# 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.

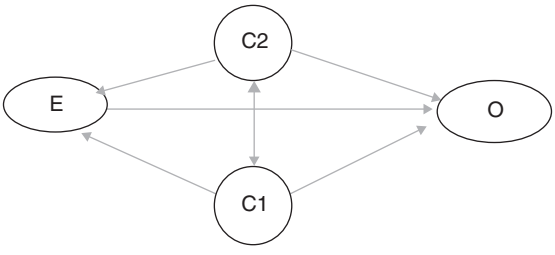


**Figure 1.1** A DAG showing confounding.

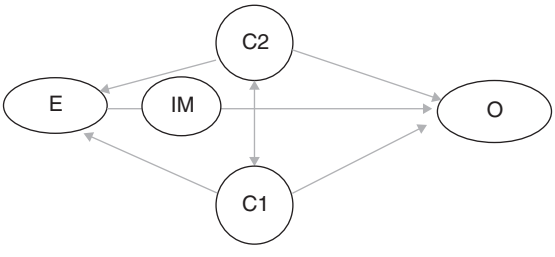


**Figure 1.2** A DAG showing an intermediate variable (IM).

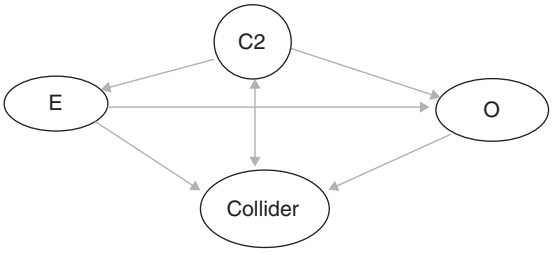


**Figure 1.3** A DAG showing a collider.

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 $\beta_0$, $\beta_1$ and $\beta_2$ 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,

which is a binary variable taking the value (say) 0 for the standard diet and 1 for the new diet. The outcome variable is weight. There are no confounding variables (in theory because this is a randomised trial). The fitted model is Weight = $b_0 + b_1 \times$ Diet + Residual. The FIT part of the model is $b_0 + b_1 \times$ Diet and is what we would predict someone's weight to be given our estimate of the effect of the diet. We assume that the residuals have an approximate Normal distribution. The null hypothesis is that the coefficient associated with diet $b_1$ is from a population with mean 0. Thus, we assume that $\beta_1$, the population parameter, is 0 and, rather than using a simple test, we can use a model. The results from a t-test and linear regression are compared in Appendix 3.

Models enable us to make our assumptions explicit. A nice feature about models, as opposed to tests, is that they are easily extended. Thus, weight at baseline may (by chance) differ in the two groups, and will be related to weight after treatment, so it could be included as a confounder variable. Similarly, smoking may be an effect modifier and so that could be included in the model as well.

This method is further described in Chapter 2 as multiple regression. The treatment of the chi-squared test as a model is described in Chapter 3 under logistic regression.

## 1.9 Many Variables

It may seem merely semantic but it is useful to distinguish between *multivariate* and *multivariable* analysis.[9] A multivariable model can be thought of as a model in which multiple variables are found on the right side of the model equation. This type of statistical model can be used to attempt to assess the relationship between a number of variables; one can assess independent relationships while adjusting for potential confounders. There is only one outcome variable. For models such as linear regression, the generalisation with many predictors is known as *multiple linear regression*.

Multivariate, by contrast, refers to the modelling of data that are often derived from longitudinal studies, wherein an outcome is measured for the same individual at multiple time points (repeated measures), or the modelling of nested/clustered data, wherein there are multiple individuals in each cluster. Techniques for multivariate analysis include *factor analysis, cluster methods* and *discriminant analysis*. In this book, we are more interested in analyses where there is only one outcome measure and a variety of possible predictor variables which can be combined into a prediction, and so we do not consider multivariate methods. Thus, it is a misnomer to call techniques such as multivariable regression as *multivariate* since there is only one outcome variable and so it is better to refer to multiple linear regression (Chapter 2) and similarly with multiple logistic regression (Chapter 3).

