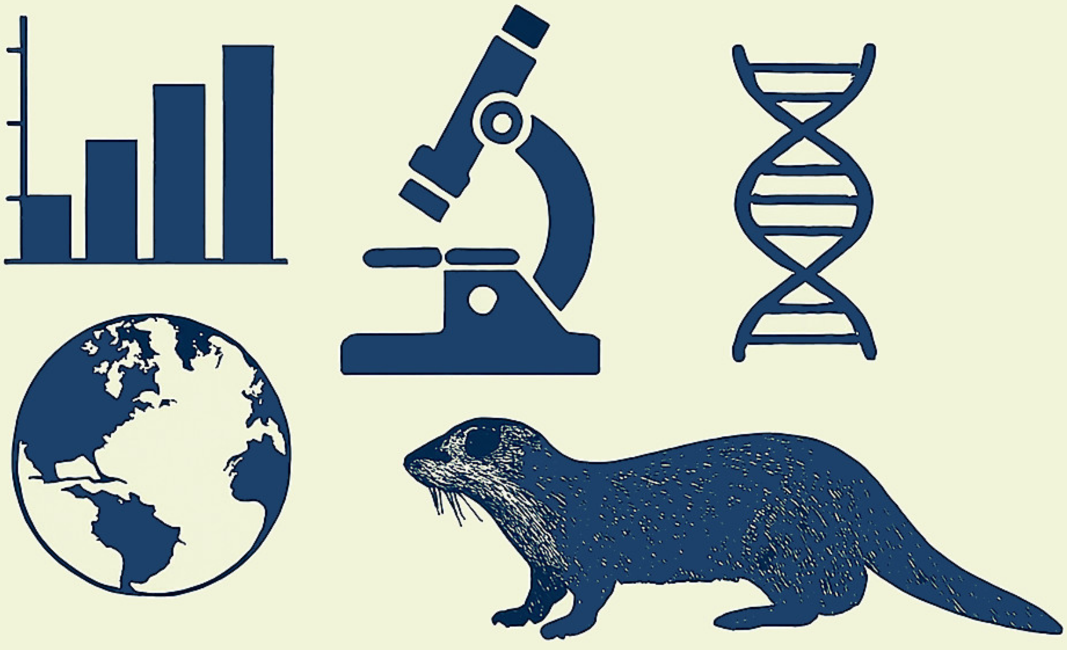
NES2303 Module Handbook

Roy Sanderson

2025-08-08

NES2303 EXPERIMENTAL DESIGN & STATISTICS

MODULE HANDBOOK BIOLOGY AND ZOOLOGY BSc



Roy Sanderson

Newcastle University

Table of contents

Welcome

This is the 2025-26 edition of the module handbook for NES2303 Experimental Design and Analysis. The handbook contains information about the module, including its aims, learning outcomes, assessment details, and schedule. You can access it via either the HTML or PDF format. If using the HTML format, you can navigate through the handbook using the sidebar on the left. The PDF version is available for download from Canvas.

I recognise that as Biology or Zoology undergraduates, you may not have a strong background in statistics. Therefore, I have designed the module to be accessible to all students, regardless of their prior knowledge. The module will provide you with the necessary statistical skills to design and analyse experiments effectively. Each week you will be expected to undertake a small amount of reading, view a short video and use an interactive website to ensure you understand the concepts. The module will also include practical sessions where you will apply the statistical techniques learned in lectures to real data sets.

Pre-requisites

Most of you will not have done A-levels or Scottish Highers in Mathematics or Statistics, and are probably uninterested in the subject! I've therefore designed the module to be accessible to all students, regardless of their prior knowledge. The module will provide you with the necessary statistical skills to design and analyse experiments effectively. The only key pre-requisite is that you know how to calculate an average (mean)!

1 Introduction

This document provides you with the aims and rationale behind the course, explains its mode of delivery, staff roles, and forms of assessment. More importantly, it explains the underlying philosophy used to teach experimental design and statistics, which will probably differ from most that you may have previously encountered, or seen in textbooks. A common approach is to provide students with a plethora of different statistical tests that they may or may not need to apply to their data, depending on how they collected their data, the aims and objectives of their experiment etc. This can be very confusing, resulting in bafflement (that was my experience as an undergraduate) with the result that some UG textbooks provide complex flowcharts to help students select the appropriate test.

A second challenge is the type of software used to undertake statistics. I am (just) old enough to remember doing some basic statistical tests with a calculator which would take a whole afternoon, even with a relatively small sample size. The arrival of computers to take the tedium out of this was a great relief. Most software commonly used for statistics, such as Microsoft Excel, Minitab and SPSS are menu-driven, making it easy to get started with data analysis and graph display. However, whilst their learning curve is initially very shallow, it soon becomes steep. Biological data is inherently complex, and often the more advanced analyses or data visualisations are difficult or impossible to undertake. An additional problem is that you have no record of what you have done. You will be taking this course in October to December of your Stage 2, but many of the skills you will learn will not be practiced in full until you are doing Stage 3 courses and projects. If you do not need to do much statistics after completing this NES2303 course, will you remember what you need in 18 months' time for your Final Year Project or other experiments/surveys? If you have menu-driven software, it is easy to learn, but easy to forget. In contrast, we will use a coding language that is now common in Biology and Zoology, called R. It is (slightly) harder to learn than menu-driven software, but you don't forget it, even 12 months later! Many careers in biology and zoology, both laboratory and field-based, expect applicants to know how to use R. We have therefore developed this course to give you the best employability skills for your CV as well.

