Statistics and its relevance to data stewardship

An interview

Interviewee: Nigel Yoccoz, UiT The Arctic University of Norway, ORCID: 0000-0003-2192-1039

Interviewer: Kathleen Alexandra Smart, UiT The Arctic University of Norway,
ORCID: 0000-0003-4231-7519

Date: September 2022

Transcript

[Kathleen:]

Welcome, everyone.

Today, we're going to be discussing how to apply statistics to research data management when you're conducting your research plan.

And we're going to have an interview with Nigel Yoccoz, who is a professor here at UiT in Statistical Ecology at the Department of Marine Biology. So, research data management is a really important tool for researchers. It helps them to effectively plan and execute research projects while also keeping in mind how to long term preserve their data.

Today we're going to discuss statistics in this topic and how statistics can promote a well-organised and successful research project. Nigel, can you give us a brief introduction to yourself and tell us what you do here at UiT?

[Nigel:]

Thanks Katie for the introduction. I have worked at the UiT since 2003. So I have been here for, you know, about 20 years. I say I'm a statistical ecologist, and that means that I have an education in biostatistics originally, but I am also a field ecologist. So I spent quite a lot of time working in the arctic, on alpine ecosystems and cold ecosystems, so being in Tromsø is really the right place.

I do quite a bit of fieldwork, so gathering my own data and having my own study designs and so on. I really liked to do everything from gathering the data to analysing the data and then publishing, of course.

[Kathleen:]

Wonderful. Thanks very much Nigel. The first thing we're going to talk about is how we might integrate statistics into the planning phase of a research project. So, at the beginning of a research project, in the planning phase, before you start collecting any data or creating any data, how do you think knowledge and skills in statistics could affect your approach to the research project?

[Nigel:]

I think the first thing is – we are doing science and when we do science, we start with a question. The first thing people should really think about is how to translate your question in a way that makes it

possible to analyse the data. That is, very often we have a relatively vague or general question and one has to really make it concrete in terms of what it will mean when you want to analyse the data. So really that's the first thing, I think, that people should start with.

And then of course, statistics is more than analysing data and what I think is really important is thinking of study design. That is what kind of information you will collect, how you will select your units. If you work in social science, what kind of people you want to involve in your surveys and so on.

Something also that I think people often forget about in terms of statistics is measurement. That is how you are going to measure what you are interested in, and thinking really of the quality of your measurements. Something that very often people forget about when starting a PhD is really, how are you going to measure things?

And then of course, the last part is really how you are going to analyse the data. And I guess we will come back to this when we think about data management plans.

[Kathleen:]

So having a very thoughtful approach to your research project is quite critical. One way we could do that, and help assist that, is by writing a very nice data management plan where you answer all those questions that you just brought up. So you're thinking about every step of the research plan, how you're going to approach data collection, how you're going to make sure the data is good quality by having standard analyses and what not.

The next topic is to think about how we might integrate statistics when we're actively doing the research. So when we're actively collecting data or generating data and we're collecting this data in order to address the hypothesis or the research question, why do you think statistics knowledge and skills are important when we're actively collecting data?

[Nigel:]

Yeah, it's really thinking of, again, the quality of the data. For example, when we collect data in social sciences or in ecology, there could be a lot of selection processes going on. That is, we have planned to acquire a given sample of data on the given population and things don't go according to the plan. And we have to read and to think carefully about how this will affect, in a way, the data we get and the results we may get at the end.

And it can be the same thing in other fields. That is, when we collect the data, really understanding the sources of biases that can affect how we are going to analyse the results. I think it's something that very often we forget when we see the final results, that there are really many choices that have been made along the process of collecting the data. And these choices are not apparent in the final product, in a way, when we publish papers. I think this is also something where the data management plan is important, because it really should aim at making these choices. This is a selection that we very often have to do, apparently.

[Kathleen:]

And then, obviously, to record these choices or this these new directions in your data collection into your data management plan. So you remember in a few months time or a few years' time when it comes to writing up these research projects into publications.

