

MOTIVATIONAL CASE STUDY VIDEOS WITH R ANALYSES OF THE DATA

John A Harraway, Matthew R Schofield, and Jessica Allen
University of Otago, New Zealand
jharraway@maths.otago.ac.nz

Twenty motivational videos that focus on applied statistics are available on the University of Otago website. These are accompanied by data and lessons that are targeted at students in schools and universities that use GenStat software. The videos and lessons cover a range of techniques and have been popular with over 100,000 page visits since 2011. We outline transitioning the lessons to R software. R is freely available and widely used, including in the statistics curriculum at the University of Otago. The updated lessons are written using the Quarto package and are available online. Relative to static documents, no downloads are required, and the lessons offer new possibilities for interaction between the students and code.

INTRODUCTION

Twenty motivational videos for analysing research data have been developed to be used by students in schools and universities. A university course in statistics is required as a prerequisite for study and research in many academic subjects. Statistics is also a subject in the New Zealand school curriculum. The videos show real-world applications that demonstrate the ability of statistics to benefit science and society. The videos and lessons are freely available and can be accessed at <https://www.stats.otago.ac.nz/videos/statistics>. Each video uses real data collected or produced by an active researcher along with a set of lessons for each study relevant to that research.

The videos have proven to be popular. Since inception in April 2011, 112,436 visits have been logged to the central website. Over the past six months, 31 countries have more than 10 hits, with the most traffic from the United States (1,478 visits) and China (738 visits), where the origin of internet traffic comes from public records.

The available lessons guide readers through the statistical analyses discussed in the videos and assume the student is using the GenStat software package (2022). GenStat was an obvious choice at the initiation of the video programme 10 years ago because Vision International provided the GenStat package for free use in every high school in New Zealand, and a dedicated group of teachers developed lessons for each video using this software. However, most visitors to the site do not have access to GenStat. This limits both its appeal and its effectiveness.

We outline conversion of GenStat lessons to those written for the R software system. The R-project (R Core Team, 2022) is freely available software that is widely used in academia and industry. R is open source and can be used after studies have been completed. The software is well supported by the statistical community. At the time of writing there are nearly 2,000 packages available for download on CRAN (comprehensive R archive network) that extend the capabilities of the software, often including state-of-the-art statistical methodology. Integrated development environments (IDEs) are available for R (e.g., RStudio Team, 2022) that are user friendly. The R software is increasingly being used by applied researchers. For example, Lai et al. (2019) report ecology publications using R software increased from 11% in 2008 to 58% in 2017. At the University of Otago, the R language is used throughout the statistics curriculum, from introductory classes to graduate level courses. It is also being used in courses in other scientific disciplines, such as zoology and psychology. Making lessons available in R means that students develop familiarity with R, gaining confidence in a software package that will benefit their future training and research in their chosen fields.

CONVERSION TO R

The current GenStat lessons are available as a document download in portable document format (PDF). Our goal was not only to transition to the R language, but also to avoid lessons downloaded as static documents. Instead, we wanted the R code and any output (graphics, tables, etc.), to be available embedded on the website. Moreover, we wanted to ensure that any content was interactive, easy to access, and reproducible. To achieve this, we use the software Quarto (Allaire, 2022). Quarto can be thought of as a next generation version of R markdown (Allaire et al., 2022). It offers a writing and publishing environment for technical content that allows integration of code from several software

languages, including R, Python, and Julia. There are several output formats, including html, pdf, and Microsoft Word. We have used Quarto rather than R markdown for two reasons. The first is that Quarto offers more flexibility in designing the layout of the new lessons. The second is that we expect Quarto to receive more development and support in future years.

The website featuring the R code is under construction. The main page introduces and provides an overview of the project. It also provides an overview of the website, provides information about the lesson structure, and offers tips about how to interact with the lessons. The next page provides information about getting started with R and does not assume prior knowledge of the software. It provides a comprehensive and easy to follow introduction to the R language. Those already familiar with R can skip this page. Each lesson includes the following.

- *Video*: A motivational video embedded in the html document. The videos focus on a real-world application and on statistical methodology. The analyses presented do not rely on any statistical package. The videos give context and appreciation for the research undertaken and motivate the corresponding lesson.
- *Learning outcomes*: A list of outcomes students can expect to learn by completing the lesson.
- *Data*: Data that are freely available to download in Excel format.
- *Lesson Task 0*: Installation of any necessary R packages and loading the data into R.
- *Tasks*: Individual lesson components, with full explanations given for each task.