Finally, this is the third year in which I have taught on this course and it is updated each year to try and improve it based on your feedback. The main changes this year is to add an additional initial practical to help you learn how to configure and 'navigate' the software we'll use. This will hopefully ensure that in the subsequent practicals you can focus on data science for biologists and zoologists, and not worry about IT. Passive learning in lectures is not particularly useful for quantitative biologists and zoologists like yourselves, and hands-on experience is more useful, so additional practicals are always valuable. If you find yourself making excellent progress, do not feel obliged to attend every practical; however repeating some exercises will embed your knowledge. Also **understanding** is key: getting the 'right' answer from an analysis is **not** sufficient. You need to know how to interpret its biological meaning.

Please provide feedback (positive or negative) as you progress through NES2303 either directly to me as course leader, or to your student representative; name(s) of student reps will be shared as soon as available.

2 Introduction to Biostatistical Modelling

To some extent the above heading would be a better title for this course than “Experimental Design and Statistics”. Some of you, especially those focussed on cell and molecular biology and / or working in the laboratories are likely to be conducting formal experiments. Others, especially those collecting ecological / environmental data, might be undertaking more survey-work, but this will still be formalised such that it can be analysed and inferences drawn. There are a wide range of experimental and survey designs available, but what they all have in common is the aim to **separate the signal from the noise**. By “signal” we are talking about questions such as ’Does this antibiotic kill MRSA bacteria?” or “Does low-fertiliser input result in more insect species?” etc. In other words, the biological question that you really want to address and answer.

Biological data are inherently noisy: if you repeat the same experiment twice you will obtain similar, but not identical results. Some biological systems, such as ecological ones, are noisier than laboratory ones, but even in the latter sample contamination, temperature variations etc. will result in slightly different results. All good experimental / survey designs have the aim of allowing you to work out where the noise in your system is coming from. If you can reduce, or understand, the noise you are more likely to be able to look at the signal which is what you are really interested in.

2.1 Cause and effect

When you undertake an experiment or survey, you will often be interested in knowing what effect a particular treatment, management regime etc. has on an outcome. For example:

- Is bird Shannon diversity affected by willow coppice management?
- How long should people self-isolate for if tested positive for Covid19?
- Is % photosynthetic efficiency affected by iron availability in the soil?
- Is wheat yield increased by both Azole and SDHI fungicides at different doses?
- What dose of insecticide kills 50% of the aphids?

For some of these questions it is easy to see how you might design an experiment or survey to answer your question, whereas for others it is harder. What they all haven in common is an underlying assumption about **cause** and **effect**, which are often referred to as **independent** and **dependent** variables, or **explanatory** and **response** variables. So for the above examples, we can view them as:

Response (effect)	Explanatory (cause)
Bird diversity	Coppice management occurs (yes/no)
Isolation time (days)	Covid test (positive / negative)
Photosynthesis (%)	Fe concentration (mg/l)
Wheat yield (t/ha)	Azole (none / low / high) and SDHI (none / low / high)
Aphids (dead/alive)	Insecticide dose (mg/l)

In some of these examples your response (dependent) variable is continuous, with numbers that contain decimal points (bird diversity, wheat yield), in some the response is a whole number (3 days isolation, 7 days, 12 days etc.), and others a value between 0 and 100 (% peak photosynthesis). Likewise sometimes the explanatory (independent) variables have two categories (coppice, covid test), some multiple categories (Azole and SDHI) whilst others are continuous (Fe concentration). Three of the examples (bird, isolation and photosynthesis) only have one explanatory variable, whereas the fourth (wheat) has two explanatory variables. In the last example, the response variable is binary (dead or alive aphids).

Irrespective of the exact type of data, all the above examples have the same pattern of:

$$\boxed{\textbf{Response variable}} = \boxed{\textbf{Explanatory variable(s)}}$$

in other words we are trying to determine whether the explanatory variable(s) are or are not changing the response variable. However, we have already stated that biological data is noisy, and the noisier the data the harder it becomes to determine cause and effect. Therefore, we need to expand our previous equation to:

$$\boxed{\textbf{Response variable}} = \boxed{\textbf{Explanatory variable(s)}} + \boxed{\textbf{Noise}}$$

We don't know what causes this noise, although we can often have a good guess. In this course we will consider simple experimental designs that help explain some of the noise, allowing you to focus more on what your explanatory variables are doing. This noise is more usually referred to as "Residual Error" or simply "Error" : this is not to imply that you have done something wrong, merely that we do not know where the variation has come from, and we need to account for it.

The majority of the statistical models that we create in this course will have this syntax of a response variable and one or more explanatory variables, with residual error that is quantified. Formally, this can be written as:

$$Y = X + \epsilon$$

where

- Y is your response variable
- X is one or more explanatory variables
- ϵ is the Greek letter Epsilon, which by convention is used to represent the 'noise' or 'residual error' or 'variability' in your data.

