Medical Statistics at a Glance

Third Edition

Aviva Petrie
Caroline Sabin

Check out the workbook as well!

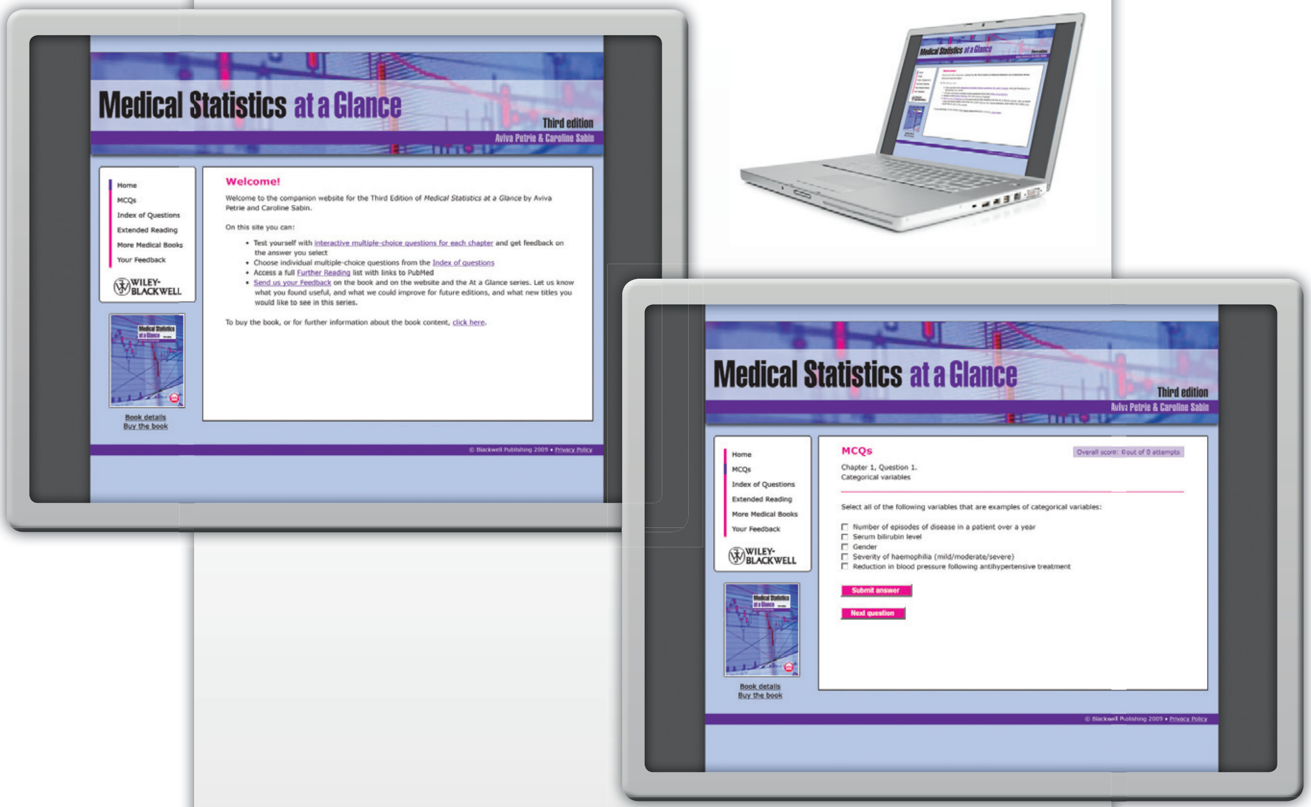
with self-test and more

WILEY Blackwell

Medical Statistics at a Glance

A companion website for this book is available at:

www.medstatsaag.com



Welcome!

Welcome to the companion website for the Third Edition of *Medical Statistics at a Glance* by Aviva Petrie and Caroline Sabin.

On this site you can:

- Test yourself with interactive multiple-choice questions for each chapter and get feedback on the answer you select
- Choose individual multiple-choice questions from the Index of questions
- Access a full Further Reading list with links to PubMed
- Send us your Feedback on the book and on the website and the At a Glance series. Let us know what you found useful, and what we could improve for future editions, and what new titles you would like to see in this series.

To buy the book, or for further information about the book content, click here.

MCQs

Overall score: 0 out of 0 attempts

Chapter 1, Question 1.
Categorical variables

Select all of the following variables that are examples of categorical variables:

□ Number of episodes of disease in a patient over a year
□ Serum bilirubin level
□ Gender
□ Severity of haemophilia (mild/moderate/severe)
□ Reduction in blood pressure following antihypertensive treatment

The site includes:

- Interactive multiple-choice questions for each chapter
- Feedback on each of the answers you select
- Extended reading lists
- A chance to send us your feedback

Medical Statistics at a Glance

Aviva Petrie

Head of Biostatistics Unit and Senior Lecturer
UCL Eastman Dental Institute
256 Gray's Inn Road
London WC1X 8LD *and*
Honorary Lecturer in Medical Statistics
Medical Statistics Unit
London School of Hygiene and Tropical Medicine
Keppel Street
London WC1E 7HT

Caroline Sabin

Professor of Medical Statistics and Epidemiology
Research Department of Infection and Population Health
Division of Population Health
University College London Medical School
Royal Free Campus
Rowland Hill Street
London NW3 2PF

Third edition

WILEY Blackwell

Registered office: John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices: 9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting a specific method, diagnosis, or treatment by health science practitioners for any particular patient. The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of fitness for a particular purpose. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. Readers should consult with a specialist where appropriate. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication Data

Petrie, Aviva.
Medical statistics at a glance / Aviva Petrie, Caroline Sabin. – 3rd ed.
p.; cm. – (At a glance series)
Includes bibliographical references and index.
ISBN 978-1-4051-8051-1 (alk. paper)
1. Medical statistics. I. Sabin, Caroline. II. Title. III. Series: At a glance series (Oxford, England)
[DNLM: 1. Statistics as Topic. 2. Research Design. WA 950 P495m 2009]

R853.S7P476 2009
610.72′7–dc22

2008052096

A catalogue record for this book is available from the British Library.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Set in 9.5 on 12 pt Times by Toppan Best-set Premedia Limited

1 2009

Contents

Handling data

Sampling and estimation

Study design

Hypothesis testing

Basic techniques for analysing data

Numerical data

Categorical data

Regression and correlation

Important considerations

Additional chapters

Appendices

Preface

Medical Statistics at a Glance is directed at undergraduate medical students, medical researchers, postgraduates in the biomedical disciplines and at pharmaceutical industry personnel. All of these individuals will, at some time in their professional lives, be faced with quantitative results (their own or those of others) which will need to be critically evaluated and interpreted, and some, of course, will have to pass that dreaded statistics exam! A proper understanding of statistical concepts and methodology is invaluable for these needs. Much as we should like to fire the reader with an enthusiasm for the subject of statistics, we are pragmatic. Our aim in this new edition, as it was in the earlier editions, is to provide the student and the researcher, as well as the clinician encountering statistical concepts in the medical literature, with a book which is sound, easy to read, comprehensive, relevant, and of useful practical application.

