STATISTICS AT SQUARE TWO

MICHAEL J. CAMPBELL
RICHARD M. JACQUES

THIRD EDITION

WILEY Blackwell

Statistics at Square Two

Statistics at Square Two

Understanding Modern Statistical Application in Medicine

Michael J. Campbell
Emeritus Professor of Medical Statistics
University of Sheffield
Sheffield, UK

Richard M. Jacques
Senior Lecture in Medical Statistics
University of Sheffield
Sheffield, UK

Third Edition

WILEY Blackwell

This edition first published 2023

Edition History
©1e, 2001 by BMJ Books
©2e, 2006 by M. J. Campbell

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data
Names: Campbell, Michael J., 1950- author. | Jacques, Richard M., author.
Title: Statistics at square two : understanding modern statistical applications in medicine / Michael J. Campbell, Emeritus Professor of Medical Statistics, University of Sheffield, Sheffield, UK, Richard M. Jacques, Senior Lecturer in Medical Statistics, University of Sheffield, Sheffield, UK.
Description: Third edition. | Hoboken : John Wiley & Sons, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022057672 (print) | LCCN 2022057673 (ebook) | ISBN 9781119401360 (paperback) | ISBN 9781119401377 (pdf) | ISBN 9781119401391 (epub) | ISBN 9781119401407 (ebook)
Subjects: LCSH: Medical statistics--Data processing. | Medical statistics--Computer programs.
Classification: LCC RA409.5 .C36 2023 (print) | LCC RA409.5 (ebook) | DDC 610.2/1--dc23/eng/20221214
LC record available at https://lccn.loc.gov/2022057672
LC ebook record available at https://lccn.loc.gov/2022057673

Cover Image: © Somatuscani/iStock/Getty Images
Cover Design: Wiley

Set in 9.5/12.5pt STIXTwoText by Integra Software Services Pvt. Ltd, Pondicherry, India

Contents

Preface

In the 16 years since the second edition of *Statistics at Square Two* was published, there have been many developments in statistical methodology and in methods of presenting statistics. MJC is pleased that his colleague Richard Jacques, who has considerable experience in more advanced statistical methods and teaching medical statistics to non- statisticians, has joined him as a co-author. Most of the examples have been updated and two new chapters have been added on meta-analysis and on time series analysis. In addition, reference is made to the many checklists which have appeared since the last edition to enable better reporting of research.

This book is intended to build on the latest edition of *Statistics at Square One*.[1] It is hoped to be a *vade mecum* for investigators who have undergone a basic statistics course, but need more advanced methods. It is also intended for readers and users of the medical literature, but is intended to be rather more than a simple "bluffer's guide". It is hoped that it will encourage the user to seek professional help when necessary. Important sections in each chapter are tips on reading and reporting about a particular technique; the book emphasises correct interpretation of results in the literature. Much advanced statistical methodology is used rather uncritically in medical research, and the data and code to check whether the methods are valid are often not provided when the investigators write up their results. This text will help readers of statistics in medical research engage in constructive critical review of the literature.

Since most researchers do not want to become statisticians, detailed explanations of the methodology will be avoided. However, equations of the models are given, since they show concisely what each model is assuming. We hope the book will prove useful to students on postgraduate courses and for this reason there are a number of exercises with answers. For students on a more elementary course for health professionals we recommend Walters *et al.*[2]

The choice of topics reflects what we feel are commonly encountered in the medical literature, based on many years of statistical refereeing. The linking theme is regression models and we cover multiple regression, logistic regression, Cox regression, random effects (mixed models), ordinal regression, Poisson regression, time series regression and meta-analysis. The predominant philosophy is frequentist, since this reflects the literature and what is available in most packages. However, a discussion on the uses of Bayesian methods is given in an Appendix 4. The huge amount of work on causal modelling is briefly referenced, but is generally beyond the scope of this book.

Most of the concepts in statistical inference have been covered in *Statistics at Square One*.[1] In order to keep this book short, reference will be made to the earlier book for basic concepts. All the analyses described in the book have been conducted in the free software R and the code is given to make the methods accessible to reserachers without commercial statistical packages.