2.2 Correlation does not indicate causation

It is very easy to be seduced by statistics, and show a nice pattern between different variables that suggests something is happening when in reality it is not. For example, the number of diagnosed brain tumours in different States in the USA is strongly correlated with the number of mobile phone users in each State. Does that imply that you should throw your Android and iPhones in the bin? No. Think about it for a moment: what matters is the number of brain tumours per capita. Once you take the size of the population of each State into account the correlation disappears. No need to panic!

This is a trivial example, but I include it simply to emphasise that you are first and foremost biologists. You must always ask whether the question you are asking makes biological sense, and not blindly accept some numbers generated by a computer.

3 Defining goals

One approach we will try and use as much as possible is to think in terms of the **goal** that you are trying to score. By a goal, we could be referring to a particular graph or plot to visualise your data, a way of tabulating a summary of your data, or undertaking a statistical model. Broadly, your thinking can be along the lines of:

$$\boxed{\text{goal}}\ \textbf{(}\ \boxed{\textbf{y}} \sim \boxed{\textbf{x}}\textbf{, data} = \boxed{\textbf{mydata}}\textbf{, ...)}$$

Let's decode this little diagram:

- **goal** This can be a plot, such as a barchart, scatterplot, boxplot etc. It could be a statistical test, it might be a simple summary statistic, for example an average (`mean`), or standard deviation (`sd`), or it might be a statistical model, such as a linear model (`lm`) or generalised linear model (`glm`). All these will be introduced to you during the course.
- **y ~ x** This represents the $Y = X$ in the equation earlier. For some graphs or simple statistics you might only need the Y component. For example, if you the average overall bird diversity across all your sites, you would just provide the Y data of bird diversity at each wood. If you want the average broken down for both coppiced and uncoppiced woodlands, you would provide both the Y and X data. Note that you do not actually need to put $+\epsilon$ as this is calculated for you where needed.
- **data = mydata** Obviously you will be working with data. Often you need to *pre-process* your data before you can achieve your goal, and we will teach you simple tricks to do this.
- **...** This represents additional options. For graphs, you might want to change colours, titles etc. In statistical models, some assumptions may need to be changed when you work with deal with count data (e.g. days after Covid19), bounded data (e.g. % photosynthesis efficiency), or presence / absence data (e.g. dead or alive aphids).

4 Signal and noise

4.1 Why do I need to know this?

One student last year asked me "Why do I need to know this?" during one of the practicals. Basically, the student didn't see any reason why data analysis was needed to show whether a laboratory treatment was having an effect on the results. The same question could be raised about a field survey of birds or mammals, and it is a fair question to ask.

4.2 Visualise the data

My advice for a good starting point is not to begin with data analysis, and instead to visualise your results: this gives you a good initial understanding of what's happening. Here are two possible scenarios, given that both Biology and Zoology students do the module, and some of you will be focussed on laboratory experiments, and others amongst you field-based studies:

- **Laboratory example** You've just done a laboratory experiment into cell growth for a bioreactor. Your response is methane production and you have one explanatory variable with two 'levels', a conventional feedstock (control), and a new bioreactor feedstock;
- **Field example** You've just done a field survey of birds. Your response is the number of birds seen in a given area, and you have one explanatory variable with two 'levels', a conventional habitat management (control), and a new wildflower management technique.

For both examples you can visualise the data using a "boxplot" which shows the median, interquartile range, and outliers. The boxplot is a good way to visualise the data because it shows the distribution of the data, and allows you to see if there are any outliers. In the boxplot below the horizontal (x) axis shows the explanatory variable (bioreactor feedstock or field management), and the vertical (y) axis shows the response variable (methane production or number of birds seen).