We believe *Medical Statistics at a Glance* will be particularly helpful as an adjunct to statistics lectures and as a reference guide. The structure of this third edition is the same as that of the first two editions. In line with other books in the *At a Glance* series, we lead the reader through a number of self-contained two-, three- or occasionally four-page chapters, each covering a different aspect of medical statistics. We have learned from our own teaching experiences and have taken account of the difficulties that our students have encountered when studying medical statistics. For this reason, we have chosen to limit the theoretical content of the book to a level that is sufficient for understanding the procedures involved, yet which does not overshadow the practicalities of their execution.

Medical statistics is a wide-ranging subject covering a large number of topics. We have provided a basic introduction to the underlying concepts of medical statistics and a guide to the most commonly used statistical procedures. Epidemiology is closely allied to medical statistics. Hence some of the main issues in epidemiology, relating to study design and interpretation, are discussed. Also included are chapters which the reader may find useful only occasionally, but which are, nevertheless, fundamental to many areas of medical research; for example, evidence-based medicine, systematic reviews and meta-analysis, survival analysis, Bayesian methods and the development of prognostic scores. We have explained the principles underlying these topics so that the reader will be able to understand and interpret the results from them when they are presented in the literature.

The chapter titles of this third edition are identical to those of the second edition, apart from Chapter 34 (now called 'Bias and confounding' instead of 'Issues in statistical modelling'); in addition, we have added a new chapter (Chapter 46 – 'Developing prognostic scores'). Some of the first 45 chapters remain unaltered in this new edition and some have relatively minor changes which accommodate recent advances, cross-referencing or re-organization of the new material. We have expanded many chapters; for example, we have included a section on multiple comparisons (Chapter 12), provided more information on different study designs, including multicentre studies (Chapter 12) and sequential trials (Chapter 14), emphasized the importance of study management (Chapters 15 and 16), devoted greater space to receiver operating characteristic (ROC) curves (Chapters 30, 38 and 46), supplied more details of how to check the assumptions underlying a logistic regression analysis (Chapter 30) and explored further some of the different methods to remove confounding in observational studies (Chapter 34). We have also reorganized some of the material. The brief introduction to bias in Chapter 12 in the second edition has been omitted from that chapter in the third edition and moved to Chapter 34, which covers this topic in greater depth. A discussion of 'interaction' is currently in Chapter 33 and the section on prognostic indices is now much expanded and contained in the new Chapter 46.

New to this third edition is a set of learning objectives for each chapter, all of which are displayed together at the beginning of the book. Each set provides a framework for evaluating understanding and progress. If you are able to complete all the bulleted tasks in a chapter satisfactorily, you will have mastered the concepts in that chapter.

As in previous editions, the description of most of the statistical techniques is accompanied by an example illustrating its use. We have generally obtained the data for these examples from collaborative studies in which we or colleagues have been involved; in some instances, we have used real data from published papers. Where possible, we have used the same data set in more than one chapter to reflect the reality of data analysis, which is rarely restricted to a single technique or approach. Although we believe that formulae should be provided and the logic of the approach explained as an aid to understanding, we have avoided showing the details of complex calculations – most readers will have access to computers and are unlikely to perform any but the simplest calculations by hand.

We consider that it is particularly important for the reader to be able to interpret output from a computer package. We have therefore chosen, where applicable, to show results using extracts from computer output. In some instances, where we believe individuals may have difficulty with its interpretation, we have included (Appendix C) and annotated the complete computer output from an analysis of a data set. There are many statistical packages in common use; to give the reader an indication of how output can vary, we have not restricted the output to a particular package and have, instead, used three well-known ones – SAS, SPSS and Stata.

There is extensive cross-referencing throughout the text to help the reader link the various procedures. A basic set of statistical tables is contained in Appendix A. Neave, H.R. (1995) *Elemementary Statistical Tables*, Routledge: London, and Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th edition, Blackwell Publishing: Oxford, amongst others, provide fuller versions if the reader requires more precise results for hand calculations. The glossary of terms in Appendix D provides readily accessible explanations of commonly used terminology.

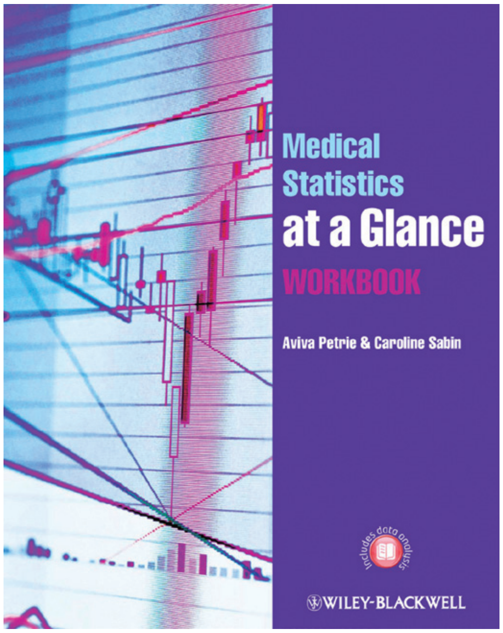
We know that one of the greatest difficulties facing non-statisticians is choosing the appropriate technique. We have therefore produced two flow charts which can be used both to aid the decision as to what method to use in a given situation and to locate a particular technique in the book easily. These flow charts are located in the plate section.

The reader may find it helpful to assess his/her progress in self-directed learning by attempting the interactive exercises on our website (www.medstatsaag.com). This website also contains a full set of references (some of which are linked directly to Medline) to supplement the references quoted in the text and provide useful background information for the examples. For those readers who wish to gain a greater insight into particular areas of medical statistics, we can recommend the following books:

- Altman, D.G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall/CRC.
- Armitage, P., Berry, G. and Matthews, J.F.N. (2001) *Statistical Methods in Medical Research.* 4th edition. Oxford: Blackwell Science.
- Kirkwood, B.R. and Sterne, J.A.C. (2003) *Essential Medical Statistics.* 2nd Edn. Oxford: Blackwell Publishing.
- Pocock, S.J. (1983) *Clinical Trials: A Practical Approach.* Chichester: Wiley.

We are extremely grateful to Mark Gilthorpe and Jonathan Sterne who made invaluable comments and suggestions on aspects of the second edition, and to Richard Morris, Fiona Lampe, Shak Hajat and Abul Basar for their counsel on the first edition. We wish to thank everyone who has helped us by providing data for the examples. Naturally, we take full responsibility for any errors that remain in the text or examples. We should also like to thank Mike, Gerald, Nina, Andrew and Karen who tolerated, with equanimity, our preoccupation with the first two editions and lived with us through the trials and tribulations of this third edition.