We are grateful to Tommy Nyberg of the Biostatistics Unit, Cambridge for feedback on his survival paper and to our colleague, Jeremy Dawson, who read and commented on the final draft. Any remaining errors are our own.

Michael J. Campbell
Richard M. Jacques
Sheffield, June 2022

References

1 Campbell MJ. *Statistics at Square One*, 12th edn. Hoboken, NJ: Wiley-Blackwell, 2021.
2 Walters SJ, Campbell MJ, Machin D. *Medical Statistics: A Textbook for the Health Sciences*, 5th edn. Chichester: John Wiley & Sons Ltd, 2020.

1

Models, Tests and Data

Summary

This chapter covers some of the basic concepts in statistical analysis, which are covered in greater depth in *Statistics at Square One*. It introduces the idea of a statistical model and then links it to statistical tests. The use of statistical models greatly expands the utility of statistical analysis. In particular, they allow the analyst to examine how a variety of variables may affect the result.

1.1 Types of Data

Data can be divided into two main types: quantitative and qualitative. *Quantitative data* tend to be either continuous variables that one can measure (such as height, weight or blood pressure) or discrete (such as numbers of children per family or numbers of attacks of asthma per child per month). Thus, count data are discrete and quantitative. Continuous variables are often described as having a Normal distribution, or being non-Normal. Having a Normal distribution means that if you plot a histogram of the data it would follow a particular "bell-shaped" curve. In practice, provided the data cluster about a single central point, and the distribution is symmetric about this point, it would be commonly considered close enough to Normal for most tests requiring Normality to be valid. Here one would expect the mean and median to be close. Non-Normal distributions tend to have asymmetric distributions (skewed) and the means and medians differ. Examples of non-Normally distributed variables include ages and salaries in a population. Sometimes the asymmetry is caused by outlying points that are in fact errors in the data and these need to be examined with care.

Note that it is a misnomer to talk of "non-parametric" data instead of non-Normally distributed data. Parameters belong to models, and what is meant by "non-parametric" data is data to which we cannot apply models, although as we shall see later, this is often a too limited view of statistical methods. An important feature of quantitative data is that you can deal with the numbers as having real meaning, so for example you can take averages of the data. This is in contrast to qualitative data, where the numbers are often convenient labels and have no quantitative value.

Qualitative data tend to be categories, thus people are male or female, European, American or Japanese, they have a disease or are in good health and can be described as

Statistics at Square Two: Understanding Modern Statistical Application in Medicine, Third Edition.
Michael J. Campbell and Richard M. Jacques.

nominal or *categorical*. If there are only two categories they are described as *binary* data. Sometimes the categories can be ordered, so for example a person can "get better", "stay the same" or "get worse". These are *ordinal* data. Often these will be scored, say, 1, 2, 3, but if you had two patients, one of whom got better and one of whom got worse, it makes no sense to say that on average they stayed the same (a statistician is someone with their head in the oven and their feet in the fridge, but on average they are comfortable!). The important feature about ordinal data is that they can be ordered, but there is no obvious weighting system. For example, it is unclear how to weight "healthy", "ill" or "dead" as outcomes. (Often, as we shall see later, either scoring by giving consecutive whole numbers to the ordered categories and treating the ordinal variable as a quantitative variable or dichotomising the variable and treating it as binary may work well.) Count data, such as numbers of children per family appear ordinal, but here the important feature is that arithmetic is possible (2.4 children per family is meaningful). This is sometimes described as having *ratio* properties. A family with four children has twice as many children as a family with two, but if we had an ordinal variable with four categories, say "strongly agree", "agree", "disagree" and "strongly disagree", and scored them 1–4, we cannot say that "strongly disagree", scored 4, is twice "agree", scored 2.

Qualitative data can also be formed by categorising continuous data. Thus, blood pressure is a continuous variable, but it can be split into "normotension" or "hypertension". This often makes it easier to summarise, for example 10% of the population have hypertension is easier to comprehend than a statement giving the mean and standard deviation of blood pressure in the population, although from the latter one could deduce the former (and more besides). Note that qualitative data is not necessarily associated with qualitative research. Qualitative research is of rising importance and complements quantitative research. The name derives because it does not quantify measures, but rather identifies themes, often using interviews and focus groups.