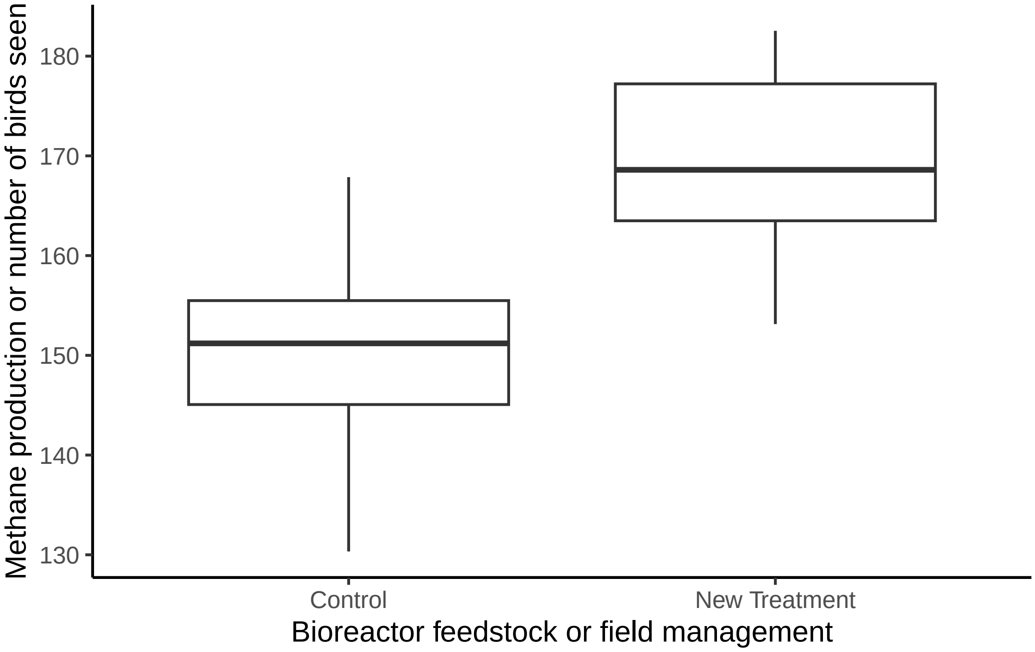


Figure 4.1: Boxplot of lab experiment or field survey

Well that looks fairly obvious: your new bioreactor feedstock or wildflower field management treatment is having the positive effect you hoped for. The median, shown by the horizontal line in each box, is roughly 150 for the control, and roughly 170 for your new treatment. Yes, there is a bit of noise in the data, but the differences are still clear. You can think of this example as:

$$\boxed{\text{Methane or birds}} = \boxed{\text{Treatment}} + \boxed{\text{noise}}$$

Your response is of course methane production or bird diversity. The **signal** from your treatment (feedstock type or field management) on your response is much more important than the random **noise**, as shown by the bigger box, so it is still obvious what is going on in your graph. Thus, when measuring the **signal to noise ratio** the signal from your treatment is much bigger than the unexplained random noise, so you will have a big signal to noise ratio. This is a good situation to be in, and you can be confident that your treatment is having an effect.

4.3 Noisier data

But what if you had obtained a graph like this?

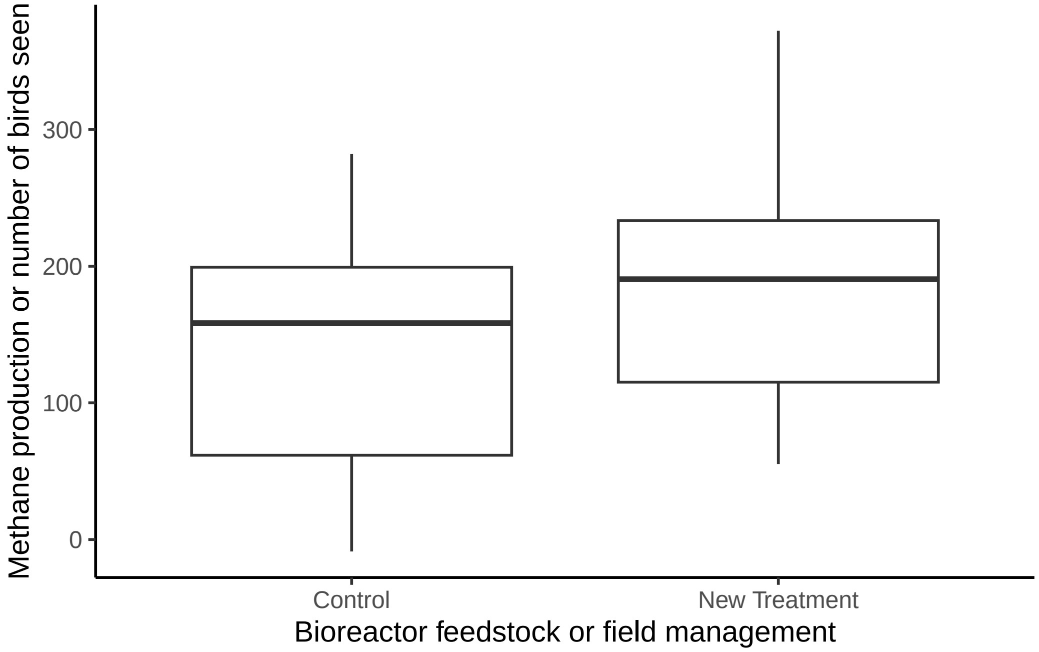


Figure 4.2: Boxplot of lab experiment or field survey with more noise

The control is still at about 150, and the new treatment still around 170. But the random noise is much bigger, as shown by the much bigger boxes and whiskers. Indeed, on the first graph the vertical axis only goes from 130 to 180, whereas here it goes from 0 to 300 to encompass the much noisier data. You can think of this as:

$$\boxed{\text{Methane or birds}} = \boxed{\text{Treatment}} + \boxed{\text{noise}}$$