Aviva Petrie
Caroline Sabin
London



Learning objectives

By the end of the relevant chapter you should be able to:

1 Types of data
- Distinguish between a sample and a population
- Distinguish between categorical and numerical data
- Describe different types of categorical and numerical data
- Explain the meaning of the terms: variable, percentage, ratio, quotient, rate, score
- Explain what is meant by censored data

2 Data entry
- Describe different formats for entering data on to a computer
- Outline the principles of questionnaire design
- Distinguish between single-coded and multi-coded variables
- Describe how to code missing values

3 Error checking and outliers
- Describe how to check for errors in data
- Outline the methods of dealing with missing data
- Define an outlier
- Explain how to check for and handle outliers

4 Displaying data diagrammatically
- Explain what is meant by a frequency distribution
- Describe the shape of a frequency distribution
- Describe the following diagrams: (segmented) bar or column chart, pie chart, histogram, dot plot, stem-and-leaf plot, box-and-whisker plot, scatter diagram
- Explain how to identify outliers from a diagram in various situations
- Describe the situations when it is appropriate to use connecting lines in a diagram

5 Describing data: the 'average'
- Explain what is meant by an average
- Describe the appropriate use of each of the following types of average: arithmetic mean, mode, median, geometric mean, weighted mean
- Explain how to calculate each type of average
- List the advantages and disadvantages of each type of average

6 Describing data: the 'spread'
- Define the following terms: percentile, decile, quartile, median, and explain their inter-relationship
- Explain what is meant by a reference interval/range, also called the normal range
- Define the following measures of spread: range, interdecile range, variance, standard deviation (SD), coefficient of variation
- List the advantages and disadvantages of the various measures of spread
- Distinguish between intra- and inter-subject variation

7 Theoretical distributions: the Normal distribution
- Define the terms: probability, conditional probability
- Distinguish between the subjective, frequentist and *a priori* approaches to calculating a probability
- Define the addition and multiplication rules of probability
- Define the terms: random variable, probability distribution, parameter, statistic, probability density function
- Distinguish between a discrete and continuous probability distribution and list the properties of each
- List the properties of the Normal and the Standard Normal distributions
- Define a Standardized Normal Deviate (SND)

8 Theoretical distributions: other distributions
- List the important properties of the *t*-, Chi-squared, *F*- and Lognormal distributions
- Explain when each of these distributions is particularly useful
- List the important properties of the Binomial and Poisson distributions
- Explain when the Binomial and Poisson distributions are each particularly useful

9 Transformations
- Describe situations in which transforming data may be useful
- Explain how to transform a data set
- Explain when to apply and what is achieved by the logarithmic, square root, reciprocal, square and logit transformations
- Describe how to interpret summary measures derived from log transformed data after they have been back-transformed to the original scale

10 Sampling and sampling distributions
- Explain what is meant by statistical inference and sampling error
- Explain how to obtain a representative sample
- Distinguish between point and interval estimates of a parameter
- List the properties of the sampling distribution of the mean
- List the properties of the sampling distribution of the proportion
- Explain what is meant by a standard error
- State the relationship between the standard error of the mean (SEM) and the standard deviation (SD)
- Distinguish between the uses of the SEM and the SD

11 Confidence intervals
- Interpret a confidence interval (CI)
- Calculate a confidence interval for a mean
- Calculate a confidence interval for a proportion
- Explain the term 'degrees of freedom'
- Explain what is meant by bootstrapping and jackknifing

12 Study design I
- Distinguish between experimental and observational studies, and between cross-sectional and longitudinal studies
- Explain what is meant by the unit of observation
- Explain the terms: control group, epidemiological study, cluster randomized trial, ecological study, multicentre study, survey, census
- List the criteria for assessing causality in observational studies
- Describe the time course of cross-sectional, repeated cross-sectional, cohort, case–control and experimental studies
- List the typical uses of these various types of study
- Distinguish between prevalence and incidence

13 Study design II

- Describe how to increase the precision of an estimate
- Explain the principles of blocking (stratification)
- Distinguish between parallel and cross-over designs
- Describe the features of a factorial experiment
- Explain what is meant by an interaction between factors
- Explain the following terms: study endpoint, surrogate marker, composite endpoint

14 Clinical trials

- Define 'clinical trial' and distinguish between Phase I/II and Phase III clinical trials
- Explain the importance of a control treatment and distinguish between positive and negative controls
- Explain what is meant by a placebo
- Distinguish between primary and secondary endpoints
- Explain why it is important to randomly allocate individuals to treatment groups and describe different forms of randomization
- Explain why it is important to incorporate blinding (masking)
- Distinguish between double- and single-blind trials
- Discuss the ethical issues arising from a randomized controlled trial (RCT)
- Explain the principles of a sequential trial
- Distinguish between on-treatment analysis and analysis by intention-to-treat (ITT)
- Describe the contents of a protocol
- Apply the CONSORT Statement guidelines

15 Cohort studies

- Describe the aspects of a cohort study
- Distinguish between fixed and dynamic cohorts
- Explain the terms: historical cohort, risk factor, healthy entrant effect, clinical cohort
- List the advantages and disadvantages of a cohort study
- Describe the important aspects of cohort study management
- Calculate and interpret a relative risk

16 Case–control studies

- Describe the features of a case–control study
- Distinguish between incident and prevalent cases
- Describe how controls may be selected for a case–control study
- Explain how to analyse an unmatched case–control study by calculating and interpreting an odds ratio
- Describe the features of a matched case–control study
- Distinguish between frequency matching and pairwise matching
- Explain when an odds ratio can be used as an estimate of the relative risk
- List the advantages and disadvantages of a case–control study

17 Hypothesis testing

- Define the terms: null hypothesis, alternative hypothesis, one- and two-tailed test, test statistic, P-value, significance level
- List the five steps in hypothesis testing
- Explain how to use the P-value to make a decision about rejecting or not rejecting the null hypothesis
- Explain what is meant by a non-parametric (distribution-free) test and explain when such a test should be used
- Explain how a confidence interval can be used to test a hypothesis
- Distinguish between superiority, equivalence and non-inferiority studies
- Describe the approach used in equivalence and non-inferiority tests

18 Errors in hypothesis testing

- Explain what is meant by an effect of interest
- Distinguish between Type I and Type II errors
- State the relationship between the Type II error and power
- List the factors that affect the power of a test and describe their effects on power
- Explain why it is inappropriate to perform many hypothesis tests in a study
- Describe different situations which involve multiple comparisons within a data set and explain how the difficulties associated with multiple comparisons may be resolved in each situation
- Explain what is achieved by a *post hoc* test
- Outline the Bonferroni approach to multiple hypothesis testing

19 Numerical data: a single group

- Explain the rationale of the one-sample t-test
- Explain how to perform the one-sample t-test
- State the assumption underlying the test and explain how to proceed if it is not satisfied
- Explain how to use an appropriate confidence interval to test a hypothesis about the mean
- Explain the rationale of the sign test
- Explain how to perform the sign test