Using a multivariable model means we can estimate the relationship between exposure (E) and outcome (O) in Figure 1.1 *controlling* for the confounders $C_1$ and $C_2$. For example, in an observational study, a confounder may be age, and the age distribution for those exposed may be different from that of those not exposed. We can use a multivariable model to estimate the relationship between E and O *allowing for* age. Similarly, we can explore how different variables interact when predicting the outcome variable. In a

clinical trial, we can explore whether the treatment effect varies by (pre-specified) subgroups. We can use models for prediction, for example predict lung function given a person's personal characteristics. The great advantage of a model is that we can predict what a particular person's expected lung function should be, given say their age, height, gender and smoking habit *even though* a person with these precise characteristics is not in the data set.

A reader should always question why covariates are included in a regression.[10] A common error is to only include covariates which are statistically significantly related to the exposure in the model. In clinical trials, the reason is that randomisation, if done correctly, will have rendered the relationship between a treatment and a covariate null, and any observed difference will, by definition, be due to chance. A more fundamental reason is that such a test does not tell us whether the covariate should be included; non-significant covariates can be important effect modifiers and covariates may act by interaction with other covariates which singular testing would not reveal. It also makes the study more difficult to repeat; should a new investigator choose covariates which a previous investigator found by individual testing, even if they are not significant in the second study? The usual advice is to pre-specify the covariates to be used in a study but this is sometimes to assign divine knowledge to the investigator, who will want to discover relationships not previously anticipated. It can help to decide on whether the study is *explanatory* or *confirmatory*, as explained in Section 1.10.

## 1.10 Model Fitting and Analysis: Exploratory and Confirmatory Analyses

There are two aspects to data analysis: confirmatory and exploratory analyses. In a *confirmatory analysis* we are testing a pre-specified hypothesis and it follows naturally to conduct significance tests. Testing for a treatment effect in a clinical trial is a good example of a confirmatory analysis. In an *exploratory analysis* we are looking to see what the data are telling us. An example would be looking for risk factors in a cohort study. A technique such as *stepwise* regression, which tries to find the best fitting model out of a set of covariates, is an example of an explanatory technique. The findings should be regarded as tentative to be confirmed in a subsequent study, and p-values are largely decorative.

Often one can do both types of analysis in the same study. For example, when analysing a clinical trial, a large number of possible outcomes may have been measured. Those specified in the protocol as primary outcomes are subjected to a confirmatory analysis, but there is often a large amount of information, say concerning side effects, which could also be analysed. These should be reported, but with a warning that they emerged from the analysis and not from a pre-specified hypothesis. It seems illogical to ignore information in a study, but also the lure of an apparent unexpected significant result can be very difficult to resist (but should be).

It may also be useful to distinguish *audit*, which is largely descriptive, intending to provide information about one particular time and place, and *research*, which tries to be generalisable to other times and places.

## 1.11 Computer-intensive Methods

Much of the theory described in the rest of this book requires some prescription of a distribution for the data, such as the Normal distribution. There are now methods available which use models but are less dependent on the actual distribution of the data. They are commonly available in computer packages and are easy to use. A description of one such method, the bootstrap, is given in Appendix 3.1. For *prediction*, there are now many methods which include *machine learning* or *artificial intelligence*, for example to predict Covid-19 progression.[11] They eschew a fixed model, and instead learn from many combinations of the data to find which combinations give the best predictions. An advantage of these methods is that real life does not behave like a model and so they may be better at predicting events than trying to find a model to do so. The problem with these methods is that because they are "black-box" they don't really explain how their inputs relate to each other, and their predictions are often not associated with measures of uncertainty. In future, they may replace models, but not as yet since their utility is still being explored.