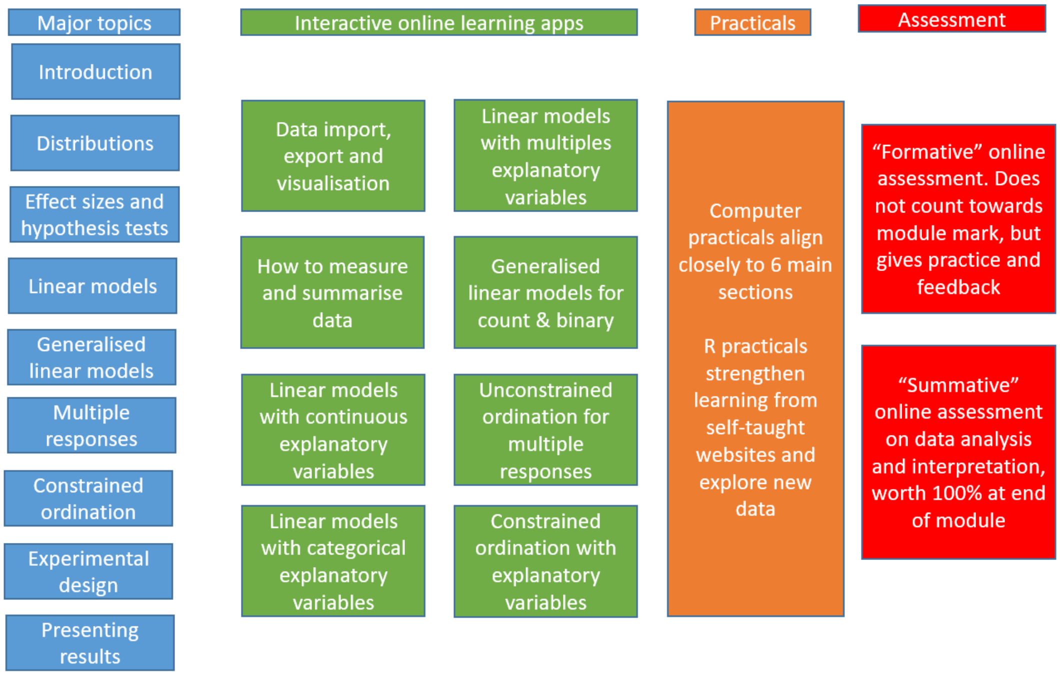
Whilst the **signal** from your treatment (feedstock type or field management) still has a big box, the random **noise** has become much worse, making it harder to understand. Thus, whilst a visual assessment of your data is useful when you have clear differences, you need more powerful methods when the data are noisy. These formal methods of data analysis are what you'll learn in this module, and they'll allow you to determine whether the differences you see in your data are statistically significant, or whether they could have occurred by chance. Most of these methods work by assessing the **signal to noise ratio** in your data, to help you decide if the differences are real. In this second example, the signal to noise ratio is much smaller, meaning that the differences you see could be due to random variation rather than a real effect of your treatment.

I've used an example here with a categorical explanatory variable (feedstock or field management), but the same principles apply to continuous explanatory variables, such as temperature or time. The key point is that you need to assess the signal to noise ratio in your data to determine whether the differences you see are real or just due to random variation.

5 NES2303 Course structure

There will be a couple of introductory lectures from Roy Sanderson at the start of the course, but the focus will be on practical sessions. If for health or other reasons you are not able to attend the computer classes on Campus, then you should still be able to complete the course without problems. **We will use Microsoft Teams** to answer questions, as students can **help each other** more easily, and staff or demonstrators can **respond more quickly**.

The following diagram summarises the overall course structure


Major topics
Introduction
Distributions
Effect sizes and hypothesis tests
Linear models
Generalised linear models
Multiple responses
Constrained ordination
Experimental design
Presenting results
Interactive online learning apps
Data import, export and visualisation
Linear models with multiples explanatory variables
How to measure and summarise data
Generalised linear models for count & binary
Linear models with continuous explanatory variables
Unconstrained ordination for multiple responses
Linear models with categorical explanatory variables
Constrained ordination with explanatory variables
Practicals
Computer practicals align closely to 6 main sections
R practicals strengthen learning from self-taught websites and explore new data
Assessment
“Formative” online assessment. Does not count towards module mark, but gives practice and feedback
“Summative” online assessment on data analysis and interpretation, worth 100% at end of module

- Blue - broad topics shown on the left
- Green - 6 major sections for the NES2303 course (see below)
- Orange - Computer practical sessions; these map onto the 6 topics
- Red - Formative and Summative assessments

Teaching staff and many demonstrators will be from the Modelling Evidence and Policy Research Group (MEP) https://www.ncl.ac.uk/nes/our-research/biology/modelling-evidence-policy/ within SNES. The group includes Biologists and Zoologists working in many different fields, from microbiology through to field zoology, but all these areas require solid data analysis and interpretation. The formative and summative tests will be hosted on the Canvas Virtual Learning Environment, but will require you to analyse and interpret data separately. The tests will take the form of short-answer / multiple choice / numeric answers.

5.1 Lecture structure

An Introductory lecture from Roy Sanderson (Module Leader) in Week 4, ‘quantitative foundations’ lecture in Week 5, and a final informal Question and Answer session in Week 14 to answer any outstanding queries on your understanding of course content, and help you prepare for the Summative Assessment in January.

5.2 Interactive websites

We have created **interactive websites** for this course, for you to learn and practice ideas **before and after** the practicals. These websites include simple quizzes to test your understanding. In the practicals you will be able to explore the example data shown on the websites in more depth, and hone your skills using new data. If you study the interactive websites **before** attending the relevant practical, you will find the latter much simpler, and gain more from it.

5.3 Computer classes

These are scheduled for seven weeks (Week 4 to Week 11 - Week 10 is ‘Enrichment Week’ with no teaching) in the main Herschel Building cluster room (Fell) on the first floor.

5.4 Ten-minute videos (TMV)

I have prepared a series of short (5 to 10 minutes) animated videos to explain in simple, non-technical terms some quantitative biology concepts. These have been gradually introduced to the course over the last 2 years, and students have found them helpful. They deliberately avoid complex maths, and build up key concepts sequentially to aid your understanding.