20 Numerical data: two related groups

- Describe different circumstances in which two groups of data are related
- Explain the rationale of the paired t-test
- Explain how to perform the paired t-test
- State the assumption underlying the test and explain how to proceed if it is not satisfied
- Explain the rationale of the Wilcoxon signed ranks test
- Explain how to perform the Wilcoxon signed ranks test

21 Numerical data: two unrelated groups

- Explain the rationale of the unpaired (two-sample) t-test
- Explain how to perform the unpaired t-test
- List the assumptions underlying this test and explain how to check them and proceed if they are not satisfied
- Use an appropriate confidence interval to test a hypothesis about the difference between two means
- Explain the rationale of the Wilcoxon rank sum test
- Explain how to perform the Wilcoxon rank sum test
- Explain the relationship between the Wilcoxon rank sum test and the Mann–Whitney U test

22 Numerical data: more than two groups

- Explain the rationale of the one-way analysis of variance (ANOVA)
- Explain how to perform a one-way ANOVA
- Explain why a *post hoc* comparison method should be used if a one-way ANOVA produces a significant result and name some different *post hoc* methods
- List the assumptions underlying the one-way ANOVA and explain how to check them and proceed if they are not satisfied

- Explain the rationale of the Kruskal–Wallis test
- Explain how to perform the Kruskal–Wallis test

23 Categorical data: a single proportion

- Explain the rationale of a test, based on the Normal distribution, which can be used to investigate whether a proportion takes a particular value.
- Explain how to perform this test
- Explain why a continuity correction should be used in this test
- Explain how the sign test can be used to test a hypothesis about a proportion
- Explain how to perform the sign test to test a hypothesis about a proportion

24 Categorical data: two proportions

- Explain the terms: contingency table, cell frequency, marginal total, overall total, observed frequency, expected frequency
- Explain the rationale of the Chi-squared test to compare proportions in two unrelated groups
- Explain how to perform the Chi-squared test to compare two independent proportions
- Calculate the confidence interval for the difference in the proportions in two unrelated groups and use it to compare them
- State the assumption underlying the Chi-squared test to compare proportions and explain how to proceed if this assumption is not satisfied
- Describe the circumstances under which Simpson's paradox may occur and explain what can be done to avoid it
- Explain the rationale of McNemar's test to compare the proportions in two related groups
- Explain how to perform McNemar's test
- Calculate the confidence interval for the difference in two proportions in paired groups and use the confidence interval to compare them

25 Categorical data: more than two categories

- Describe an $r \times c$ contingency table
- Explain the rationale of the Chi-squared test to assess the association between one variable with r categories and another variable with c categories
- Explain how to perform the Chi-squared test to assess the association between two variables using data displayed in an $r \times c$ contingency table
- State the assumption underlying this Chi-squared test and explain how to proceed if this assumption is not satisfied
- Explain the rationale of the Chi-squared test for trend in a $2 \times k$ contingency table
- Explain how to perform the Chi-squared test for trend in a $2 \times k$ contingency table

26 Correlation

- Describe a scatter diagram
- Define and calculate the Pearson correlation coefficient and list its properties
- Explain when it is inappropriate to calculate the Pearson correlation coefficient if investigating the relationship between two variables
- Explain how to test the null hypothesis that the true Pearson correlation coefficient is zero
- Calculate the 95% confidence interval for the Pearson correlation coefficient
- Describe the use of the square of the Pearson correlation coefficient
- Explain when and how to calculate the Spearman rank correlation coefficient
- List the properties of the Spearman rank correlation coefficient

27 The theory of linear regression

- Explain the terms commonly used in regression analysis: dependent variable, explanatory variable, regression coefficient, intercept, gradient, residual
- Define the simple (univariable) regression line and interpret its coefficients
- Explain the principles of the method of least squares
- List the assumptions underlying a simple linear regression analysis
- Describe the features of an analysis of variance (ANOVA) table produced by a linear regression analysis
- Explain how to use the ANOVA table to assess how well the regression line fits the data (goodness of fit) and test the null hypothesis that the true slope of the regression line is zero.
- Explain what is meant by regression to the mean

28 Performing a linear regression analysis

- Explain how to use residuals to check the assumptions underlying a linear regression analysis
- Explain how to proceed in a regression analysis if one or more of the assumptions are not satisfied
- Define the terms 'outlier' and 'influential point' and explain how to deal with each of them
- Explain how to assess the goodness of fit of a regression model
- Calculate the 95% confidence interval for the slope of a regression line
- Describe two methods for testing the null hypothesis that the true slope is zero
- Explain how to use the regression line for prediction
- Explain how to (1) centre and (2) scale an explanatory variable in a regression analysis
- Explain what is achieved by centring and scaling.

29 Multiple linear regression

- Explain the terms: covariate, partial regression coefficient, collinearity
- Define the multiple (multivariable) linear regression equation and interpret its coefficients
- Give three reasons for performing a multiple regression analysis
- Explain how to create dummy variables to allow nominal and ordinal categorical explanatory variables with more than two categories of response to be incorporated in the model
- Explain what is meant by the reference category when fitting models that include categorical explanatory variables
- Describe how multiple regression analysis can be used as a form of analysis of covariance
- Give a rule of thumb for deciding on the maximum number of explanatory variables in a multiple regression equation
- Use computer output from a regression analysis to assess the goodness of fit of the model, and test the null hypotheses that all the partial regression coefficients are zero and that each partial regression coefficient is zero
- Explain the relevance of residuals, leverage and Cook's distance in identifying outliers and influential points

30 Binary outcomes and logistic regression

- Explain why multiple linear regression analysis cannot be used for a binary outcome variable
- Define the logit of a proportion
- Define the multiple logistic regression equation
- Interpret the exponential of a logistic regression coefficient
- Calculate, from a logistic regression equation, the probability that a particular individual will have the outcome of interest
- Describe two ways of assessing whether a logistic regression coefficient is statistically significant
- Describe various ways of testing the overall model fit, assessing predictive efficiency and investigating the underlying assumptions of a logistic regression analysis
- Explain when the odds ratio is greater than and when it is less than the relative risk
- Explain the use of the following types of logistic regression: multinomial, ordinal, conditional

31 Rates and Poisson regression

- Define a rate and describe its features
- Distinguish between a rate and a risk, and between an incidence rate and a mortality rate
- Define a relative rate and explain when it is preferred to a relative risk
- Explain when it is appropriate to use Poisson regression
- Define the Poisson regression equation and interpret the exponential of a Poisson regression coefficient
- Calculate, from the Poisson regression equation, the event rate for a particular individual
- Explain the use of an offset in a Poisson regression analysis
- Explain how to perform a Poisson regression analysis with (1) grouped data and (2) variables that change over time
- Explain the meaning and the consequences of extra-Poisson dispersion
- Explain how to identify extra-Poisson dispersion in a Poisson regression analysis

