Medical Statistics at a Glance

WORKBOOK

Aviva Petrie & Caroline Sabin

Includes data analysis

WILEY-BLACKWELL

Medical Statistics at a Glance Workbook

This title is also available as an e-book.
For more details, please see
www.wiley.com/buy/9780470658482
or scan this QR code:

Companion website

Additional resources are available at:

www.medstatsaag.com

featuring:

- Excel datasets to accompany the data analysis sections
- Downloadable PDFs of two analysis templates
- Links to online further reading
- Supplementary MCQs

Medical Statistics at a Glance Workbook

Aviva Petrie

Head of Biostatistics Unit and Senior Lecturer
UCL Eastman Dental Institute
256 Grays 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
Royal Free Campus
Rowland Hill Street
London NW3 2PF

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

This edition first published 2013 © 2013 by Aviva Petrie and Caroline Sabin

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

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.

Library of Congress Cataloging-in-Publication Data

Petrie, Aviva.
Medical statistics at a glance workbook / Aviva Petrie, Caroline Sabin.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-65848-2 (pbk. : alk. paper) 1. Medical statistics. I. Sabin, Caroline. II. Title.
R853.S7P4762 2013
610.72'7–dc23

2012025027

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.

Cover design by Nathan Harris

Set in 9/11.5 pt Times by Toppan Best-set Premedia Limited

1 2013

Contents

Companion website

Additional resources are available at:

www.medstatsaag.com

featuring:
- Excel datasets to accompany the data analysis sections
- Downloadable PDFs of two analysis templates
- Links to online further reading
- Supplementary MCQs

Introduction

This workbook is a companion volume to the third edition of *Medical Statistics at a Glance*. Although primarily directed at undergraduate medical students preparing for statistics examinations, we believe that the workbook will also be of use to others working in the biomedical disciplines who simply want to brush up on their analytical and interpretation skills (e.g. other medical researchers, postgraduates in the biomedical disciplines and pharmaceutical industry personnel). Our aim for this workbook is therefore for it to act as a revision aid, equip readers with the skills necessary to read and interpret the published literature and give them the confidence to tackle their own statistical analyses. Although designed as an accompanying text to *Medical Statistics at a Glance*, it is not indelibly linked to it and can be used as a stand-alone text or in conjunction with any reputable text on statistics.

We believe that the optimal way to learn statistics is to put the theory into practice by undertaking an analysis of a data set, but recognise that this may not always be practical. Instead, the use of carefully constructed exercises in a variety of formats can help to test and fully evaluate the reader's understanding of the material (and identify any gaps that remain). As the *At a Glance* textbook presents information in a concise manner, there is limited space in it for worked examples and no room for exercises. Our workbook amends this insufficiency by providing an extensive set of questions, as well as templates for critical appraisal and descriptions of the statistical analyses of two data sets. Where possible, we have based questions on published studies in the medical and dental fields, and references are provided so that the reader may consult the original source material if interested.

The structure of the workbook

This workbook is divided into six parts:

Part 1

This section of the workbook contains multiple-choice questions (MCQs) that are generally brief, each testing the reader's knowledge of a single theoretical concept or aspect of study interpretation. Only one of the five possible answers provided is correct: an explanation is given in Part 5 for each correct and incorrect answer. The ordering of the MCQs generally follows that of the chapters in the third edition of *Medical Statistics at a Glance*. To aid readers who may wish to focus on specific topics in the *At a Glance* textbook, we provide a list of MCQs and the related chapters in Appendix I.

Part 2

This section of the workbook contains structured questions that are longer than the MCQs and provide a more in-depth exploration of the reader's knowledge of several statistical concepts. The questions may include elements that test a reader's understanding of the theory, as well as his or her ability to interpret study findings and, in some instances, to perform basic statistical calculations. The questions are similar to those that we have set in the past for exams: detailed model answers are provided in Part 5. As the longer structured questions may relate to information contained in several diverse chapters of the textbook, these do not follow any particular order, but in Appendix II, to aid readers who may wish to focus on specific topics in the *At a Glance* textbook, we provide a list of the structured questions and their related chapters.