5.5 Cheatsheet

There is a lot of new information to absorb in this module so I've prepared a 2-page quick-reference "cheatsheet" with the main statistical analyses. This will probably be incomprehensible to you at the start of this course (don't panic!), but should soon be a valuable resource to help you as you progress and develop your understanding.

5.6 Managing your workload

This is a 10-credit module so assumes you will devote 100 hours of study time. There are 7 practicals and two lectures, so you would theoretically do 86 hours of self-directed study. In practice, you won't need 86 hours (phew!) but please do make use of the extensive online resources on Canvas, especially the Ten-Minute Videos and Interactive Websites. I have provided a page on Canvas to help you plan and manage the workload for NES2303, with a week-by-week diary throughout Semester One.

6 Detailed structure of this course

Materials are all provided on Canvas in the following sections (Canvas 'modules'). Each contains a set of presentations, interactive websites, a related practical, and five include an online Canvas test. You will not be able to access the practicals (or tests) until you have reviewed the interactive websites. The area on Canvas called "Timetable - a guide to managing your workload for NES2303" gives exact times, dates, venues for all activities.

6.1 Section Zero - Orientation practical

This will simply be to help you setup R/RStudio on an NUIT PC, understand files, folders, filetypes, R packages etc.

6.2 Section One - Summarising and displaying data

This will cover simple statistics, including measures of central tendency and variation in your data.

- Statistical power and hypothesis tests will be covered.
- How do you know if two treatments are different?
- How do you know if one variable is correlated with an other?
- What are p-values, what are their uses, and their abuses.
- What are effect sizes?

This section will also provide guidance on good methods to visualise your data. Data visualisation is essential both prior to data analysis, as well as to communicate your results.

6.2.1 Interactive websites

Introducing R and RStudio Data summaries and visualisation Understanding measures of central tendancy and variation

6.2.2 Practical

Practice examples from interactive websites

6.3 Section Two - Linear models and ANOVA

Learn about the concept of linear models, the R `lm` and `anova` functions, and how to interpret them. Understand the assumptions behind linear models, and look at examples where the explanatory variables are continuous or categorical.

6.3.1 Interactive websites

Linear models with continuous explanatory variables Linear models with categorical explanatory variables

6.3.2 Practical

Practice examples from interactive webistes

6.4 Section Three - Dealing with multiple explanatory variables

This section explores situations where you might have multiple explanatory variables. These might not be independent in their effects, so we explore interactions between them, and how to understand them. We also consider ways in which you can improve your experimental design by blocking, and look at what to do if things go wrong and you lose some data. We also consider how ANOVA operates - what are the basic concepts of an ANOVA table, and how are we able to study treatment effects by looking at variances?

6.4.1 Interactive websites

Multiple explanatory variables, interactions, blocks and missing data Understanding how sums of squares are used in linear models

6.4.2 Practical

Practice examples from interactive webistes

6.5 Section Four - Generalised linear models

The assumptions behind linear models, accessed through the `lm()` function in R, are breached when our response data consists of counts, proportions, or yes / no, true / false, dead / alive. We can extend linear models by generalising them using the `glm()` function. This sections shows you the two key statistical distributions relevant to these types of data, the Poisson and Binomial distributions, and how they relate the the traditional Gaussian (normal) distribution used in linear models.

6.5.1 Interactive website

Generalised linear models

6.5.2 Practical

Practice examples from interactive webistes

6.6 Section Five - Multiple response data; unconstrained ordination

Sometimes you will not have a single response measurement, but might have 10, 20, 50, or hundreds. If you are dealing with genomic or bioinformatic data, you might have thousands of gene sequences that you want to understand. If you do surveys of plants, birds or insects you may have large numbers of species. It is not correct to do hundreds of separate (g)lms on these data, but multivariate techniques (MVT) provide an effective way to help you interpret them. In this section we introduce you to unconstrained ordination, which allows you to visualise relationships amongst your data. You can also make simple comparisons with potential explanatory variables.

6.6.1 Interactive website

Unconstrained ordination

6.6.2 Practical

Practice examples from interactive webistes

6.7 Section Six - Multiple response data; constrained ordination

Here we demonstrate more advanced methods of using your explanatory data when you are working with multiple response variables. You will learn about permutation tests, which allow you to 'simulate' conventional ANOVA tests, such that you can infer the importance of your explanatory variables. We also discuss the best ways of visualising the results of these analyses. Advice from Roy Sanderson (SNES) advice on report-writing, planning research projects in Stage 3 etc.

6.7.1 Interactive website

Constrained ordination

6.7.2 Practical

Practice examples from interactive websites

6.8 Online Appendix on "Other Tests"

Other tests, such as t.tests, paired t.tests, rank-based tests, Chi-squared tests. These all fit into the (generalised) linear model framework. Comparison of the standard (base) R commands and the equivalent via `lm()` or `glm()`

6.8.1 Interactive website

Appendix - other tests

7 Assessments

There will be two assessments, both delivered through Canvas, to test your skills and understanding. The format will be multiple choice and numeric answers (based on datasets for you to download and analyse)

- “formative” assessment, probably mid- to late-November. This does not count towards your mark for NES2303, but will give you practice in using the Canvas system for online assessments. Probably released in late November.
- “summative” assessment, worth 100% of the course marks. This can be completed in your own time, during the standard Semester One assessment period. Again, it is a Canvas-based test.