[Nigel:]

Exactly. So really to have a true tracking. These different choices that we make, so that we can integrate this idea, and other people can really understand and compare to what they did.

[Kathleen:]

Perfect, thank you so much.

So if we want to use statistics to maximise our research efficacy. Sometimes doing data collection analysis, like you just said, we have to make choices. We may find we need more data, may need to change our approach and our collection methodology, or may need to run some different standards, but this may prove very costly or logistically even impossible if you have to collect data only in a certain season or you're controlled by some external factor. How can we use our statistics to help us plan and conduct research projects to, sort of, be the most effective when conducting our research?

[Nigel:]

Ideally, one should try to, in a way, find somebody before one starts the study, to simulate data according to what we hope to get when we do the real fieldwork for example, all the surveys and so on, so that one can really assess if it will be possible to get the information that we are aiming at. It's a concern that in many fields in science, people collect too little, too few data and you get what is called technically this kind of low powered study. That is, the precision of the estimates that you get at the end will be much too low. And basically you end up with studies where … It's not that you can't answer that this is not happening. It's that in fact, you don't know what is happening. That is, that you have very little idea about if, for example, your effect is small or large or close to zero.

And I think doing this kind of preliminary work to assess, really, if we have a reasonable chance to get the answer we are after is something that people should invest more time in.

[Kathleen:]

So what you're saying is that it's really important, if you can, to sort of model what you need to do. And then when you're in the field or in the lab, whatever your method is, to collect a lot of data.

[Nigel:]

Yeah. It's a question that I have gotten hundreds of times: "How large should my sample be", and I always answer… Yeah, the obvious answer, as large as you can make it. But I think it's important that people don't start studies where they have very, very little chance to get a sample which in fact will be large enough to give a good answer.

So I think it's a waste of time on the research, of money, and especially for the PhD candidates. It's really not very nice when you start a study and at the end you realise that you could not answer your question.

[Kathleen:]

So you want to decrease the risk by thinking about what you need to do and if it's feasible before you do it. That sounds like a good idea.

So, in your opinion, when we talk about statistics and research data management, when you're analysing the data, what would be, in your opinion, some pitfalls in data analysis that result from a lack of knowledge in statistics? So, maybe you didn't write a thorough data management plan. Maybe you didn't think about the sample size that's required in order to thoroughly evaluate

whatever your research question or your hypothesis is. So what are the common pitfalls in data analysis do you think – that you encounter?

[Nigel:]

I think there are two aspects here. You have the more technical aspects. That is, you choose the wrong method to analyse the data. You ignore very important sources of biases, for example. And that's clearly … that's always a technical issue. It is a very important one. When I say it is technical, it doesn't mean that you should ignore them, but that's the first aspect, that is ignoring this kind of problem.

The other aspect is that when scientists analysed data, as we discussed, they have to make many choices. Some people call it degrees of freedom in the sense that you have to choose which method you use, which part of the sample you use.

A famous statistician, Andrew Gelman, talked about the "garden of forking paths". That is really, you have this kind of decision: You could go left, you could go right, you go left and you go right – these are mainly decisions you have to make. And if you ignore it in a way, in the final product, that what you have is a result of all these choices. You might get a very biased view of, in a way, the results, in fact, that you could draw from your study.

So ignoring all this selection … and that has been shown now in many fields of sciences. This kind of selection process has led to results that we cannot reproduce. They reflect more the choices made by the scientists than really what is happening in the systems that people study.

[Kathleen:]

That is very interesting that the bias can exert such control on the results. And reproducibility is a big issue. And you always want it to be reproducible, so people can use your work and integrate it into their research and get citations, obviously, which is very important, especially when you're an early career researcher.

The next thing I want to talk about is statistics and the Open Science movement. If you are familiar with the Open Science movement, it ensures that not only do you publish your research data, but also any notes or any supporting information that goes along with the holistic research project. So this can include methodology, maybe those choices that you just discussed where you took the forks in the path. It can also be any statistical methods that you use to process, clean, and analyse data. So the movement towards Open Science, what do you think? How does that intersect with your expectations of people's statistical knowledge? Do you think having Open Science will improve researchers' level of statistical knowledge?

