INVITED FEATURE: STATISTICS PRIMER

What statistical software should I use? And does it actually matter?

Andrew R. Gray, Claire Cameron, Ari Samaranayaka, Robin M. Turner

Introduction

With so many options available, a question that researchers often ask us as biostatisticians is, "what statistical software should I use?" When faced with performing statistical analyses, if you search online, talk to your colleagues or supervisors, or browse your favourite bookstore, you will potentially find many, many options. Wikipedia lists over one hundred![1] This article will focus on general advice and look at software commonly used in the health sciences, specifically (in alphabetical order), GraphPad Prism, IBM SPSS Statistics, Microsoft Excel, R, SAS, and Stata.

An important point to keep in mind is that *using statistical software* is not the same thing as *doing biostatistics*. As Stuart Pocock, a notable medical statistician, said nearly forty years ago:

> *"I would like to refer briefly to the frequent misuse of statistical packages. Since they make each analysis task so easy to perform, there is a real danger that the user requests a whole range of analyses without any clear conception of what [they are] looking for… Thus, my main message here is that use of computers is no substitute for clear thought."*[2]

We should carefully think about our statistical analyses long before we have data in hand, preferably when we're first starting to design our study. These analyses are important when we're preparing funding applications, study registrations, ethics applications, protocol papers, and (most obviously) statistical analysis plans. We should also keep in mind that we will need to adhere to reporting guidelines, such as the Consolidated Standards of Reporting Trials (CONSORT) and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statements.

A decision-making process

We suggest that there are four steps to consider when deciding what statistical software to use. It's useful to keep in mind that on occasion we find that more than one statistical programme will be needed.

1. What do I need to do? A good start is understanding where you want to go in terms of your statistical analyses. This will require a solid understanding of your research question(s), the type(s) of data you anticipate collecting, the analyses you will need to precisely address your research question(s), and how you plan to communicate your findings. Most of the time, it will be useful to involve a biostatistician early in the project to make sure you're not missing something that will cause problems further down the line.
2. What do I already know how to do? If you already have experience with particular statistical software, it could be efficient to keep using that (as long as it will do what you need, of course).
3. What resources do I have (or can I get) access to? If you have identified that you will need to use software you're not already familiar with, or unfamiliar parts of software you have used before, you'll need to consider how you can reach the necessary level of expertise. If you have time, you might be able to find courses that will help you close the gap between what you already know and what you need to know. Some software might incur a cost depending on your institution's software licences. We find it's worth asking the information technology specialists at your institution in case a licence is available that makes commercial software either free, or very cheap. If you're planning to do the analyses yourself, the software used by your colleagues or supervisor(s) might be another consideration.
4. What do I want to be able to do in the future? If the first two points match up, you might be tempted to stop thinking about statistical software. Planning and implementing a research project is time-consuming, and it can be very easy to forget, at least momentarily, that this might not be the only time you need to do your own statistical analyses. It's reassuring to keep in mind that learning to use one programme will help develop your confidence with other programmes[3] in the future. It's also important to remember that whatever software you choose to use now might not exist, in its current form, or at all, for your entire research career! This is another excellent point to discuss with a biostatistician.

But at least I can use whatever software I want and get the same results, right?

While this is broadly true, it's important to note that software packages will sometimes give different results due to different default options. There are several options when we perform even simple analyses such as Chi-squared and Mann-Whitney U tests (such as continuity corrections and exact versus asymptotic tests). Sometimes you can change these options yourself, and other times, the designers of the software have already made these decisions for you. Forgetting this point can lead to frustration when a co-investigator, who might be a biostatistician, using different software, keeps getting slightly different results.

A danger with software that presents multiple versions of statistical tests alongside each other, as SAS and SPSS Statistics sometimes do, is that we might (subconsciously) choose the method or option that will give the most favourable p-value, having not decided in advance which is the most appropriate. Having a sufficiently detailed plan of the statistical analyses before they are performed is essential.

Writing your statistical methods

Once you've used your chosen software to perform your analyses, you will need to provide enough detail in your work (whether this is a journal article, thesis, or something else) so that the reader can understand what you did. A good goal to keep in mind when doing

this is to provide enough information so that a competent (bio)statistician with access to your data, but not your actual computer code or step-by-step instructions for using a point-and-click interface, could reproduce your results without extensive trial and error. This includes citing the software used and any user-written packages you've relied on (including the versions for both). It's crucial to explain when non-default options were used. Similarly, if you have chosen between approaches depending on what you find when you explore your data or statistical models, it's essential that you explain clearly how and why you have done this.

Quick overviews of some software options

This section will provide an overview of some statistical software commonly used in health research (in approximate order of complexity). All of these programmes have recent updates, and we anticipate that new versions of each will be released in the future. There are many other statistics programmes that we do not have space to cover and come across less frequently.