It is a parody to suggest that statisticians prefer not to dichotomise data and researchers always do it, but there is a grain of truth in it. Decisions are often binary: treat or not treat. It helps to have a "cut-off", for example treat with anti-hypertensive if diastolic blood pressure is >90 mmHg, although more experienced clinicians would take into account other factors related to the patient's condition and use the cut-off as a point when their likelihood of treating increases. However, statisticians point out the loss of information when data are dichotomised, and are also suspicious of arbitrary cut-offs, which may have been chosen to present a conclusion desired by a researcher. Although there may be good reasons for a cut-off, they are often opaque, for example deaths from Covid are defined as deaths occurring within 30 days of a positive Covid test. Why 30 days, and not 4 weeks (which would be easier to implement) or 3 months? Clearly ten years is too long. In this case it probably matters little which period of time is chosen but it shows how cut-offs are often required and the justification may be lost.

1.2 Confounding, Mediation and Effect Modification

Much medical research can be simplified as an investigation of an input–output relationship. The inputs, or explanatory variables, are thought to be related to the outcome or effect. We wish to investigate whether one or more of the input variables are plausibly

causally related to the effect. The relationship is complicated by other factors that are thought to be related to both the cause and the effect; these are confounding factors. A simple example would be the relationship between stress and high blood pressure. Does stress cause high blood pressure? Here the causal variable is a measure of stress, which we assume can be quantified either as a binary or continuous variable, and the outcome is a blood pressure measurement. A confounding factor might be gender; men may be more prone to stress, but they may also be more prone to high blood pressure. If gender is a confounding factor, a study would need to take gender into account. A more precise definition of a confounder states that a confounder should "not be on the causal pathway". For example stress may cause people to drink more alcohol, and it is the increased alcohol consumption which causes high blood pressure. In this case alcohol consumption is not a confounder, and is often termed a *mediator*.

Another type of variable is an effect modifier. Again, it is easier to explain using an example. It is possible that older people are more likely than younger people to suffer high blood pressure when stressed. Age is not a confounder if older people are not more likely to be stressed than younger people. However, if we had two populations with different age distributions our estimate of the effect of stress on blood pressure would be different in the two populations if we didn't allow for age. Crudely, we wish to remove the effects of confounders, but study effect modifiers.

An important start in the analysis of data is to determine which variables are outputs and which variables are inputs, and of the latter which do we wish to investigate as causal, and which are confounders or effect modifiers. Of course, depending on the question, a variable might serve as any of these. In a survey of the effects of smoking on chronic bronchitis, smoking is a causal variable. In a clinical trial to examine the effects of cognitive behavioural therapy on smoking habit, smoking is an outcome. In the above study of stress and high blood pressure, smoking may also be a confounder.

A common error is to decide which of the variables are confounders by doing significance tests. One might see in a paper: "only variables that were significantly related to the output were included in the model." One issue with this is it makes it more difficult to repeat the research; a different researcher may get a different set of confounders. In later chapters we will discuss how this could go under the name of "stepwise" regression. We emphasise that significance tests are not a good method of choosing the variable to go in a model.

In summary, before any analysis is done, and preferably in the original protocol, the investigator should decide on the causal, outcome and confounder variables. An exploration of how variables relate in a model is given in Section 1.10.

1.3 Causal Inference

Causal inference is a new area of statistics that examines the relationship between a putative cause and an outcome. A useful and simple method of displaying a causal model is with a Direct Acyclic Graph (DAG).[1] They can be used to explain the definitions given in the previous section. There are two key features to DAGs: (1) they show direct relationships using lines and arrows and are usually read from left to right and (2) they don't allow feedback, that is, you can't get back to where you started following the arrows.