Part 3

The ability to critically appraise the published literature is an essential skill that is required by anyone in the medical and dental professions (or, indeed, anyone involved in research more generally) and, consequently, is an important objective of a statistics course. Many aspects of statistics must be considered when evaluating the evidence provided in a research article, for example biases that might arise from inappropriate designs, sample size, outcome measures, the choice of statistical analysis, the presentation of the data and the conclusions drawn. Whilst *Medical Statistics at a Glance* presents a brief introduction to critical appraisal in Chapter 40 (Evidence-based medicine), Part 3 of our workbook supplements this by providing structured templates that can be used when reviewing and/or assessing the published literature. We suggest that the reader use these templates to critically appraise two published articles: a randomised controlled trial and an observational study. Our own evaluation of these articles is to be found in Part 5. Whilst we cannot hope to cover all possible topics within these two appraisals, we hope that they will at least provide a basic structure for appraisal that readers may find helpful.

Part 4

In our experience, one of the most common complaints from our students and junior research colleagues is that they just do not know where to start when analysing a substantial data set. To address this need, we have included in Part 4 a detailed description of the analyses of two data sets, the latter being available on the accompanying website (www.medstatsaag.com) as Excel files. Each analysis starts with a description of the clinical problem, and then takes the reader through the various steps that would be undertaken when performing the analysis, from the initial exploratory and descriptive analyses to the final sensitivity analyses that assess the robustness of study findings. We believe that this is an innovative approach and hope that readers will find it useful.

Part 5

This section of the workbook contains solutions to the MCQs in Part 1, model answers for the structured questions in Part 2, and our own critical appraisals of the randomised controlled trial (Paper 1) and the observational study (Paper 2) in Part 3. The pages in Part 5 are shaded so that the reader is easily able to navigate to the solutions and model answers.

Appendices

In Appendix I, we provide an ordered list of the MCQs and show which chapters they relate to, with an indication of the material included in each question. Appendix II is similar but identifies the associated chapters for the structured questions. For those readers who require exercises that relate to specific chapters of the *At a Glance* textbook, we provide a list of the chapters and indicate which multiple-choice and structured questions are relevant to them in Appendix III.

Further information

In addition to the workbook, we remind readers that the companion website to *Medical Statistics at a Glance* (www.medstatsaag.com) also contains an extensive set of interactive exercises, with references to many published papers that may be of interest.

Acknowledgements

Special thanks are due to Drs Laura Silveira-Moriyama and Angus Pringle who very kindly lent us their data sets for the analyses in Part 4 of the workbook. We are most appreciative of the extremely helpful comments and suggestions that they made during the development of the analyses, but we take full responsibility for any errors or misconceptions in the final presentations. We are also indebted to the authors and publishers of the two papers that we used for critical appraisal for allowing us to reproduce the articles, thereby providing useful exercises for our readers, and apologise if any of our criticisms cause offence. We acknowledge the generosity of the many authors and publishers who have kindly assented to our adapting or reproducing material for the multiple-choice and structured questions, and are grateful to the publishing team at Wiley-Blackwell both for suggesting that we write this workbook and for their ideas and support along its route to publication. Our acknowledgements would not be complete without thanking our students over the years from whom we have learnt the art of teaching, and Mike, Gerald, Nina, Andrew and Karen for their forbearance, encouragement and good humour during our absorption with this manuscript.