MICROSOFT EXCEL

While Microsoft Excel is very limited in terms of the statistical analyses it can do, it has the advantage of familiarity for many researchers. It can be used for activities such as data entry, data cleaning, "mechanical" calculations (where the same formula is repeatedly applied to data, such as with standardisation), descriptive statistics, and constructing tables and figures. There are inbuilt functions and extensions adding more statistical methods, but the authors wouldn't recommend using these over dedicated statistical software. We sometimes see researchers who started using Microsoft Excel because it seemed easier, but who then found that they couldn't do what they needed, or what reviewers or examiners asked them to do.

GRAPHPAD PRISM

GraphPad Prism has less functionality compared to the other options below, reflecting the statistical tests, models, and graphs that its designers think their users will want. It encourages a point-and-click approach, and while this can be very appealing for beginners, we find that this approach makes it much harder for researchers to update or expand the statistical analyses they've performed. Since reviewers and examiners will often request further or different analyses, we strongly prefer software that allows the use of code, as the four packages described below do.

IBM SPSS STATISTICS

While not used by the authors, SPSS Statistics is very popular among researchers and provides a point-and-click interface alongside a code-based approach. Part of the reason for the authors not using it is that we find SPSS Statistics syntax to be less straightforward than other software, making it harder for researchers to learn, write, read, and understand and make changes to their analyses in the future. If you're using SPSS Statistics, we strongly recommend that you save the code generated by the dialogue boxes so that you can reproduce your analyses on updated data or with modifications to the analyses themselves. The output from SPSS Statistics is often voluminous. For us, finding precisely the information we need to report can sometimes be more challenging than performing the analysis itself!

STATA

This is the software we use the most. We find that it provides most of what most biostatisticians want, and close to, if not everything, a non-biostatistician should want to do. Using fairly simple syntax, which is generally consistent between functions, you can perform interactive analyses, or submit a batch of code all at once. There is also the option of doing almost everything through a point-and-click interface, which conveniently displays the code that would have done the same thing, helping you to learn new commands and options as you go. A particular virtue of Stata is that it provides succinct output where you need to ask if you want more. While this might initially seem frustrating, it helps you to focus on what you currently need to know and reduces the risk of second-guessing options based on results. Another attractive feature of Stata is its user-friendly documentation that explains its commands in both non-technical language and with the technical details that statisticians sometimes want. Stata has an active user community, with many user-written programmes available. Over time, some of these have been added to Stata itself.

SAS

SAS is especially strong in data management and when working with multiple data sets. Like Stata, it's feature-rich, but its syntax is verbose, and we find it can take more time to learn. Like SPSS Statistics and other programmes that were developed when computers were more expensive and less accessible, its output is often lengthy and sometimes includes results using multiple options.

R

The freely available software R is, at heart, a programming language that has excellent support for statistical methods. Sometimes it's the first way to use new (bio)statistical methods. One challenge with R is the paradox of choice. R has a huge number of user-written packages. While other software might provide one standard way of doing something, community-written packages for R will often present the researcher with an unavoidable choice about exactly how they want to perform their analyses. Sometimes, different package authors will have their own preferences about options and extensions. This can lead to different results, or a researcher realising part-way through their statistical analyses that another package would have been a better choice. At the same time, there has been a lot of work towards making tasks more consistent (e.g. the Tidyverse) and enhancing the quality of packages (e.g. the Comprehensive R Archive Network (CRAN) and Bioconductor). Like Stata, R encourages you to work interactively with your data, and is succinct in its output.

What next?

Biostatistics features throughout the research cycle,[4] and different statistical software makes different things possible or easier. By carefully thinking through what you want to do, you can determine the statistical analyses you'll need (although these requirements might change later if your research doesn't go to plan), and then decide what software is best for performing these analyses.

We often find ourselves recommending Stata to researchers wanting, or at least needing, to perform their own statistical analyses, as it has almost all of the methods we would recommend, encourages good practice in being code-based (but with menus and dialogue boxes to remind us of the commands when we forget, alongside some excellent manuals and books), and facilitates focused exploration of our data and statistical models.

As always, talking to a biostatistician early in your research can avoid many problems further down the line. If possible, ensuring that a biostatistician is part of your research team will optimise the match between the research question, the statistical approaches, and the statistical software used.

References

1. Wikipedia Contributors. List of statistical software [Internet]. 2021 [cited 2021 Oct 26]. Available from https://en.wikipedia.org/wiki/List_of_statistical_software.

2. Pocock SJ. Clinical trials: a practical approach. Chichester: John Wiley & Sons Ltd; 1983.

3. Rode JB, Ringel MM. Statistical software output in the classroom: a comparison of R and SPSS. Teach Psychol. 2019 Oct;46(4):319-27.

4. Samaranayaka A, Cameron C, Turner R. Sample size in health research. New Zealand Medical Student Journal. 2021 Apr;(32):61-2.

About the authors

› Mr Andrew Gray, BCom(Hons), BA, is a Senior Research Fellow and Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Associate Professor Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is a Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Dr Ari Samaranayaka, BSc, MPhil, PhD, is a Senior Research Fellow and Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Professor Robin Turner, BSc(Hons), MBiostat, PhD, is the Director of the Biostatistics Centre, Division of Health Sciences, University of Otago.

Correspondence

Andrew Gray: andrew.gray@otago.ac.nz