## 1.12 Missing Values

Missing values are the bane of a statistical analyst's life and are usually not discussed in elementary textbooks. In any survey, for example, some people will not respond; at worst we need to know how many are missing and at best we would like some data on them, say their age and gender. Then we can make some elementary checks to see if the subjects who did respond are typical. (One usually finds that the worst responders are young and male). One then has to decide whether anything needs to be done. For longitudinal data, it is important to distinguish values missing in the main outcome variables and values missing in covariates. For the outcome variables, missing values are often characterised into one of three groups: (1) missing completely at random (MCAR); (2) missing at random (MAR) and (3) not missing at random (NMAR) or non-ignorable (NI).[12] The crucial difference between (1) and (2) is that for (2) the reason for a value being missing can depend on previously recorded input and outcome variables, but must not depend on the value that is missing beyond what can be explained by other variables in the model. Thus a blood pressure value would be NMAR (NI) if it were missing every time the blood pressure exceeded 180 mmHg (which we cannot measure but can say that it made the patient too ill to turn up). However, if it were missing because the previous value exceeded 180 mmHg and the patient was then taken out of the study then it may be MAR. The important point to be made is that the reason for missing in MAR is independent of the actual value of the observation *conditional* on previous observations.

In longitudinal clinical trials it used to be traditional to ignore subjects if the values were missing. However, this can lead to biased and inefficient treatment estimates. The usual method of dealing with missing values is called *last observation carried forward* (LOCF), which does exactly what it says. However, this can also lead to bias and a number of other techniques have been developed including *imputation*, where the missing value is guessed from other values. *Multiple imputation* gives a distribution of possible values and enables

uncertainty about the missing values to be incorporated in the analysis. The use of random effects models is also a way of analysing all the data measured for some longitudinal designs (see Chapter 5). Care, thought and sensitivity analyses are needed with missing data. For further details see, for example, Little and Rubin.[12]

## 1.13 Bayesian Methods

The model-based approach to statistics leads one to statements such as: "given model M, the probability of obtaining data D is P." This is known as the *frequentist* approach. This assumes that population parameters are fixed. However, many investigators would like to make statements about the probability of model M being true, in the form "given the data D, what is the probability that model M is the correct one?" Thus, one would like to know, for example, what is the probability of a diet working. A statement of this form would be particularly helpful for people who have to make decisions about individual patients. This leads to a way of thinking known as "Bayesian", which allows population parameters to vary. It has particular uses in health economics and health decision making where we would like to know the probability of a drug being cost effective.[13] Bayesian thinking is often facilitated by the use of DAGs.

This book is largely based on the frequentist approach. Most computer packages are also based on this approach, although most now enable Bayesian analysis as well. Further discussion on Bayesian methods is given in Chapter 5 and Appendix 4.

## 1.14 Causal Modelling

The models described earlier could be described as causal models. However, there is a specific area of *causal modelling* which is increasingly becoming more prevalent in biomedicine. It is beyond the scope of this book and here we just mention some areas that we have used in our own research, so that the reader will recognise them when reading the medical literature.

The models described in this book have the dependent variable on the left of the equation and the independent variables on the right. In econometrics the independent variables are termed *exogenous* and the dependent variables *endogenous* (probably better terms than the ones used here) and econometricians have long dealt with the observational data where the endogenous variables can be on the right-hand side as well. For example, consider a series of cross-sectional studies of children where we are interested in obesity.[14] We may be interested in factors influencing the obesity of children at age seven conditional of the obesity of the children when they were younger. The interest is obesity at age seven, but obesity at younger ages are endogenous variables to be included in the model. Each endogenous variable will have its own set of predictor (exogenous) variables, which gives a series of simultaneous equations. This is known as a *structural equation model*. These models often look at concepts which cannot be observed directly, or are measured with error. These are known as *latent variables*. For example Grey *et al*.[14] considered "family lifestyle" as a latent

variable. They concluded that family lifestyle has a significant influence on all outcomes in their study, including diet, exercise and parental weight status. Family lifestyle accounted for 11.3% of the variation in child weight by the age of seven.

A useful philosophical idea for thinking about causality is that of the *counterfactual*, that is, consider two parallel universes: in one universe a person gets a treatment and in the other they get a control. The difference in outcomes for the *same* person is the effect of treatment. The closest we can get to this is a cross-over trial, where we randomise the order of treatment to a single person, but this requires that after the first treatment the person can return to a pre-treatment state before receiving the second. The next closest is a parallel randomised trial. Counterfactuals are used in treatment switching trials, comparing two treatments and where the length of survival is the outcome.[15] Data gathered earlier in the trial might indicate that one treatment is superior to the another and so all patients still in the trial are switched to the superior treatment. However, to estimate the cost-benefit of treatment we would wish to know how long people would have survived *if* they had stayed on the original treatment and so need to estimate the counterfactual using some form of predictive model based on patient characteristics.

