	Calibration lines	386
8.7	Intrinsically linear models	389
8.8	Summary	393
8.8.1	Notation	393
8.8.2	Summary of main results	393
8.8.3	MATLAB and R commands	394
8.9	Exercises	395

9	Multiple regression	403
9.1	Introduction	403
9.2	Multivariate data	404
9.3	Multiple regression model	405
9.3.1	The linear model	405
9.3.2	Random vectors	406
9.3.2.1	Linear transformations of a random vector	406
9.3.2.2	Multivariate normal distribution	407
9.3.3	Matrix formulation of the linear model	407
9.3.4	Geometrical interpretation	407
9.4	Fitting the model	408
9.4.1	Principle of least squares	408
9.4.2	Multivariate calculus - Three basic results	409
9.4.3	The least squares estimator of the coefficients	410
9.4.4	Estimating the coefficients	411
9.4.5	Estimating the standard deviation of the errors	416
9.4.6	Standard errors of the estimators of the coefficients	417
9.5	Assessing the fit	418
9.5.1	The residuals	419
9.5.2	R-squared	420
9.5.3	F-statistic	421
9.5.4	Cross validation	422
9.6	Predictions	422
9.7	Building multiple regression models	424
9.7.1	Interactions	424
9.7.2	Categorical variables	428
9.7.3	F-test for an added set of variables	433
9.7.4	Quadratic terms	440
9.7.5	Guidelines for fitting regression models	447
9.8	Time series	450
9.8.1	Introduction	450
9.8.2	Aliasing and sampling intervals	450
9.8.3	Fitting a trend and seasonal variation with regression	451
9.8.4	Auto-covariance and auto-correlation	456
9.8.4.1	Defining auto-covariance for a stationary times series model	457
9.8.4.2	Defining sample auto-covariance and the correlogram	458
9.8.5	Auto-regressive models	459
9.8.5.1	AR(1) and AR(2) models	460
9.9	Non-linear least squares	465
9.10	Generalized linear model	468
9.10.1	Logistic regression	468
9.10.2	Poisson regression	470
9.11	Summary	474
9.11.1	Notation	474
9.11.2	Summary of main results	474
9.11.3	MATLAB and R commands	475
9.12	Exercises	476

10 Statistical quality control	491
10.1 Continuous improvement	491
10.1.1 Defining quality	491
10.1.2 Taking measurements	492
10.1.3 Avoiding rework	493
10.1.4 Strategies for quality improvement	494
10.1.5 Quality management systems	494
10.1.6 Implementing continuous improvement	495
10.2 Process stability	496
10.2.1 Runs chart	496
10.2.2 Histograms and box plots	499
10.2.3 Components of variance	501
10.3 Capability	510
10.3.1 Process capability index	510
10.3.2 Process performance index	511
10.3.3 One-sided process capability indices	512
10.4 Reliability	514
10.4.1 Introduction	514
10.4.1.1 Reliability of components	514
10.4.1.2 Reliability function and the failure rate	515
10.4.2 Weibull analysis	517
10.4.2.1 Definition of the Weibull distribution	517
10.4.2.2 Weibull quantile plot	518
10.4.2.3 Censored data	522
10.4.3 Maximum likelihood	524
10.4.4 Kaplan-Meier estimator of reliability	529
10.5 Acceptance sampling	530
10.6 Statistical quality control charts	533
10.6.1 Shewhart mean and range chart for continuous variables	533
10.6.1.1 Mean chart	533
10.6.1.2 Range chart	535
10.6.2 p-charts for proportions	538
10.6.3 c-charts for counts	539
10.6.4 Cumulative sum charts	542
10.6.5 Multivariate control charts	544
10.7 Summary	548
10.7.1 Notation	548
10.7.2 Summary of main results	548
10.7.3 MATLAB and R commands	550
10.8 Exercises	550
11 Design of experiments with regression analysis	559
11.1 Introduction	559
11.2 Factorial designs with factors at two levels	562
11.2.1 Full factorial designs	562
11.2.1.1 Setting up a 2^k design	562
11.2.1.2 Analysis of 2^k design	565
11.3 Fractional factorial designs	580
11.4 Central composite designs	585
11.5 Evolutionary operation (EVOP)	593
11.6 Summary	597