Part 1: Multiple-choice questions

Handling data

M1

To collect information on an individual's ability to function physically, investigators identified six daily tasks, each relating to a different aspect of physical functioning. For every task, respondents were asked to say whether they generally experienced 'no problems' (allocated a score of 0), 'some problems' (score of 1) or 'many problems' (score of 2) when performing the task; by summing the six individual scores, the investigators generated a total physical functioning score variable, which ranged from 0 to 12. Which one of the following statements is true?

a) The variable is best described as a continuous variable.

b) When capturing data on this score, only the final total score should be recorded on the data capture form.

c) Although this is strictly an ordinal categorical variable, for the purposes of analysis, it may be possible to treat this variable as a numerical variable.

d) The most suitable summary measure of the 'average' value for this variable would be the mode.

e) For the purposes of analysis, it would be preferable to re-categorise this final score into three categories: good functioning (scores of 0 to 4), average functioning (scores of 5 to 8) and poor functioning (scores of 9 to 12).

M2

Which one of the following statements is true?

a) A qualitative variable comprises two categories which may be ordinal or numerical.

b) An ordinal variable comprises categories which cannot be ordered.

c) The age groups 'young', 'middle aged' and 'old' relate to a nominal categorical variable.

d) Blood group is classified as a nominal categorical variable.

e) It may be difficult to distinguish a continuous numerical variable from an ordinal variable when the ordinal variable has many categories.

M3

As part of an epidemiological study investigating the association between consumption of dairy products in adolescence and the onset of cardiovascular disease later in life, study investigators plan to collect information on weekly egg consumption from a sample of children aged 14–17 years using self-administered questionnaires. Which one of the following would be the best approach for collecting this information?

a) Respondents are asked to indicate the number of eggs they consumed in the previous week and are asked to leave the entry blank if they do not know the answer.

b) Respondents are asked to tick the box that best describes the number of eggs they have consumed in the previous week: 0, 1–3, 4–7, >7 or 'unknown'.

c) Respondents are asked to indicate the number of eggs they consumed in the previous week, and to record a value of 9 if they do not know the answer.

d) Respondents are asked to tick the box that best describes the number of eggs they consumed in the previous week: 0, 1–3, 4–7 or >7; if they do not know the answer, they are asked to leave the response blank.

e) Respondents are asked to indicate the number of eggs they consumed in the previous week, and to record a value of 999 if they do not know the answer.

M4

Which one of the following statements which relate to the information provided in a questionnaire is true?

a) Having data available as an ASCII file is inflexible because many people have not heard of ASCII.

b) A multi-coded question has more than two possible responses, but the respondent can provide only one answer to it.

c) Dates must be entered into a computer spreadsheet as day/month/year.

d) Missing data for a particular respondent must always be entered on the computer spreadsheet as 9, 99 or 999.

e) It is often necessary to assign numerical codes to a categorical variable before entering the data into the computer.

M5

The number of eggs consumed by an adolescent in a week was collected from a sample of 40 adolescents aged 14–17 years with a view to estimating average weekly egg consumption in such adolescents. Information on egg consumption was missing for two adolescents; the data from the remaining 38 subjects are as follows: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 7, 7, 7, 8, 11, 14, 15, 21, 25, 27 and 71. Which one of the statements below is true?

a) The entry of 71 is an outlier but is likely to be a correct value of weekly egg consumption.

b) As it is unlikely that any individual will consume more than three eggs per day, the investigators should exclude any values greater than 21 before analysing the data.

c) As it is unlikely that any individual will consume more than three eggs per day, the investigators should replace any values >21 with the value 21 before analysing the data.

d) The authors suspect that the value of 71 is a typing error in the data set and so plan to replace this value with the value 9 before conducting any analyses.

e) The authors believe that the value of 71 is an error in the data set and so could consider running their analyses both including and excluding this outlying value.

M6

Which one of the following statements is true?

a) One approach to handling outliers in a data set is to analyse the data both with and without the outliers and see whether the results are similar.

b) It is never sensible to transform data to overcome the problem of a skewed distribution as the parameter estimates obtained from the transformed data cannot be interpreted.
c) Outliers in data should be omitted from the analysis as they may skew the results.
d) An outlier is an extreme value which is incompatible with the main body of the data and is always greater than all the other values in the data set.
e) The only ways of dealing with outliers in a data set are to analyse the data both with and without the outliers and determine the effect of the omission, to omit the outlier(s) from the analysis or to transform the data.