A common problem in observational data is to evaluate a treatment where randomisation is not possible. Consider a study of the effect of breastfeeding of babies on the obesity of the child seven years later.[14] The issue is that there are confounding factors that determine whether a mother breastfeeds which may also determine whether the child is subsequently obese, such as income or education. This is usually tackled using linear or logistic regression which are discussed in Chapters 2 and 3. However, an alternative method is to use *propensity scores*.[16] Here models use the characteristics of a mother to predict the probability that she will breastfeed. Mothers who breastfeed are then matched with those who don't and a measure of obesity such as the body mass index of the children at seven are compared. There are various methods of matching, including pair matching, stratification of the propensity score or the use of the score as a covariate in a regression model. Propensity score methods rely on the assumption that we have measured *all* the covariates that determine breastfeeding to give unbiased results.[17] Some empirical studies, comparing estimates obtained from observational data using propensity scores and those from conventional trials, have shown that trials show more modest effects than estimates using propensity scores.[17] Traditional epidemiologists are very cautious about the term "causal" and will invoke criteria such as those of Bradford-Hill to bolster their claims. These criteria include the strength of the relationship, the plausibility of it, a dose response and the exposure must come before the outcome (temporality). For a discussion of how these criteria work in modern molecular biology, see the article by Fedak *et al*.[18] In trials where randomisation occurs, there are fewer inhibitions in making conclusions but it is perhaps unfortunate that some users of propensity score methods, which are, after all, based on observational data, take on the confident tone of trialist with regard to causality.

The whole issue of causal modelling can become rather "black-box" where insight as to how inferences are obtained can rest on many unacknowledged assumptions. For further reading see, for example, Imbens and Rubin.[19]

## 1.15 Reporting Statistical Results in the Medical Literature

The reporting of statistical results in the medical literature often leaves something to be desired. Here we will briefly give some tips that can be generally applied. In subsequent chapters we will consider specialised analyses.

Lang and Secic[20] give an excellent description of a variety of methods for reporting statistics in the medical literature. Checklists for reading and reporting statistical analyses are also given in Mansournia *et al.*[21] A huge variety of checklists for reporting results from different types of study are given on the Equator website.[22] These include the CONSORT statement for clinical trials[23] and the STROBE statement for observational studies.[24]

- Always describe how the subjects were recruited and how many were entered into the study and how many dropped out. For clinical trials one should say how many were screened for entry and describe the drop-outs by treatment group.
- Describe the model used and assumptions underlying the model and how these were verified. Always give an estimate of the main effect, with a measure of precision, such as a 95% CI as well as the p-value. It is important to give the right estimate. Thus, in a clinical trial, while it is of interest to have the mean of the outcome, by treatment group, the main measure of the effect is the difference in means and a CI for the difference. This can often not be derived from the CIs of the means for each treatment.
- It is often sensible to apply several models for the same data to show how including variables changes the model estimates. In the stress example, give the coefficients for separate models allowing sequentially for, for example, gender and alcohol consumption, and by age.
- Describe how the p-values were obtained (Wald, likelihood ratio or score) or the actual tests and similarly with CIs.
- It is sometimes useful to describe the data using binary data (e.g. percentage of people with hypertension), but analyse the continuous measurement (e.g. blood pressure).
- Describe which computer package was used. This will often explain why a particular test was used. Results from "home-grown" programs may need further verification. It is very helpful to make the code and data available for future analysts.