8 Software required

You will use free software called R/RStudio in this course. R was originally designed as a free version as a purely statistical commercial package (called S), and over the last 20 years has become the most popular analytical modelling software used by biologists in academia and industry. A huge number of free online training resources are available, and people have written what are known as 'packages' to provide extra functionality. These include data analysis in bioinformatics, ecology, zoology, animal tracking, genomics etc. We will make extensive use of two R packages: `mosaic` and `ggformula`. The `mosaic` package eases the transition for university students to use a code-based data science system like RStudio. `gg_formula` allows you to build complex graphs, using a series of simple commands.

The R software, and `mosaic` and `ggformula` packages, are pre-installed on University NUIT cluster-rooms. They can also be used on your own PC or MacBook but due to technical differences between makes / models we will **not** provide you with any support with these.

8.1 Why R rather than Excel, Minitab or SPSS?

Commercial packages such as Excel, Minitab and SPSS have a "shallow" initial learning curve, as they are based on familiar point-and-click menu-driven systems. R is much tougher initially, as you have to type in commands to, for example, calculage a mean or standard deviation, display a graph etc. However, R has several big advantages:

- you can save your set of commands in an ordinary text file. When you want to repeat or modify any analyses or graphs 6 months later, simply go back and make minor edits.
- Your analysis is therefore "reproducible".
- In Excel, Minitab etc. you need to remember exactly which check-box and menus to click: if you are like me you will have forgotten;
- even moderately complex analyses, including some that you will do in NES2303, are impossible in Excel etc;
- Government bodies (including NHS, Natural England), media organisations including BBC, Financial Times, The Guardian), Research Institutes (including Centre for Ecology and Hydrology, Francis Crick Institute), all universities (internationally and UK) use to analyse and visualise R

- **Increases your employability** It looks great on your CV after you graduate as it demonstrates you are also a data scientist!

9 Stage 3 Research Projects

Many of the skills you learn in NES2303 will be essential for your Stage 3 research or literature review projects. This is not just your computer and analytical skills, but also *critical thinking* and *problem solving*. Indeed, a not uncommon conversation between academic staff and students when scoping their projects is along the lines of:

> **Student**: I'd like to do a project on badgers / DNA / bacteria / phylogeny (*delete as applicable*)
>
> **Staff**: What exactly do you want to find out about badgers / DNA / bacteria / phylogeny?
>
> **Student**: I'm really interested in them!
>
> **Staff**: But what is the research question?

As you progress through the rest of Stage 2, and especially in Stage 3, you will find that you are designing more experiments and surveys yourself, rather than being given a laboratory manual or field guide to follow like a cookery recipe. This, of course, is what real scientific research is like: there isn't an absent-minded academic stood nearby to tell you exactly what to do. You need to be able to deconstruct what you plan to do, in order to create a sensible, practical experiment or survey, that provide data of sufficient quality for you to interpret.

When Stage 3 students approach me wanting to do a research project under my supervision, a common starting point is them saying something along the lines of "I want to do a research project on butterflies", or "I'm interested in pollinator conservation", or "I want to follow a career in mammal conservation". These might be valid statements, but my immediate response is to ask "What is the question?", in other words, "What is it specifically that you want to determine?". So, taking those examples, it would be better to have something along the lines of:

- Are the numbers of butterflies higher in roadside verges where the vegetation is cut monthly compared to once per year?
- Do bumblebees occur more frequently near fields of mass-flowering crops such as oilseed rape?
- Is the frequency of small mammals detected in camera traps associated with supplementary feed for game birds?

It is only when the students' initial statements are re-phrased into questions that we can start to sketch out possible routes by which you can develop formal scientific hypotheses to test with your statistical model, develop a sampling strategy, do the experiment / survey etc. Another common mistake is to **confuse aims and objectives**. The three bullet points above are still quite general, and might be considered as overall aims. Objectives are much more specific, and might encompass pilot studies to test methodology, or detailed studies of other variables. For example, in the butterfly study, two obvious objectives might be to see if the numbers of species of plants or the numbers of flowers m^2differed according to cutting regime.

Therefore, when planning your research, a good strategy is to:

- Scope subject area, including literature review
- Define overall aim and the hypothesis / hypotheses you want to test
- Define specific objectives
- Pilot study (ideally) or use literature. This is useful to give you an understanding of the 'effect size' (more later) you are trying to detect, and thus the numbers or frequency of replicate samples that might be needed
- Design your experiment or survey.
- Undertake your experiment or survey, keeping careful laboratory or field records.
- Enter your data into a computer (often Excel, but if using maps specialist GIS software may be better).
- Import your data into R. If needed pre-process it to make it suitable for graphing or analysis. This **always** takes longer than you expect. Think your analyses **before** you data collection to minimise pre-processing
- Create summary tables, charts, figures. Get a broad understanding of your data.
- Statistical model of your data. Formal tests of hypotheses.
- Check assumptions of your model. All statistical models make certain assumptions about the input data which need to be checked;
- Interpret model; if necessary modify and re-run different models and compare them. Remember: **"All models are wrong, but some are useful!"**
- Write up your report (see tips below)