32 Generalized linear models

- Define the equation of the generalized linear model (GLM)
- Explain the terms 'link function' and 'identity link'
- Specify the link functions for the logistic and Poisson regression models
- Explain the term 'likelihood' and the process of maximum likelihood estimation (MLE)
- Explain the terms: saturated model, likelihood ratio
- Explain how the likelihood ratio statistic (LRS), i.e. the deviance or –2log likelihood, can be used to:
  - ○ assess the adequacy of fit of a model
  - ○ compare two models when one is nested within the other
  - ○ assess whether all the parameters associated with the covariates of a model are zero (i.e. the model Chi-square)

33 Explanatory variables in statistical models

- Explain how to test the significance of a nominal explanatory variable in a statistical model when the variable has more than two categories
- Describe two ways of incorporating an ordinal explanatory variable into a model when the variable has more than two categories, and:
  - ○ state the advantages and disadvantages of each approach
  - ○ explain how each approach can be used to test for a linear trend
- Explain how to check the linearity assumption in multiple, Poisson and logistic regression analyses
- Describe three ways of dealing with non-linearity in a regression model
- Explain why a model should not be over-fitted and how to avoid it
- Explain when it is appropriate to use automatic selection procedures to select the optimal explanatory variables
- Describe the principles underlying various automatic selection procedures
- Explain why automatic selection procedures should be used with caution
- Explain the meaning of interaction and collinearity
- Explain how to test for an interaction in a regression analysis
- Explain how to detect collinearity

34 Bias and confounding

- Explain what is meant by bias
- Explain what is meant by selection bias, information bias, funding bias and publication bias
- Describe different forms of bias which comprise either selection bias or information bias
- Explain what is meant by the ecological fallacy
- Explain what is meant by confounding and what steps may be taken to deal with confounding at the design stage of a study
- Describe various methods of dealing with confounding at the analysis stage of a study
- Explain the meaning of a propensity score
- Discuss the advantages and disadvantages of the various methods of dealing with confounding at the analysis stage
- Explain why confounding is a particular issue in a non-randomized study
- Explain the following terms: causal pathway, intermediate variable, time-varying confounding

35 Checking assumptions

- Name two tests and describe two diagrams that can be used to assess whether data are Normally distributed
- Explain the terms homogeneity and heterogeneity of variance
- Name two tests that can be used to assess the equality of two or more variances
- Explain how to perform the variance ratio *F*-test to compare two variances
- Explain how to proceed if the assumptions under a proposed analysis are not satisfied
- Explain what is meant by a robust analysis
- Explain what is meant by a sensitivity analysis
- Provide examples of different sensitivity analyses

36 Sample size calculations

- Explain why it is necessary to choose an optimal sample size for a proposed study
- Specify the quantities that affect sample size and describe their effects on it
- Name five approaches to calculating the optimal sample size of a study
- Explain how information from an internal pilot study may be used to revise calculations of the optimal sample size
- Explain how to use Altman's nomogram to determine the optimal sample size for a proposed *t*-test (unpaired and paired) and Chi-squared test

- Explain how to use Lehr's formula for sample size calculations for the comparison of two means and of two proportions in independent groups
- Write an appropriate power statement
- Explain how to adjust the sample size for losses to follow-up and/or if groups of different sizes are required
- Explain how to increase the power of a study for a fixed sample size

37 Presenting results

- Explain how to report numerical results
- Describe the important features of good tables and diagrams
- Explain how to report the results of a hypothesis test
- Explain how to report the results of a regression analysis
- Indicate how complex statistical analyses should be reported
- Locate and follow the guidelines for reporting different types of study

38 Diagnostic tools

- Distinguish between a diagnostic test and a screening test and explain when each is appropriate
- Define 'reference range' and explain how it is used
- Describe two ways in which a reference range can be calculated
- Define the terms: true positive, false positive, true negative, false negative
- Estimate (with a 95% confidence interval) and interpret each of the following: prevalence, sensitivity, specificity, positive predictive value, negative predictive value
- Construct a receiver operating characteristic (ROC) curve
- Explain how the ROC curve can be used to choose an optimal cut-off for a diagnostic test
- Explain how the area under the ROC curve can be used to assess the ability of a diagnostic test to discriminate between individuals with and without a disease and to compare two diagnostic tests
- Calculate and interpret the likelihood ratio for a positive and for a negative test result if the sensitivity and specificity of the test are known.

39 Assessing agreement

- Distinguish between measurement variability and measurement error
- Distinguish between systematic and random error
- Distinguish between reproducibility and repeatability
- Calculate and interpret Cohen's kappa for assessing the agreement between paired categorical responses
- Explain what a weighted kappa is and when it can be determined
- Explain how to test for a systematic effect when comparing pairs of numerical responses
- Explain how to perform a Bland and Altman analysis to assess the agreement between paired numerical responses and interpret the limits of agreement
- Explain how to calculate and interpret the British Standards Institution reproducibility/repeatability coefficient
- Explain how to calculate and interpret the intraclass correlation coefficient and Lin's concordance correlation coefficient in a method comparison study
- Explain why it is inappropriate to calculate the Pearson correlation coefficient to assess the agreement between paired numerical responses

40 Evidence-based medicine

- Define evidence-based medicine (EBM)
- Describe the hierarchy of evidence associated with various study designs
- List the six steps involved in performing EBM to assess the efficacy of a new treatment, and describe the important features of each step
- Explain the term number needed to treat (NNT)
- Explain how to calculate the NNT
- Explain how to assess the effect of interest if the main outcome variable is binary
- Explain how to assess the effect of interest if the main outcome variable is numerical
- Explain how to decide whether the results of an investigation are important

41 Methods for clustered data

- Describe, with examples, clustered data in a two-level structure
- Describe how such data may be displayed graphically
- Describe the effect of ignoring repeated measures in a statistical analysis
- Explain how summary measures may be used to compare groups of repeated measures data
- Name two other methods which are appropriate for comparing groups of repeated measures data
- Explain why a series of two-sample *t*-tests is inappropriate for analysing such data

42 Regression methods for clustered data

- Outline the following approaches to analysing clustered data in a two-level structure: aggregate level analysis, analysis using robust standard errors, random effects (hierarchical, multilevel, mixed, cluster-specific, cross-sectional) model, generalized estimating equations (GEE)
- List the advantages and disadvantages of each approach
- Distinguish between a random intercepts and a random slopes random effects model
- Explain how to calculate and interpret the intraclass correlation coefficient (ICC) to assess the effect of clustering in a random effects model
- Explain how to use the likelihood ratio test to assess the effect of clustering