M7

Consider the data relating to the number of eggs consumed in a week described in Questions M3 and M5. Which one of the following diagrams would be best for displaying the information?
a) A bar chart
b) A histogram
c) A pie chart
d) A scatter diagram
e) A segmented bar chart

M8

Consider the data on the number of eggs consumed in a week described in Questions M3 and M5. Which one of the following best describes the distribution of this variable?
a) Skewed to the right
b) Normally distributed
c) Skewed to the left
d) Uniformly distributed
e) Negatively skewed

M9

Which one of the following statements is true?
a) A pie chart is one in which a circular 'pie' is split into sectors, one for each category of a categorical variable, so that the area of each sector is equal.
b) A sensible way of displaying continuous numerical data is to draw a bar chart.
c) A histogram is a chart in which separate vertical (or horizontal) bars are drawn with gaps between the bars; the width (height) of each bar relates to a specific range of values of the variable, and its height (width) is proportional to the associated frequency of observations.
d) The distribution of a variable is right skewed if a histogram of observed values has a long tail to the right with one or a few high values.
e) A box-and-whisker plot comprises a vertical or horizontal rectangle indicating the interquartile range, within which is the median; the ends of the 'whiskers' represent the upper and lower limits of the 95% confidence interval for the median.

M10

The authors of the egg consumption study (Questions M3 and M5) now wish to summarise the data on the number of eggs consumed in a week. Which one of the following approaches would be the best way to summarise these data?
a) The arithmetic mean and range
b) The median and interquartile range
c) The median and range
d) The arithmetic mean and standard deviation
e) The mode

M11

Which one of the following statements is true?
a) The median is greater than the arithmetic mean if the data are skewed to the right.
b) The median value of n observations is equal to the $(n + 1)/2$th value in the ordered set if n is odd.
c) The median and the weighted mean are always identical if the weights used in the calculation of the weighted mean are equal.
d) The logarithmic transformation of left-skewed data will often produce a symmetrical distribution when the transformed data are plotted on an arithmetic scale.
e) The geometric mean of a data set is equal to the arithmetic mean of the log-transformed data.

M12

Study investigators collected information on haemoglobin levels in a sample of 212 healthy women of mixed ethnicity. The investigators calculated the median value, and used the 2.5th and 97.5th percentile values to generate a reference range. Which one of the following statements is true?
a) The authors generated the reference range using the percentile approach as the number of subjects in their study was small.
b) Healthy individuals in the population will not have a value of haemoglobin that falls below the lower limit of the reference range.
c) Use of the mean and standard deviation to generate the reference range would have provided a more suitable reference range.
d) Individuals in the population with an underlying health condition that has an impact on haemoglobin levels will always have values that fall outside the reference range.
e) An individual in the population with an underlying health condition that has an impact on haemoglobin levels is likely to have a value that falls outside the reference range.

M13

When numerical data are arranged in order of magnitude, which one of the following statements is true?
a) The interquartile range is the difference between the first and fourth percentiles.
b) The interdecile range contains the central 80% of the ordered observations.
c) The middle observation is always equal to the arithmetic mean.
d) The 50th percentile is equal to the fifth quartile.
e) The first percentile is always equal to the minimum value.

M14

If a set of observations follow the Normal or Gaussian distribution, which one of the following statements is true?
a) Its mean and variance are equal.
b) Its observations are derived from healthy individuals.
c) Its mean and variance are always equal to zero and one, respectively.

d) 95% of the observations lie between the mean ± 1.96 times the variance.
e) Approximately 68% of the observations lie between the mean ± the standard deviation.

M15

Which one of the following statements is true?
a) A Binomial random variable is the count of the number of events that occur randomly and independently in time or space at some fixed average rate.
b) The two parameters that characterise a Poisson distribution are the number of individuals in the sample (or repetitions of a trial) and the true probability of success for each individual (or in each trial).
c) The Chi-squared distribution is based on a categorical random variable.
d) When the logarithm of observations which follow the Lognormal distribution are taken, the transformed observations follow the Normal distribution.
e) The Lognormal distribution is highly skewed to the left.