10 The last part of any quantitative modelling

Depending on your preferences, this is either the easiest part, or the most difficult! Analysing and interpretating your data is not the whole story: you have to write it up as a report. Personally, I often find this the most difficult step, as I easily get writer's block, especially if I start up Microsoft Word, type in the word **Introduction** and look at a blank, white screen. Most reports you will write on this and other courses, especially Stage 3, and for your Stage 3 project, will probably follow the format of

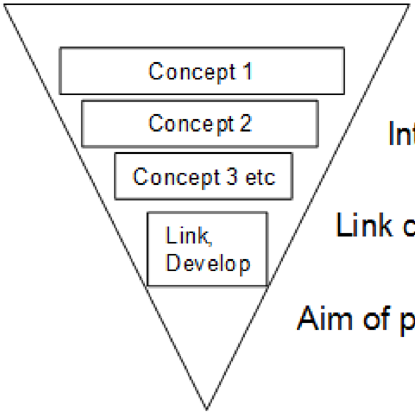
- Abstract
- Introduction
- Methods
- Results
- Discussion
- References

That might be how it is written, but the order I often find myself writing scientific papers is along the lines of:

1. Methods (I know what I did, so this is easy to write. Phew! One section written…)
2. Results (I know what I've obtained from my analyses, and have created various graphs and tables to support them. Making progress, two sections written…)
3. Discussion (now I can go ahead and interpret my results in the context of existing scientific literature, discuss where it accords or contrasts with previous research, and explore possibilities for further research…)
4. Introduction (given that the rest has now been written, this is much easier to write as I know what I'm about to introduce.)
5. References (of course, before you even begin your experiments / surveys, you need to have a good knowledge of the literature. I recommend you also use bibliographic software, such as Zotero, which automatically creates references for you)
6. Abstract (a good abstract is suprisingly difficult to write. It needs to be concise, precises, and quickly communicate your findings. A good title is of course also essential.)

Many scientists use the “hourglass” model as a way of thinking about how to structure scientific papers which you may have come across in Stage 1:

Introduction



Broad vision with respect to concepts and ideas

Introduce basic concept - background details

Link concepts together- build-up to aim

Aim of project - hypotheses, questions, predictions

Methods

eg
- Study site
- Data collection
- Statistics

Group methods into sub-sections:
eg: Where/what with
Data collection protocols
Data collation and statistical analysis

Results

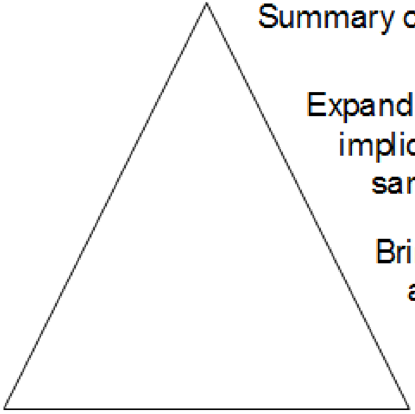
Distribution of x on study site

Behaviour of x

How x responded to y

Interpret each experiment/ set of observations, and state what conclusions can be drawn. Use sections if appropriate. Refer to all results presented. This section should have a clear narrative taking the reader through the work.

Discussion



Summary of main conclusions from each results section

Expand on each main conclusion in turn - what implications are, whether other studies show same or different thing, how novel, etc.

Bring main conclusions together- what are advances in knowledge, what new ideas may result

Broad vision on where we go from here

References

Format exactly as instructed

I also particularly recommend you look at this post which describes the **5 key paragraphs of any scientific paper**. It is written by an ecologist, but is actually applicable to any discipline. The same site also provides some useful tricks for clear writing.

11 Final remarks: how to do well

11.1 Getting the most out of this module

First, don't be afraid of the data, basic statistics or R code. You'll learn some really useful critical thinking and transferable skills, but the module is designed to allow you to go at your own pace. Do not panic if you don't understand everything at first, as the module has a steep learning curve. There is plenty of demonstrator support. Based on previous years, I'd also add:

- **Don't copy and paste** It's very tempting to copy and paste R code from my instructions into your workspace to complete the practicals quickly. You'll finish the 2 hour practical session quickly, but will learn less. So I've been really nasty and have included some deliberate typos in the code in later practicals in the module to reduce the temptation to do this!
- **Prepare for the practicals** The practicals are designed to be completed in 2 hours, but you will need to prepare for them. This means watching the "Ten Minute Videos" (TMV) associated with each practical beforehand. These videos are designed to be user-friendly and accessible
- **Use the interactive websites** Some concepts are difficult to show in a PowerPoint or static screen, so the interactive websites for each website will help you with this. Please ask if you still have problems. The interactive websites also contain quizzes to allow you to test your understanding as you progress.

11.2 Provide feedback

This module has only been running for 3 years and I update it every year on the basis of your feedback. You can email me directly with comments to roy.sanderson@newcastle.ac.uk or you can contact your student rep, whose name is on your University app. Please **do not** wait until the end of the module to provide feedback, as it will be too late for me to correct problems. Let me know of issues as soon as possible and I will do my best to resolve them.