43 Systematic reviews and meta-analysis

- Define a systematic review and explain what it achieves
- Describe the Cochrane Collaboration
- Define a meta-analysis and list its advantages and disadvantages
- List the four steps involved in performing a meta-analysis
- Distinguish between statistical and clinical heterogeneity
- Explain how to test for statistical homogeneity
- Explain how to estimate the average effect of interest in a meta-analysis if there is evidence of statistical heterogeneity
- Explain the terms: fixed effects meta-analysis, random effects meta-analysis, meta-regression
- Distinguish between a forest plot and a funnel plot
- Describe ways of performing a sensitivity analysis after performing a meta-analysis

44 Survival analysis

- Explain why it is necessary to use special methods for analysing survival data

- Distinguish between the terms ‘right-censored data’ and ‘left-censored data’
- Describe a survival curve
- Distinguish between the Kaplan–Meier method and lifetable approaches to calculating survival probabilities
- Explain what the log-rank test is used for in survival analysis
- Explain the principles of the Cox proportional hazards regression model
- Explain how to obtain a hazard ratio (relative hazard) from a Cox proportional hazards regression model and interpret it
- List other regression models that may also be used to describe survival data
- Explain the problems associated with informative censoring and competing risks

45 Bayesian methods

- Explain what is meant by the frequentist approach to probability
- Explain the shortcomings of the frequentist approach to probability
- Explain the principles of Bayesian analysis
- List the disadvantages of the Bayesian approach
- Explain the terms: conditional probability, prior probability, posterior probability, likelihood ratio
- Express Bayes theorem in terms of odds
- Explain how to use Fagan’s nomogram to interpret a diagnostic test result in a Bayesian framework

46 Developing prognostic scores

- Define the term ‘prognostic score’
- Distinguish between a prognostic index and a risk score
- Outline different ways of deriving a prognostic score
- List the desirable features of a good prognostic score
- Explain what is meant by assessing overall score accuracy
- Describe how a classification table and the mean Briar score can be used to assess overall score accuracy
- Explain what is meant by assessing the ability of a prognostic score to discriminate between those that do and do not experience the event
- Describe how classifying individuals by their score, drawing an ROC curve and calculating Harrell’s *c* statistic can each be used to assess the ability of a prognostic score to discriminate between those that do and do not experience the event
- Explain what is meant by correct calibration of a prognostic score
- Describe how the Hosmer–Lemeshow goodness of fit test can be used to assess whether a prognostic score is correctly calibrated
- Explain what is meant by transportability of a prognostic score
- Describe various methods of internal and external validation of a prognostic score

MCQ 1 2 16

SQ 1

1 Types of data

Data and statistics

The purpose of most studies is to collect **data** to obtain information about a particular area of research. Our data comprise **observations** on one or more variables; any quantity that varies is termed a **variable**. For example, we may collect basic clinical and demographic information on patients with a particular illness. The variables of interest may include the sex, age and height of the patients.

Our data are usually obtained from a **sample** of individuals which represents the **population** of interest. Our aim is to condense these data in a meaningful way and extract useful information from them. **Statistics** encompasses the methods of collecting, summarizing, analysing and drawing conclusions from the data: we use statistical techniques to achieve our aim.

Data may take many different forms. We need to know what form every variable takes before we can make a decision regarding the most appropriate statistical methods to use. Each variable and the resulting data will be one of two types: **categorical** or **numerical** (Fig. 1.1).

Categorical (qualitative) data

These occur when each individual can only belong to one of a number of distinct categories of the variable.

- **Nominal data** – the categories are not ordered but simply have names. Examples include blood group (A, B, AB and O) and marital status (married/widowed/single, etc.). In this case, there is no reason to suspect that being married is any better (or worse) than being single!
- **Ordinal data** – the categories are ordered in some way. Examples include disease staging systems (advanced, moderate, mild, none) and degree of pain (severe, moderate, mild, none).

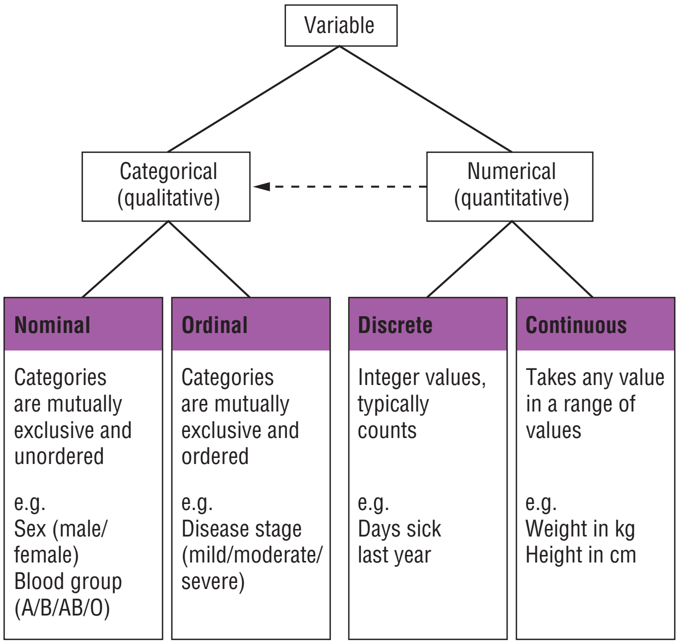

Variable
Categorical (qualitative)
Numerical (quantitative)
Nominal
Categories are mutually exclusive and unordered
e.g.
Sex (male/female)
Blood group (A/B/AB/O)
Ordinal
Categories are mutually exclusive and ordered
e.g.
Disease stage (mild/moderate/severe)
Discrete
Integer values, typically counts
e.g.
Days sick last year
Continuous
Takes any value in a range of values
e.g.
Weight in kg
Height in cm

Figure 1.1 Diagram showing the different types of variable.

A categorical variable is **binary** or **dichotomous** when there are only two possible categories. Examples include 'Yes/No', 'Dead/Alive' or 'Patient has disease/Patient does not have disease'.

Numerical (quantitative) data

These occur when the variable takes some numerical value. We can subdivide numerical data into two types.

- **Discrete data** – occur when the variable can only take certain whole numerical values. These are often counts of numbers of events, such as the number of visits to a GP in a particular year or the number of episodes of illness in an individual over the last five years.
- **Continuous data** – occur when there is no limitation on the values that the variable can take, e.g. weight or height, other than that which restricts us when we make the measurement.

Distinguishing between data types

We often use very different statistical methods depending on whether the data are categorical or numerical. Although the distinction between categorical and numerical data is usually clear, in some situations it may become blurred. For example, when we have a variable with a large number of ordered categories (e.g. a pain scale with seven categories), it may be difficult to distinguish it from a discrete numerical variable. The distinction between discrete and continuous numerical data may be even less clear, although in general this will have little impact on the results of most analyses. Age is an example of a variable that is often treated as discrete even though it is truly continuous. We usually refer to 'age at last birthday' rather than 'age', and therefore, a woman who reports being 30 may have just had her 30th birthday, or may be just about to have her 31st birthday.