We start with a cause, which might be an exposure (E) or a treatment (T). This is related to an outcome (O) or disease (D) but often just denoted Y. Confounders (C) are variables related to both E and Y which may change the relationship between E and Y. In randomised trials, we can in theory remove the relationship between C and T by randomisation, so making causal inference is easier. For observational studies, we remove the link between C and T using models, but models are not reality and we may have omitted to measure key variables, so confounding and bias may still exist after modelling.

Figure 1.1 shows a simple example. We want to estimate the relationship between an exposure (E) and an outcome (O). C1 and C2 are confounders in that they may affect one another and they both affect E and O. Note that the direction of the arrows means that neither C1 nor C2 are affected by E or O. Thus, E could be stress as measured by the Perceived Stress Scale (PSS) and O could be high blood pressure. Then C1 could be age and C2 ethnicity. Although age and ethnicity are not causally related, in the UK ethnic minorities tend to be younger than the rest of the population. Older people and ethnic minorities may have more stress and have higher blood pressure for reasons other than stress. Thus, in a population that includes a wide range of ages and ethnicities we need to allow for these variable when considering whether stress causes high blood pressure.

An important condition for a variable to be a confounder is that it is not on the direct casual path. This is shown in Figure 1.2, where an intermediate variable (IM) is on the causal path between E and O. An example might be that stress causes people to drink alcohol and alcohol is the actual cause of high blood pressure. To control for alcohol, one might look at two models with different levels of drinking. One might fit a model with and without the intermediate factor, to see how the relationship between E and O changes.

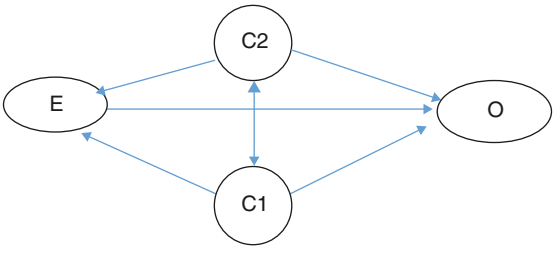

Figure 1.1 A DAG showing confounding.

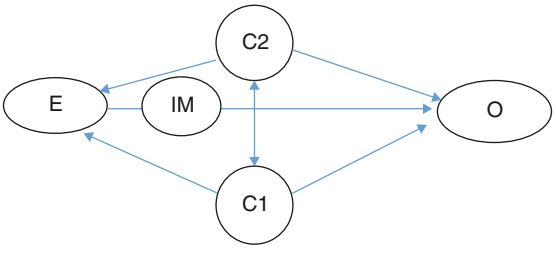

Figure 1.2 A DAG showing an intermediate variable (IM).

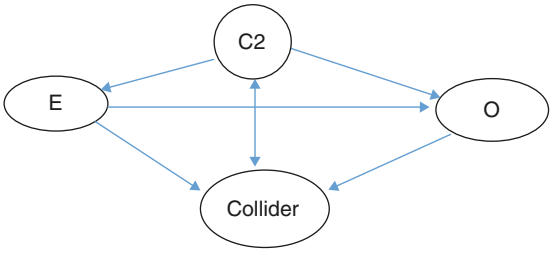

Figure 1.3 A DAG showing a collider.

One use of DAGs is to identify what is known as *Berkson's bias*. This is where the arrows are reversed going to one particular variable, and so they collide at this variable; this variable is called a *collider* (see Figure 1.3). This is the situation where having both O and E increases your chance of the collider. To extend the stress example, a hospital may run a cardiovascular clinic and so the investigators might choose cases of high blood pressure from the clinic and controls as people not in the clinic. However, stress may cause symptoms of cardiovascular disease and so stressed people are more likely to attend the clinic, which sets up a spurious association between stress and high blood pressure.

In general, allowing for confounders in models gives a better estimate of the strength of a causal relationship, whereas allowing for IMs and colliders does not and so it is important to identify which are which. DAGs are a qualitative way of expressing relationships, and one doesn't often see them in publications. They also have their limitations, such as in displaying effect modifiers.[2] Relationships can also depend on how a variable is coded, such as an absolute risk or a relative risk. Statistical models are useful for actually quantifying and clarifying these relationships.