11.6.1	Notation	597
11.6.2	Summary of main results	597
11.6.3	MATLAB and R commands	598
11.7	Exercises	598
12	Design of experiments and analysis of variance	605
12.1	Introduction	605
12.2	Comparison of several means with one-way ANOVA	605
12.2.1	Defining the model	606
12.2.2	Limitation of multiple t-tests	606
12.2.3	One-way ANOVA	607
12.2.4	Testing H_0^O	610
12.2.5	Follow up procedure	610
12.3	Two factors at multiple levels	613
12.3.1	Two factors without replication (two-way ANOVA)	614
12.3.2	Two factors with replication (three-way ANOVA)	618
12.4	Randomized block design	621
12.5	Split plot design	626
12.6	Summary	636
12.6.1	Notation	636
12.6.2	Summary of main results	637
12.6.3	MATLAB and R commands	637
12.7	Exercises	638
13	Probability models	649
13.1	System reliability	649
13.1.1	Series system	649
13.1.2	Parallel system	650
13.1.3	k -out-of- n system	651
13.1.4	Modules	652
13.1.5	Duality	653
13.1.6	Paths and cut sets	655
13.1.7	Reliability function	656
13.1.8	Redundancy	658
13.1.9	Non-repairable systems	658
13.1.10	Standby systems	659
13.1.11	Common cause failures	661
13.1.12	Reliability bounds	661
13.2	Markov chains	662
13.2.1	Discrete Markov chain	663
13.2.2	Equilibrium behavior of irreducible Markov chains	667
13.2.3	Methods for solving equilibrium equations	670
13.2.4	Absorbing Markov chains	675
13.2.5	Markov chains in continuous time	681
13.3	Simulation of systems	684
13.3.1	The simulation procedure	685
13.3.2	Drawing inference from simulation outputs	689
13.3.3	Variance reduction	692
13.4	Summary	694
13.4.1	Notation	694
13.4.2	Summary of main results	694

13.5 Exercises	696
14 Sampling strategies	699
14.1 Introduction	699
14.2 Simple random sampling from a finite population	702
14.2.1 Finite population correction	702
14.2.2 Randomization theory	703
14.2.2.1 Defining the simple random sample	703
14.2.2.2 Mean and variance of sample mean	704
14.2.2.3 Mean and variance of estimator of population total	705
14.2.3 Model based analysis	707
14.2.4 Sample size	708
14.3 Stratified sampling	708
14.3.1 Principle of stratified sampling	709
14.3.2 Estimating the population mean and total	709
14.3.3 Optimal allocation of the sample over strata	711
14.4 Multi-stage sampling	713
14.5 Quota sampling	716
14.6 Ratio estimators and regression estimators	716
14.6.1 Introduction	716
14.6.2 Regression estimators	716
14.6.3 Ratio estimator	716
14.7 Calibration of the unit cost data base	718
14.7.1 Sources of error in an AMP	718
14.7.2 Calibration factor	719
14.8 Summary	721
14.8.1 Notation	721
14.8.2 Summary of main results	721
14.9 Exercises	722
Appendix A - Notation	727
A.1 General	727
A.2 Probability	727
A.3 Statistics	728
A.4 Probability distributions	729
Appendix B - Glossary	731
Appendix C - Getting started in R	745
C.1 Installing R	745
C.2 Using R as a calculator	745
C.3 Setting the path	747
C.4 R scripts	747
C.5 Data entry	747
C.5.1 From keyboard	747
C.5.2 From a file	748
C.5.2.1 Single variable	748
C.5.2.2 Several variables	748
C.6 R vectors	749
C.7 User defined functions	750
C.8 Matrices	750

C.9	Loops and conditionals	751
C.10	Basic plotting	752
C.11	Installing packages	753
C.12	Creating time series objects	753
Appendix D - Getting started in MATLAB		755
D.1	Installing MATLAB	755
D.2	Using MATLAB as a calculator	755
D.3	Setting the path	756
D.4	MATLAB scripts (m-files)	756
D.5	Data entry	757
D.5.1	From keyboard	757
D.5.2	From a file	757
D.5.2.1	Single variable	757
D.5.2.2	Several variables	758
D.6	MATLAB vectors	758
D.7	User defined functions	761
D.8	Matrices	761
D.9	Loops and conditionals	761
D.10	Basic plotting	763
D.11	Creating time series objects	764
Appendix E - Experiments		765
E.1	How good is your probability assessment?	765
E.1.1	Objectives	765
E.1.2	Experiment	765
E.1.3	Question sets	765
E.1.4	Discussion	767
E.1.5	Follow up questions	767
E.2	Buffon's needle	767
E.2.1	Objectives	767
E.2.2	Experiment	767
E.2.3	Questions	768
E.2.4	Computer simulation	768
E.2.5	Historical note	768
E.3	Robot rabbit	768
E.3.1	Objectives	768
E.3.2	Experiment	769
E.3.3	Data	770
E.3.4	Discussion	770
E.3.5	Follow up question	772
E.4	Use your braking brains	772
E.4.1	Objectives	772
E.4.2	Experiment	772
E.4.3	Discussion	772
E.5	Predicting descent time from payload	773
E.5.1	Objectives	773
E.5.2	Experiment	773
E.5.3	Discussion	774
E.5.4	Follow up question	774
E.6	Company efficiency, resources and teamwork	774

E.6.1	Objectives	774
E.6.2	Experiment	774
E.6.3	Discussion	776
E.7	Factorial experiment – reaction times by distraction, dexterity and distinctness	776
E.7.1	Aim	776
E.7.2	Experiment	776
E.7.3	Analysis	776
E.7.4	Discussion	777
E.7.5	Follow up questions	777
E.8	Weibull analysis of cycles to failure	778
E.8.1	Aim	778
E.8.2	Experiment	778
E.8.3	Weibull plot	778
E.8.4	Discussion	779
E.9	Control or tamper?	779
E.10	Where is the summit?	781
References		783
Index		789