Do not be tempted to record numerical data as categorical at the outset (e.g. by recording only the range within which each patient's age falls rather than his/her actual age) as important information is often lost. It is simple to convert numerical data to categorical data once they have been collected.

Derived data

We may encounter a number of other types of data in the medical field. These include:

- **Percentages** – These may arise when considering improvements in patients following treatment, e.g. a patient's lung function (forced expiratory volume in 1 second, FEV1) may increase by 24% following treatment with a new drug. In this case, it is the level of improvement, rather than the absolute value, which is of interest.
- **Ratios** or **quotients** – Occasionally you may encounter the ratio or quotient of two variables. For example, body mass index (BMI), calculated as an individual's weight (kg) divided by her/his height squared (m^2), is often used to assess whether s/he is over- or underweight.
- **Rates** – Disease rates, in which the number of disease events occurring among individuals in a study is divided by the total number of years of follow-up of all individuals in that study (Chapter 31), are common in epidemiological studies (Chapter 12).
- **Scores** – We sometimes use an arbitrary value, such as a score, when we cannot measure a quantity. For example, a series of responses to questions on quality of life may be summed to give some overall quality of life score on each individual.

All these variables can be treated as numerical variables for most analyses. Where the variable is derived using more than one value (e.g. the numerator and denominator of a percentage), it is important to record all of the values used. For example, a 10% improvement in a marker following treatment may have different clinical relevance depending on the level of the marker before treatment.

Censored data

We may come across **censored** data in situations illustrated by the following examples.

- If we measure laboratory values using a tool that can only detect levels above a certain cut-off value, then any values below this cut-off will not be detected, i.e. they are censored. For example, when measuring virus levels, those below the limit of detectability will often be reported as 'undetectable' or 'unquantifiable' even though there may be some virus in the sample. In this situation, if the lower cut-off of a tool is x, say, the results may be reported as '$<x$'. Similarly, some tools may only be able to reliably quantify levels below a certain cut-off value, say y; any measurements above that value will also be censored and the test result may be reported as '$>y$'.
- We may encounter censored data when following patients in a trial in which, for example, some patients withdraw from the trial before the trial has ended. This type of data is discussed in more detail in Chapter 44.

MCQ 1 3 4

SQ 1

2 Data entry

When you carry out any study you will almost always need to enter the data into a computer package. Computers are invaluable for improving the accuracy and speed of data collection and analysis, making it easy to check for errors, produce graphical summaries of the data and generate new variables. It is worth spending some time planning data entry – this may save considerable effort at later stages.

Formats for data entry

There are a number of ways in which data can be entered and stored on a computer. Most statistical packages allow you to enter data directly. However, the limitation of this approach is that often you cannot move the data to another package. A simple alternative is to store the data in either a spreadsheet or database package. Unfortunately, their statistical procedures are often limited, and it will usually be necessary to output the data into a specialist statistical package to carry out analyses.

A more flexible approach is to have your data available as an **ASCII** or **text** file. Once in an ASCII format, the data can be read by most packages. ASCII format simply consists of rows of text that you can view on a computer screen. Usually, each variable in the file is separated from the next by some **delimiter**, often a space or a comma. This is known as **free format**.

The simplest way of entering data in ASCII format is to type the data directly in this format using either a word processing or editing package. Alternatively, data stored in spreadsheet packages can be saved in ASCII format. Using either approach, it is customary for each row of data to correspond to a different individual in the study, and each column to correspond to a different variable, although it may be necessary to go on to subsequent rows if data from a large number of variables are collected on each individual.

Planning data entry

When collecting data in a study you will often need to use a form or questionnaire for recording the data. If these forms are designed carefully, they can reduce the amount of work that has to be done when entering the data. Generally, these forms/questionnaires include a series of boxes in which the data are recorded – it is usual to have a separate box for each possible digit of the response.

Categorical data

Some statistical packages have problems dealing with non-numerical data. Therefore, you may need to assign numerical codes to categorical data before entering the data into the computer. For example, you may choose to assign the codes of 1, 2, 3 and 4 to categories of 'no pain', 'mild pain', 'moderate pain' and 'severe pain', respectively. These codes can be added to the forms when collecting the data. For binary data, e.g. yes/no answers, it is often convenient to assign the codes 1 (e.g. for 'yes') and 0 (for 'no').

- **Single-coded** variables – there is only one possible answer to a question, e.g. 'is the patient dead?'. It is not possible to answer both 'yes' and 'no' to this question.
- **Multi-coded** variables – more than one answer is possible for each respondent. For example, 'what symptoms has this patient experienced?'. In this case, an individual may have experienced any of a number of symptoms. There are two ways to deal with this type of data depending upon which of the two following situations applies.
  - **There are only a few possible symptoms, and individuals may have experienced many of them.** A number of different binary variables can be created which correspond to whether the patient has answered yes or no to the presence of each possible symptom. For example, 'did the patient have a cough?', 'did the patient have a sore throat?'
  - **There are a very large number of possible symptoms but each patient is expected to suffer from only a few of them.** A number of different nominal variables can be created; each successive variable allows you to name a symptom suffered by the patient. For example, 'what was the first symptom the patient suffered?', 'what was the second symptom?'. You will need to decide in advance the maximum number of symptoms you think a patient is likely to have suffered.

Numerical data

Numerical data should be entered with the same precision as they are measured, and the unit of measurement should be consistent for all observations on a variable. For example, weight should be recorded in kilograms or in pounds, but not both interchangeably.

Multiple forms per patient

Sometimes, information is collected on the same patient on more than one occasion. It is important that there is some unique identifier (e.g. a serial number) relating to the individual that will enable you to link all of the data from an individual in the study.