1.4 Statistical Models

The relationship between inputs and outputs can be described by a mathematical model that relates the inputs, which we have described earlier with causal variables, confounders and effect modifiers (often called "independent variables" and denoted by X), with the output (often called "dependent variable" and denoted by Y). Thus, in the stress and blood pressure example above, we denote blood pressure by Y, and stress and gender are both X variables. Here the X does not distinguish between confounding and causality. We wish to know if stress is still a good predictor of blood pressure when we know an individual's gender. To do this we need to assume that gender and stress combine in some way to affect blood pressure. As discussed in *Statistics at Square One*, we describe the models at a *population* level. We take samples to get estimates of the population values. In general we will refer to population values using Greek letters and estimates using Roman letters.

The most commonly used models are known as "linear models". They assume that the X variables combine in a linear fashion to predict Y. Thus, if X_1 and X_2 are the two independent

variables we assume that an equation of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ is the best predictor of Y where β_0, β_1 and β_2 are constants and are known as parameters of the model. The method often used for estimating the *parameters* is known as regression and so these are the *regression parameters*. The estimates are often referred to as the "regression coefficients". Slightly misleadingly, the X variables do not need to be independent of each other so another confounder in the stress/blood pressure relationship might be employment, and age and employment are related, so for example older people are more likely to be retired. This can be seen in Figure 1.1, where confounding variables may be linked. Another problem with the term "linear" is that it may include interactions, so the model may be of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. An effect modifier, described in the previous section, may be modelled as an interaction between a possible cause X_1 and a possible confounder X_2.

Of course, no model can predict the Y variable perfectly and the model acknowledges this by incorporating an *error* term. These linear models are appropriate when the outcome variable is continuous. The wonderful aspect of these models is that they can be generalised so that the modelling procedure is similar for many different situations, such as when the outcome is non-Normal or discrete. Thus, different areas of statistics, such as t-tests and chi-squared tests are unified and dealt with in a similar manner using a method known as "generalised linear models".

When we have taken a sample, we can estimate the parameters of the model, and get a fit to the data. A simple description of the way that data relate to the model is given by Chatfield.[3]

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The FIT is what is obtained from the model given the predictor variables. The RESIDUAL is the difference between the DATA and the FIT. For the linear model the residual is an estimate of the error term. For a generalised linear model this is not strictly the case, but the residual is useful for diagnosing poor fitting models, as we shall see later.

Models are used for two main purposes, *estimation* and *prediction*. For example we may wish to estimate the effect of stress on blood pressure, or predict what the lung function of an individual is given their age, height and gender.

Do not forget, however, that models are simply an approximation to reality. "All models are wrong, but some are useful."

The subsequent chapters describe different models where the dependent variable takes different forms: continuous, binary, a survival time, a count and ordinal and when the values are correlated such as when they are clustered or measurements are repeated on the same unit. A more advanced text covering similar material is that by Frank Harrell.[4] The rest of this chapter is a quick review of the basics covered in *Statistics at Square One*.

1.5 Results of Fitting Models

Models are fitted to data using a variety of methods. The oldest is the method of least squares, which finds values for the parameters that minimise the sum of the squared residuals. Another is *maximum likelihood*, which finds the values of the parameters that gives

the highest *likelihood* of the data given the parameter estimates (see Appendix 2 for more details). For Normally distributed data the least squares method is also the maximum likelihood method. The output from a computer package will be an estimate of the parameter with an estimate of its variability (the standard error or SE). There will usually be a p-value and a confidence interval (CI) for the parameter. A further option is to use a *robust* standard error, or a *bootstrap standard error*, which are less dependent on the model assumptions and are described in Appendix 3. There will also be some measures as to whether the model is a good fit to the data.

