

# How to Display Data

*This page intentionally left blank*

# How to Display Data

**Jenny V. Freeman**

School of Health and Related Research  
University of Sheffield  
Sheffield, UK

**Stephen J. Walters**

School of Health and Related Research  
University of Sheffield  
Sheffield, UK

**Michael J. Campbell**

School of Health and Related Research  
University of Sheffield  
Sheffield, UK

BMJ|Books



**Blackwell**  
Publishing

© 2008 Jenny V. Freeman, Stephen J. Walters, Michael J. Campbell

Published by Blackwell Publishing

BMJ Books is an imprint of the BMJ Publishing Group Limited, used under licence

Blackwell Publishing, Inc., 350 Main Street, Malden, Massachusetts 02148-5020, USA

Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK

Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia

The right of the Author to be identified as the Author of this Work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

First published 2008

1 2008

Library of Congress Cataloging-in-Publication Data

Freeman, Jenny.

How to display data / Jenny Freeman, Stephen J. Walters, Michael J. Campbell.

p. ; cm.

ISBN 978-1-4051-3974-8 (pbk. : alk. paper)

1. Medical writing. 2. Medical statistics. 3. Medicine—Research—Statistical methods. I. Walters, Stephen John. II. Campbell, Michael J., PhD. III. Title. [DNLM: 1. Research Design. 2. Data Display. 3. Data Interpretation, Statistical. 4. Statistics. W 20.5 F869h 2007]

R119.F76 2007

610.72'7—dc22

2007032641

ISBN: 978-1-4051-3974-8

A catalogue record for this title is available from the British Library

Set by Charon Tec Ltd (A Macmillan Company), Chennai, India

Printed and bound in Singapore by Utopia Press Pte Ltd

Commissioning Editor: Mary Banks

Editorial Assistant: Victoria Pittman

Development Editor: Simone Dudziak

Production Controller: Rachel Edwards

For further information on Blackwell Publishing, visit our website:

<http://www.blackwellpublishing.com>

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

Blackwell Publishing makes no representation, express or implied, that the drug dosages in this book are correct. Readers must therefore always check that any product mentioned in this publication is used in accordance with the prescribing information prepared by the manufacturers. The author and the publishers do not accept responsibility or legal liability for any errors in the text or for the misuse or misapplication of material in this book.

# Contents

Preface, vii

**1** Introduction to data display, 1

**2** How to display data badly, 9

**3** Displaying univariate categorical data, 17

**4** Displaying quantitative data, 29

**5** Displaying the relationship between two  
continuous variables, 43

**6** Data in tables, 59

**7** Reporting study results, 66

**8** Time series plots and survival curves, 90

**9** Displaying results in presentations, 98

Index, 107

*This page intentionally left blank*

# Preface

The best method to convey a message from a piece of research in health is via a figure. The best advice that a statistician can give a researcher is to first plot the data. Despite this, conventional statistics textbooks give only brief details on how to draw figures and display data. The purpose of this book is to give advice on the best methods to display data which have arisen from a variety of different sources. We have tried to make the book concise and easy to read. By displaying data badly one can very easily give misleading messages (or hide inconvenient truths) and we try to highlight how consumers of data have to be aware of these problems. We have also included advice on displaying data for posters and talks.

Researchers who want to display the results of their studies in figures or tables particularly for publication in a journal will find this book useful. Readers of the research literature, who wish to critically appraise a piece of work will find useful tips on interpreting figures that they encounter. People who have to deliver a talk or a conference presentation should also find good advice on displaying their results.

We would like to thank Mary Banks and Simone Dudziak from Blackwell for their patience and advice.

Jenny V. Freeman  
Stephen J. Walters  
Michael J. Campbell  
Medical Statistics Group, ScHARR, Sheffield  
June 2007

*This page intentionally left blank*



# Chapter 1 **Introduction to data display**

## **1.1 Introduction**

This book has arisen from our extensive experience as researchers and teachers of medical statistics. We have frequently been appalled by the poor quality of data display even in major medical journals. While there is already a wealth of information about how to display data, it is scattered across many sources. Our purpose in writing this book is to bring together this information into a single volume and provide clear accessible advice for both researchers, and students alike.

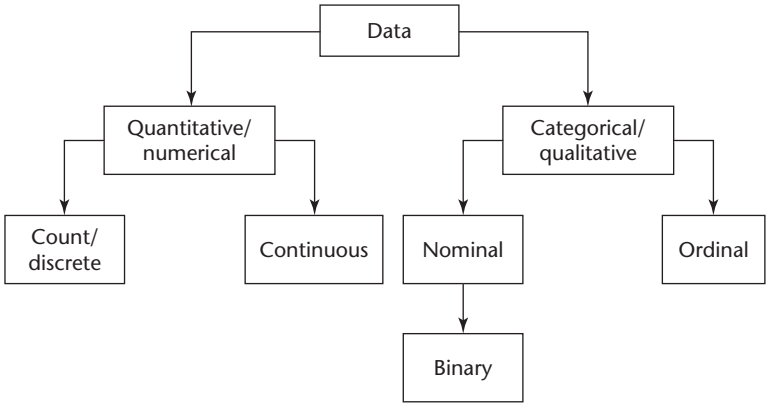
Well-displayed data can clearly illuminate and enhance the interpretation of a study, while badly laid out data and results can obscure the message or at worst seriously mislead. Although the appropriate display of data in tables and graphs is an essential part of any report, paper or presentation, little space is devoted to it in the majority of textbooks. The purpose of this book is to address this deficit and give clear guidelines on appropriate methods for displaying quantitative information, using both graphs and tables.