M16

The distribution of age at menopause tends to be skewed to the left. Study investigators wish to identify demographic and socioeconomic factors that are independently associated with age at menopause. Which one of the following statements relating to the analysis of age at menopause is true?
a) The optimal analytical approach is always to use a nonparametric method due to the skewness of the distribution.
b) Use of the logarithmic transformation would permit a parametric analysis based on the Normal distribution.
c) Use of the square transformation may help to achieve Normality.
d) By using a square transformation, we can ensure that the assumptions underlying a parametric analysis based on the Normal distribution are met.
e) The study investigators would be best advised to categorise age at menopause before performing the analysis.

M17

Which one of the following statements is true?
a) The logistic transformation linearises a sigmoid curve.
b) The logistic transformation is generally applied to counts which follow the Poisson distribution.
c) If a numerical variable, y, is skewed to the right, the distribution of $z = y^2$ is often approximately Normal.
d) If a numerical variable, y, is skewed to the left, $z = \log y$ is often approximately Normally distributed.
e) The square transformation has properties which are similar to those of the logarithmic transformation.

Sampling and estimation

M18

Which one of the following statements is true? The sampling distribution of the mean:
a) represents the mean of the distribution obtained by taking many repeated samples of a fixed size from the population of interest and plotting the observations so obtained;
b) has a mean which is an unbiased estimate of the true mean in the population;
c) will follow a Normal distribution only if the distribution of the original data is Normal;
d) has a standard deviation which is larger than the standard error of the mean; or
e) cannot be drawn if the sample size of the repeated samples is small.

M19

Study investigators have collected data on the heights of a sample of 137 women in Thailand. Which one of the following statements is true?
a) The true mean height in the Thai female population will be equal to the mean height of the women in the sample.
b) If the investigators were to calculate the range of values determined by the mean height ± 1.96 × standard deviation, they would be able to assess from this range of values the precision of the estimated mean height in their sample.
c) To enable other research groups to compare the distribution of the height values in their own studies to those from the investigators' study, the investigators should calculate and present the median height and its associated confidence interval.
d) If the heights are approximately Normally distributed, the authors may calculate and present the mean height and its standard deviation. This will allow them to describe the distribution of height values in their sample.
e) By calculating the confidence interval for the mean, the investigators will be able to determine whether the height values in their sample are Normally distributed.

M20

Jensen *et al.* (2011) conducted a retrospective cohort study to assess the incidence of wound complications among patients undergoing lower-limb arthroplasty, before and after a change in clinical practice from the use of low-molecular-weight heparin to rivaroxaban. Prior to the switch to rivaroxaban, 9 of 489 patients (1.8%, 95% confidence interval 0.9 to 3.5%) returned to theatre with wound complications within 30 days compared to 22 of the 559 patients (3.9%, 95% confidence interval 2.6 to 5.9%) who received rivaroxaban. Which one of the following statements is true?
a) The confidence interval for the wound complication rate prior to the switch to rivaroxaban is asymmetrical, indicating that the outcome is not Normally distributed.
b) The true percentage of wound complications prior to the switch to rivaroxaban lies between 0.9% and 3.5%.
c) The 95% confidence intervals for the two periods overlap, indicating that there was no significant change in the wound complication rate after the switch to rivaroxaban.
d) Had the number of wound complications been greater in each period, the confidence intervals would have been wider.
e) Had the number of patients in each period been greater, the confidence intervals would have been narrower.

Jensen CD, Steval A, Partington PF, Reed MR, Muller SD. Return to theatre following total hip and knee replacement, before and after the introduction of rivaroxaban: a retrospective cohort study. J Bone Joint Surg Br 2011; 93: 91–5.