## 1.16 Reading Statistics in the Medical Literature

- From what population are the data drawn? Are the results generalisable?
- Can you tell how much of the original data collected appeared in the report and was it a high proportion? Did many people refuse to cooperate? Were missing values investigated using imputation or sensitivity analyses?
- Is the analysis confirmatory or exploratory? Is it research or audit?
- Have the correct statistical models been used?
- Do not be satisfied with statements such as, "a significant effect was found". Ask what is the size of the effect and will it make a difference to patients (often described as a "clinically significant effect")?
- Verify that the assumptions of the model are met. For example, if a linear model is used, what evidence is there for linearity. If paired data are used, is pairing reflected in the analysis?

- Are the results critically dependent on the assumptions about the models? Often the results are quite "robust" to the actual model, but this needs to be considered. For example if the data set is large the assumption of Normality for residuals is less critical.
- How were the confounders chosen? As stated earlier it is a classic mistake to use individual significance tests to see whether a potential confounder differs between treatment and control (in a trial) or between exposed and not exposed (in an observational study). The problem is that the significance test does not tell us whether a variable should be included in a model. Ideally, variables used in a model should be specified in advance. However, this may not be possible, and the analysis should be described as *exploratory*.
- Were different models shown with and without possible confounders to describe the effects of confounding and effect modification?

## References


1 Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.

2 Weinberg CR. Can DAGs clarify effect modification? *Epidemiology* 2007; **18**(5): 569–72. doi: 10.1097/EDE.0b013e318126c11d

3 Chatfield C. *Problem Solving: A Statistician's Guide*. London: Chapman and Hall, 1995.

4 Harrell FE. *Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis*. New York: Wiley, Springer Series is Statistics, 2015.

5 Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process and purpose. *Am Stat* 2016; **70**: 129–33.

6 Altman D, Machin D, Bryant T, Gardner M, eds. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. Chichester: John Wiley & Sons, 2013.

7 Polack FP, Thomas SJ, Kitchin N, *et al*. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020; **383**: 2603–15.

8 Hahn S. Understanding non-inferiority trials. *Korean J Pediatr* 2012; **55**(11): 403–7. doi: 10.3345/kjp.2012.55.11.403

9 Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health* 2013; **103**(1): 39–40.

10 Nojima M, Tokunaga M, Nagamura F. Quantitative investigation of inappropriate regression model construction and the importance of medical statistics experts in observational medical research: a cross-sectional study. *BMJ Open* 2018; **8**: e021129. doi: 10.1136/bmjopen-2017-021129

11 Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19* 2021; 381–97. doi: 10.1016/B978-0-12-824536-1.00027-7

12 Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Chichester: John Wiley, 2019.

13 Baio G. *Bayesian Methods in Health Economics*. Boca Raton, FL: CRC Press, 2013.

14 Gray LA, Alava MH, Kelly MP, Campbell MJ. Family lifestyle dynamics and childhood obesity: evidence from the millennium cohort study. *BMC Public Health* 2018; **18**(1): 500.

**15** Latimer NR, Abrams KR, Lambert PC, *et al*. Adjusting for treatment switching in randomised controlled trials–a simulation study and a simplified two-stage method. *Stat Methods Med Res* 2017; **26**(2): 724–51.
**16** Gibson LA, Hernández Alava M, Kelly MP, Campbell MJ. The effects of breastfeeding on childhood BMI: a propensity score matching approach. *J Public Health* 2017; **39**(4): e152–60.
**17** Campbell MJ. What is propensity score modelling? *Emerg Med J* 2017; **34**: 129–31. doi: 10.1136/emermed-2016-206542
**18** Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol* 2015; **12**: 14. doi: 10.1186/s12982-015-0037-4
**19** Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press, 2015.
**20** Lang TA, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors and Reviewers*. Philadelphia, PA: American College of Physicians, 2006.
**21** Mansournia MA, Collins GS, Nielsen RO, *et al*. A checklist for statistical assessment of medical papers (the CHAMP statement): explanation and elaboration. *Br J Sports Med* 2021; **55**: 1009–17. doi: 10.1136/bjsports-2020-103652
**22** The EQUATOR Network. Enhancing the QUAlity and Transparency Of Health Research (equator-network.org)
**23** CONSORT Consort – Welcome to the CONSORT Website (www.consort-statement.org)
**24** STROBE Checklists – STROBE (www.strobe-statement.org)