There are many different types of graph and table available for displaying data; their purposes will be outlined in subsequent chapters. This chapter will outline the reasons why it is important to get display right, good principles to adhere to when displaying data and the types of data that will be covered in the rest of the book. The second chapter will cover some of the many ways in which the display of information can be badly done and the following chapters will then unpick these, and give clear guidance on how to do it well.

## **1.2 Types of data**

To display data appropriately, one must first understand what types of data there are, as this determines the best method of displaying them. Figure 1.1 shows a basic hierarchy of data types, although there are others. Data are either *categorical* or *quantitative*. Data are described as categorical when they can

## 2 How to Display Data



**Figure 1.1** Types of data.

be categorised into distinct groups, such as ethnic group or disease severity. Although categorical data may be coded numerically, for example gender may be coded 1 for male and 2 for female, these codes have no intrinsic numerical value; it would be nonsense to calculate an average gender. Categorical data can be divided into either *nominal* or *ordinal*. Nominal data have no natural ordering and examples include eye colour, marital status and area of residence. *Binary* data is a special subcategory of nominal data, where there are only two possible values, for example male/female, yes/no, dead/alive. Ordinal data occurs when there can be said to be a natural ordering of the data values, such as better/same/worse, grades of breast cancer and social class.

Quantitative data can be either counted or continuous. *Count* data are also known as discrete data and, as the name implies, occur when the data can be counted, such as the number of children in a family or the number of visits to a GP in a year. Count data are similar to categorical data as they can only take discrete whole numbers. *Continuous* data are data that can be measured and they can take any value on the scale on which they are measured; they are limited only by the scale of measurement and examples include height, weight and blood pressure.

### 1.3 Where to start?

When displaying information visually, there are three questions one will find useful to ask as a starting point (Box 1.1). Firstly and most importantly, it is vital to have a clear idea about what is to be displayed; for example, is it important to demonstrate that two sets of data have different distributions or

**Box 1.1** Useful questions to ask when considering how to display information

- What do you want to show?
- What methods are available for this?
- Is the method chosen the best? Would another have been better?

that they have different mean values? Having decided what the main message is, the next step is to examine the methods available and to select an appropriate one. Finally, once the chart or table has been constructed, it is worth reflecting upon whether what has been produced truly reflects the intended message. If not, then refine the display until satisfied; for example if a chart has been used would a table have been better or vice versa? This book will help you answer these questions and provide you with the means to best display your data.

## 1.4 Recommendations for the presentation of numbers

When summarising categorical data, both frequencies and percentages can be used. However, if percentages are reported, it is important that the denominator (i.e. total number of observations) is given. To summarise continuous numerical data, one should use the mean and standard deviation, or if the data have a skewed distribution use the median and range or interquartile range. However, for all of these calculated quantities it is important to state the total number of observations on which they are based.

In the majority of cases it is reasonable to treat count data, such as number of children in a family or number of visits to the GP in a year, as if they were continuous, at least as far as the statistical analysis goes. Ideally there should be a large number of different possible values, but in practice this is not always necessary. However, where ordered categories are numbered, such as stage of disease or social class, the temptation to treat these numbers as statistically meaningful must be resisted. For example, it is not sensible to calculate the average social class of a sample or stage of cancer for a group of patients, and in such cases the data should be treated in statistical analyses as if they are ordered categories.<sup>1</sup>

Numerical precision should be consistent throughout and summary statistics such as means and standard deviations should not have more than one extra decimal place (or significant digit) compared to the raw data. Spurious precision should be avoided although when certain measures are to be used for further calculations or when presenting the results of analyses, greater precision may sometimes be appropriate.<sup>2</sup>

## 4 How to Display Data

### 1.5 Recommendations for presenting data and results in tables

There are a few basic rules of good presentation, both within the text of a document or presentation, and within tables, as outlined in Box 1.2. Tufte, in 1983, outlined a fundamental principle: always try to get as much information into a figure consistent with legibility. In other words, one should maximise the ratio of the amount of information given to the amount of ink used.<sup>3</sup> Tables, including column and row headings, should be clearly labelled and a brief summary of the contents of a table should always be given in words, either as part of the title or in the main body of the text.

#### **Box 1.2** Recommendations when presenting data and results in tables

- The amount of information should be maximised for the minimum amount of ink.
- Numerical precision should be consistent throughout a paper or presentation, as far as possible.
- Avoid spurious accuracy. Numbers should be rounded to two effective digits.
- Quantitative data should be summarised using either the mean and standard deviation (for symmetrically distributed data) or the median and interquartile range or range (for skewed data). The number of observations on which these summary measures are based should be included.
- Categorical data should be summarised as frequencies and percentages. As with quantitative data, the number of observations should be included.
- Each table should have a title explaining what is being displayed and columns and rows should be clearly labelled.
- Solid lines in tables should be kept to a minimum.
- Where variables have no natural ordering, rows and columns should be ordered by size.

Solid lines should not be used in a table except to separate labels and summary measures from the main body of the data. However, their use should be kept to a minimum, particularly vertical gridlines, as they can interrupt eye movements, and thus the flow of information. White space can be used to separate data, such as different variables, from each other.<sup>4</sup>

The information in tables is easier to comprehend if the columns (rather than the rows) contain similar information, such as means or standard deviations, as it is easier to scan down a column than across a row.<sup>4</sup> However, it