1.6 Significance Tests

Significance tests such as the chi-squared test and the t-test, and the interpretation of p-values were described in *Statistics at Square One*. The usual format of statistical significance testing is to set up a *null hypothesis* and then collect data. Using the null hypothesis, we test if the observed data are consistent with the null hypothesis. As an example, consider a randomised clinical trial to compare a new diet with a standard diet to reduce weight in obese patients. The null hypothesis is that there is no difference between the two treatments in weight changes of the patients. The outcome is the difference in the mean weight after the two treatments. We can calculate the probability of getting the observed mean difference (or one more extreme), if the null hypothesis of no difference in the two diets is true. If this probability (the p-value) is sufficiently small we reject the null hypothesis and assume that the new diet differs from the standard. The usual method of doing this is to divide the mean difference in weight in the two diet groups by the estimated SE of the difference and compare this ratio to either a t-distribution (small sample) or a Normal distribution (large sample).

The test as described above is known as Student's t-test, but the form of the test, whereby an estimate is divided by its SE and compared to a Normal distribution, is known as a *Wald test* or a *z-test*.

There are, in fact, a large number of different types of statistical test. For Normally distributed data, they usually give the same p-values, but for other types of data they can give different results. In the medical literature there are three different tests that are commonly used, and it is important to be aware of the basis of their construction and their differences. These tests are known as the *Wald test*, the *score test* and the *likelihood ratio test*. For non-Normally distributed data, they can give different p-values, although usually the results converge as the data set increases in size. The basis for these three tests is described in Appendix 2.

In recent times there has been much controversy over significance tests.[5] They appear to answer a complex question with a simple answer, and as a consequence are often misused and misinterpreted. In particular, a non-significant p-value is supposed to indicate a lack of an effect, and a significant p-value to indicate an important effect. These misconceptions are discussed extensively in *Statistics at Square One*. The authors of this book believe they are still useful and so we will use them. It is one of our goals that this book will help reduce their misuse.

1.7 Confidence Intervals

One problem with statistical tests is that the p-value depends on the size of the data set. With a large enough data set, it would be almost always possible to prove that two treatments differed significantly, albeit by small amounts. It is important to present the results of an analysis with an estimate of the mean effect, and a measure of precision, such as a CI.[6] To understand a CI we need to consider the difference between a population and a sample. A population is a group to whom we make generalisations, such as patients with diabetes or middle-aged men. Populations have *parameters*, such as the mean HbA1c in people with diabetes or the mean blood pressure in middle-aged men. Models are used to model populations and so the parameters in a model are population parameters. We take samples to get *estimates* for model parameters. We cannot expect the estimate of a model parameter to be exactly equal to the true model parameter, but as the sample gets larger we would expect the estimate to get closer to the true value, and a CI about the estimate helps to quantify this. A 95% CI for a population mean implies that if we took 100 samples of a fixed size, and calculated the mean and 95% CI for each, then we would expect 95 of the intervals to include the true population parameter. The way they are commonly understood from a single sample is that there is a 95% chance that the population parameter is in the 95% CI. Another way of interpreting a CI is to say it is a set of values of the null hypothesis, from which the observed data would not be statistically significant. This points out that just as there are three commonly used methods to find p-values, there are also a number of different methods to find CIs, and the method should be stated.[6]

In the diet trial example given above, the CI will measure how precisely we can estimate the effect of the new diet. If in fact the new diet were no different from the old, we would expect the CI for the effect measure to contain 0.

Cynics sometimes say that a CI is often used as a proxy for a significance test, that is, the writer simply reports whether the CI includes the null hypothesis. However, using CIs emphasises *estimation* rather than *tests* and we believe this is the important goal of analysis, that is, it is better to say being vaccinated reduces your risk of catching Covid by a factor of 95% (95% CI 90.3 to 97.6) than to simply say vaccination protects you from Covid ($P < 0.001$).[7]

CIs are also useful in *non-inferiority* studies, where one might want to show that two treatments are effectively equivalent, but perhaps one has fewer side effects than the other. Here one has to specify a non-inferiority margin, and conclude non-inferiority if the CI does not include the margin but does include a difference of zero. The concepts of null and alternative hypotheses are reversed and so require careful thought. Further discussion is given, for example, by Hahn.[8]

1.8 Statistical Tests Using Models

A t-test compares the mean values of a continuous variable in two groups. This can be written as a linear model. In the example of the trial of two diets given above, weight after treatment was the continuous variable. Here the primary predictor variable X is Diet,