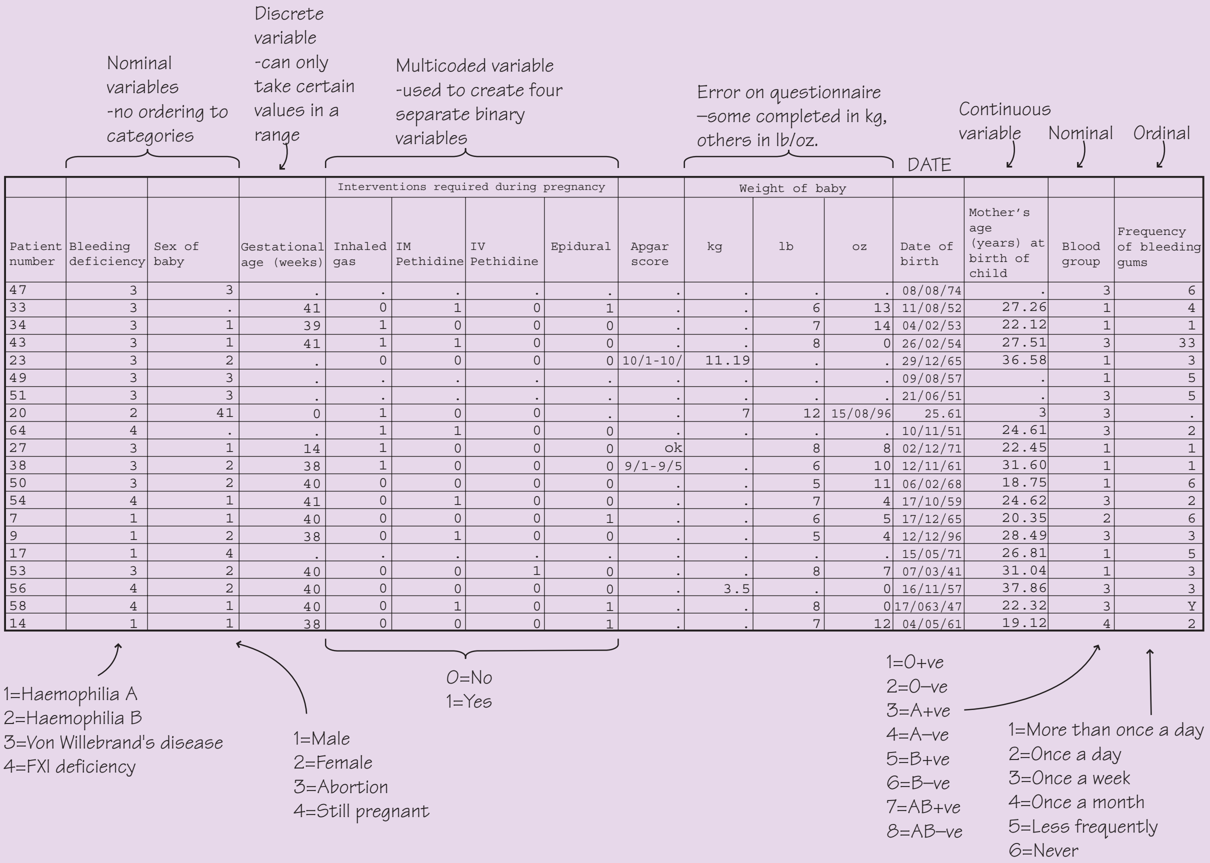
Problems with dates and times

Dates and times should be entered in a consistent manner, e.g. either as day/month/year or month/day/year, but not interchangeably. It is important to find out what format the statistical package can read.

Coding missing values

You should consider what you will do with missing values before you enter the data. In most cases you will need to use some symbol to represent a missing value. Statistical packages deal with missing values in different ways. Some use special characters (e.g. a full stop or asterisk) to indicate missing values, whereas others require you to define your own code for a missing value (commonly used values are 9, 999 or –99). The value that is chosen should be one that is not possible for that variable. For example, when entering a categorical variable with four categories (coded 1, 2, 3 and 4), you may choose the value 9 to represent missing values. However, if the variable is 'age of child' then a different code should be chosen. Missing data are discussed in more detail in Chapter 3.

Example


Nominal variables -no ordering to categories
Discrete variable -can only take certain values in a range
Multicoded variable -used to create four separate binary variables
Error on questionnaire –some completed in kg, others in lb/oz.
DATE
Continuous variable
Nominal
Ordinal
1=Haemophilia A
2=Haemophilia B
3=Von Willebrand's disease
4=FXI deficiency
1=Male
2=Female
3=Abortion
4=Still pregnant
0=No
1=Yes
1=O+ve
2=O–ve
3=A+ve
4=A–ve
5=B+ve
6=B–ve
7=AB+ve
8=AB–ve
1=More than once a day
2=Once a day
3=Once a week
4=Once a month
5=Less frequently
6=Never

				Interventions required during pregnancy					Weight of baby						
Patient number	Bleeding deficiency	Sex of baby	Gestational age (weeks)	Inhaled gas	IM Pethidine	IV Pethidine	Epidural	Apgar score	kg	lb	oz	Date of birth	Mother's age (years) at birth of child	Blood group	Frequency of bleeding gums
47	3	3	.	.	.	.	.	.	.	.	.	08/08/74	.	3	6
33	3	.	41	0	1	0	1	.	.	6	13	11/08/52	27.26	1	4
34	3	1	39	1	0	0	0	.	.	7	14	04/02/53	22.12	1	1
43	3	1	41	1	1	0	0	.	.	8	0	26/02/54	27.51	3	33
23	3	2	.	0	0	0	0	10/1-10/	11.19	.	.	29/12/65	36.58	1	3
49	3	3	.	.	.	.	.	.	.	.	.	09/08/57	.	1	5
51	3	3	.	.	.	.	.	.	.	.	.	21/06/51	.	3	5
20	2	41	0	1	0	0	.	.	7	12	15/08/96	25.61	3	3	.
64	4	.	.	1	1	0	0	.	.	.	.	10/11/51	24.61	3	2
27	3	1	14	1	0	0	0	ok		8	8	02/12/71	22.45	1	1
38	3	2	38	1	0	0	0	9/1-9/5	.	6	10	12/11/61	31.60	1	1
50	3	2	40	0	0	0	0	.	.	5	11	06/02/68	18.75	1	6
54	4	1	41	0	1	0	0	.	.	7	4	17/10/59	24.62	3	2
7	1	1	40	0	0	0	1	.	.	6	5	17/12/65	20.35	2	6
9	1	2	38	0	1	0	0	.	.	5	4	12/12/96	28.49	3	3
17	1	4	.	.	.	.	.	.	.	.	.	15/05/71	26.81	1	5
53	3	2	40	0	0	1	0	.	.	8	7	07/03/41	31.04	1	3
56	4	2	40	0	0	0	0	.	3.5	.	0	16/11/57	37.86	3	3
58	4	1	40	0	1	0	1	.	.	8	0	17/063/47	22.32	3	Y
14	1	1	38	0	0	0	1	.	.	7	12	04/05/61	19.12	4	2

Figure 2.1 Portion of a spreadsheet showing data collected on a sample of 64 women with inherited bleeding disorders.

As part of a study on the effect of inherited bleeding disorders on pregnancy and childbirth, data were collected on a sample of 64 women registered at a single haemophilia centre in London. The women were asked questions relating to their bleeding disorder and their first pregnancy (or their current pregnancy if they were pregnant for the first time on the date of interview). Fig. 2.1 shows the data from a small selection of the women after the data have been entered onto a spreadsheet, but before they have been checked for errors. The coding schemes for the categorical variables are shown at the bottom of Fig. 2.1. Each row of the spreadsheet represents a separate individual in the study; each column represents a different variable. Where the woman is still pregnant, the age of the woman at the time of birth has been calculated from the estimated date of the baby's delivery. Data relating to the live births are shown in Chapter 37.

Data kindly provided by Dr R.A. Kadir, University Department of Obstetrics and Gynaecology, and Professor C.A. Lee, Haemophilia Centre and Haemostasis Unit, Royal Free Hospital, London.