M21

Which of the following statements is true for a sample of size $n > 1$?

a) The 99% confidence interval for the mean is narrower than the 95% confidence interval for the mean.

b) The 95% confidence interval for the mean of a particular variable is narrower than the reference interval for that variable.

c) If the true standard deviation is known, the 95% confidence interval for the mean is calculated as the mean ± 1.96 times the standard deviation.

d) The 95% confidence interval for the mean represents the interval within which the sample mean falls with 95% certainty.

e) The 95% confidence interval for the mean represents the interval which contains the central 95% of the observations in the population.

Study design

M22

Which one of the following studies would be best described as a cohort study?

a) A study in which cells are stimulated with three different types of growth inducing protein.

b) A study of medical students who are followed from entering medical school to the end of their first year to describe the associations between lifestyle factors and end-of-first-year exam results.

c) A study of medical students who are split by the study investigators into two groups: those with surnames beginning with the letters A to M received regular counselling support over the first year, and those with surnames beginning with the letters O to Z did not. The outcome of the study was the proportion of students who passed their end-of-first-year exams, and this proportion was to be compared between the two groups.

d) Medical students who fail their end-of-first-year exams are interviewed about their lifestyles over the first year; a random sample of students who passed their end-of-first-year exams are also interviewed, and the results compared to assess the effects of lifestyle factors on failing the end-of-first-year exams.

e) Study investigators compared end-of-first-year exam pass rates at 10 different medical schools in the United Kingdom, and correlated these pass rates with the number of bars and nightclubs in the vicinity of each medical school.

M23

In medical research we are often interested in determining whether exposure to a factor causes an effect (e.g. a disease). Which one of the following criteria is a necessary component for assessing the cause of disease?

a) The cause and effect must take place simultaneously.

b) The association between cause and effect can be assessed on the basis of statistical results alone, independently of biological reasoning.

c) If feasible, removing the potential causative factor of interest should reduce the risk of disease.

d) The effect cannot be causal if the association between the cause and effect is small.

e) It is usually sufficient to imply causation on the basis of the results from a single study, provided the association between the cause and effect in that study is strong.

M24

Investigators conducted a randomised cross-over study to compare two appliances for the prevention of snoring. Every trial participant used each appliance for a period of one month, with a 2-week washout period between the two study periods. Which one of the following statements is true?

a) The investigators chose a cross-over design as the appliances are likely to have a long-term impact on snoring symptoms.

b) By using a cross-over design, the investigators were able to shorten the length of treatment time that was required.

c) Because they had used a cross-over design rather than a parallel group study design, the investigators had to increase the size of their sample.

d) A 2-week washout period was incorporated to allow the trial participants sufficient time to clean and return their appliances.

e) By choosing a cross-over design, the investigators were able to use each participant as his or her own control, thus reducing variability.

M25

In randomised trials of new human immunodeficiency virus therapies, investigators may use a composite endpoint known as the time to loss of virological response. Patients are deemed to meet the endpoint after the first of a series of events occurs: a new acquired immunodeficiency syndrome event, death, the patient is lost to follow-up or the patient experiences virological failure on treatment. At that point, the patient exits the trial and follow-up ceases on the patient. Which one of the following statements is true?

a) Investigators use a composite endpoint as they cannot make a decision in advance about which is the most important outcome.

b) Composite endpoints simplify the analysis of randomised trials.

c) If one or more components of the composite endpoint are deemed to have greater clinical relevance than others, then appropriate analytical methods which take this into consideration must be used when analysing a trial that utilises a composite endpoint.

d) A study that uses such a composite endpoint can provide reliable information about the frequency of occurrence of each component of the composite; thus, this type of trial provides good value for money.

e) If a composite endpoint is used instead of basing the analysis on each component of the composite, the length of the trial must be increased.

M26

Which one of the following statements is true?

a) A factorial design is one in which there is a single factor of interest.

b) A statistical interaction exists in a clinical trial when one or more of the treatments produce side effects.