[Nigel:]

Well, I think the Open Science movement or whatever you call it – I think it's extremely important. And I'm really convinced that this will really improve how statistical methods are used in science in general. Because, as we just discussed, avoiding that these choices that are made are hidden in the final papers that people publish. It is something we in all groups, for example, we are setting up in what we called an observatory.

So, take COAT – Climate ecological Observatory of Arctic Tundra. It's a bit of a mouthful, but it's an observatory, and we make available both the data that we acquire in the field, and also all the analytical steps that we use to produce papers or to make figures that we make available for people. And we make available all the different choices – we have what we call scripts where how the

analyses have been done is really explicit. So people can assess everything we did, from the fieldwork to the final product.

There is no right way, often, to do things. It's very important that people can compare and say, okay, what would happen if I had done it differently? Fine. That's really something important. And, I think, the Open Science movement from having access to data, access to analytical choices, and then really how it links to the questions we ask. This is really, I think, the way forward and I do hope that the new generation will be much more, in a way of, going forward to do this.

[Kathleen:]

I think Open Science is really a powerful movement and I think it's important to know people shouldn't be afraid of open science. It's really a good thing. As you said, there's no right or wrong way, but it's important that you show your path, sort of the how the data was generated, how it was processed. And that helps you understand the interpretations of the study and lets you also assess your interpretations of your own data.

[Nigel:]

I think you had the right words, that some people are really afraid of it. And I think especially for PhD candidates, you should really not be afraid of Open Science …

[Kathleen:]

No one's going to be angry at you for showing your methodology.

[Nigel:]

And, I think sharing your data, the experience we have had, in fact – it means that people are using them. And this is really a way to get to know other people or to get your work to be known instead of having your data in a way hidden. Very often they get forgotten, right. So don't be afraid of Open Science!

[Kathleen:]

Publish your data, publish your methodology. It's good.

Okay, lastly, let's talk specifically about the application of statistics for students or candidates who are doing their PhD. So when you're doing your PhD, it's not just you alone. Usually you're included in a team of skilled researchers. So your supervisor, the professor, other students, maybe some postdoctoral researchers, and maybe you're all conducting sort of separate parts of a more holistic research project. So, usually when you conduct any research project, you have to do some statistical analysis of your data. But in a larger team, you know, perhaps there's a feeling students can maybe ask other people to help them with that sort of statistical treatment. But how much statistical knowledge do you think PhD candidates should know, should have with them during their degree?

[Nigel:]

That's a difficult question because of course, it depends a lot on the field that you work in. But I think all PhD candidates should have a really basic understanding of the statistical principles. So, for example, when we talk about statistical evidence, what does it mean? So, when they talk to a statistician, and they try to understand: “Do I have in my data, in my research project … Do I have good evidence that this or that is happening?” So I think there is there is really some basic understanding of principles that should be common to all PhD candidates.

And then, of course, depending on the field, you have to have some understanding of the techniques, not the technical aspects, but real understanding of why they are used in that way, and how it relates to, as we discussed before, measurement, quality, or sources of biases in the way you have selected your samples and so on.

So at least that they understand the different choices that are made when analysing the data and writing papers so that they can… When you are co-author of any paper, people have a tendency to forget that you stand for the whole paper. You don't stand only for your small bits in the paper. So you have to understand the whole process from acquiring the data to writing the paper.

[Kathleen:]

Well, that's been really interesting. I think we've learned some things today. It is to think thoughtfully about your research project and plan as much as you can, think about how you're going to collect the data. Make sure you have enough data. And then also when you're doing the research project, to actively record all the steps and choices that you make in the data management plan or another platform, and really prepare yourself to share your data openly, you know, with publication, but also with, with Open Science and making sure that all researchers can see how you handled your data.

So thanks so much, Nigel, today for joining us. It's been really instructive and I hope everyone learned something from this talk today.

[Nigel:]

Thank you, Katie.