We expect the website to be complete in 2022. The content of the lessons themselves are largely unchanged (albeit converted from GenStat to R). Any modifications were made to allow easier conversion for use with R, or to ensure consistency between lessons. To illustrate many of the features of the website we use screenshots from a lesson based on a study that considers data from two populations of dolphins. One of these populations is based off the South Island of New Zealand, the other off the North Island of New Zealand. For each, measurements were collected for a sample of animals. Of interest is the evidence that the two populations are different species. Screenshots for this lesson are presented in Figures 1 and 2. We note that the lesson is still under development and that the final version may differ from that shown.

The lessons have been arranged hierarchically, which differs from the existing website where lessons are ordered chronologically. As new lessons were developed, they were added to the end of the list, which was convenient but far from optimal for student learning. The new website separates videos into three categories: (a) continuous data, (b) count data, and (c) time series data. Within each grouping, videos are arranged by difficulty of the problem. Earlier videos allow for students to learn core concepts that are reinforced in later lessons in the grouping. To assist with this, we make use of alert boxes. An orange alert box means that the current task is related to content from a previous lesson. There is a link to the previous lesson, allowing student to easily navigate and interact with the previous material (see Figure 2). Blue alert boxes inform the student that their current task is related to an earlier task in the same lesson.

The use of alert boxes focuses students' attention on important concepts. We use a yellow learning box to emphasize the learning outcomes at the start of each lesson (see Figure 1). The learning outcomes could be statistical (e.g., interpreting a simple linear regression), or they could be related to coding in R. Other important information is given in green alert boxes.

Having lessons available in html gives greater flexibility than lessons available as static documents. For example, we make use of tabs. Each task consists of at least three tabs. These correspond to 'task,' 'code,' and 'solution.' (See Figure 2.) The task tab is the tab that is initially open to the student. This tab outlines the objective of the task, e.g., 'Load data into R.' The second tab (which can be opened by clicking on the heading) is the code tab that displays the code necessary to complete the task. The use of Quarto enables two features in this tab. The first is easy copying of the R code via a clipboard button (see Figure 2(b)), which can be useful for those learning R because typographical and formatting errors can otherwise hinder learning and progress. The second feature is code folding, where certain 'chunks' of code can be shown or hidden with the click of a button. We set the default behavior of the code folding (whether the code is hidden or available) based on where the task falls in the lesson hierarchy. For tasks that involve multiple coding steps, we may choose to have steps that are familiar to students be initially hidden, whereas new material is shown for ease of learning. Users can reveal any hidden code with a single mouse click.

The third tab is the solution tab. This shows the output generated by running the code and includes any discussion or interpretation (if appropriate). In addition to providing balance for the website design, the tabs allow for self-directed, scaffolded learning (Al Mamun et al., 2020). In early lessons students can make extensive use of the code tab to assist with learning R. Similarly, the solution tabs can help with understanding statistical content. As students gain confidence, they can attempt to answer the questions before checking their answers by selecting the tabs.

A fourth tab entitled ‘extension’ is available in some tasks, particularly in later lessons. It contains extension questions that allow students to build their confidence, knowledge, and understanding, making use of what they have learned in the main task.

Statistics Video Presentations

- 1 Introduction
- 2 Getting started with R
- Continuous Data >
- 3 Cockles
- 4 Infrared Thermography
- 5 Māui's Dolphin
- 6 Altitude Adjustment
- 7 Property Sales
- 8 Trace Metals in Oysters
- 9 Rejection and Self-Esteem
- Count Data >
- 10 Epidemiology
- 11 Otago Stadium
- 12 Burden of Rotavirus Disease
- 13 Canine Cancer Detection
- Practice: Continuous and Count Data >
- 14 Iron Deficiency
- 15 Dangerous Driving
- Time Series Data >
- 16 Economic Analysis
- 17 Natural Resource Accounts

5 Māui's Dolphin


This lesson looks at the measurements of Māui's dolphins from the North and South Islands to see if these form separate species. This research is presented by Adam N. H. Smith (NIWA).

Māui's Dolphin Video

Code

Table of contents

- Māui's Dolphin Video
- Māui's Dolphin Data
- Māui's Dolphin Tasks



Māui's Dolphin Data

There is 1 file associated with this presentation. It contains the data you will need to complete the lesson tasks.

[Download Dolphins data.xls](#)

Māui's Dolphin Tasks

Learning Objectives

New skills and concepts:

1. Summarise response variable by levels of predictor variable.
2. Plots comparing several variables.
3. Wide to long data conversion.

Reinforcing skills and concepts seen in earlier lessons:

1. Read and format data.
2. Confidence interval and hypothesis test for difference in means.
3. Linear regression and ANOVA for single predictor.

Figure 1. Screenshot of the start of Lesson 5: Māui's Dolphin, with learning outcomes stated in the yellow alert box

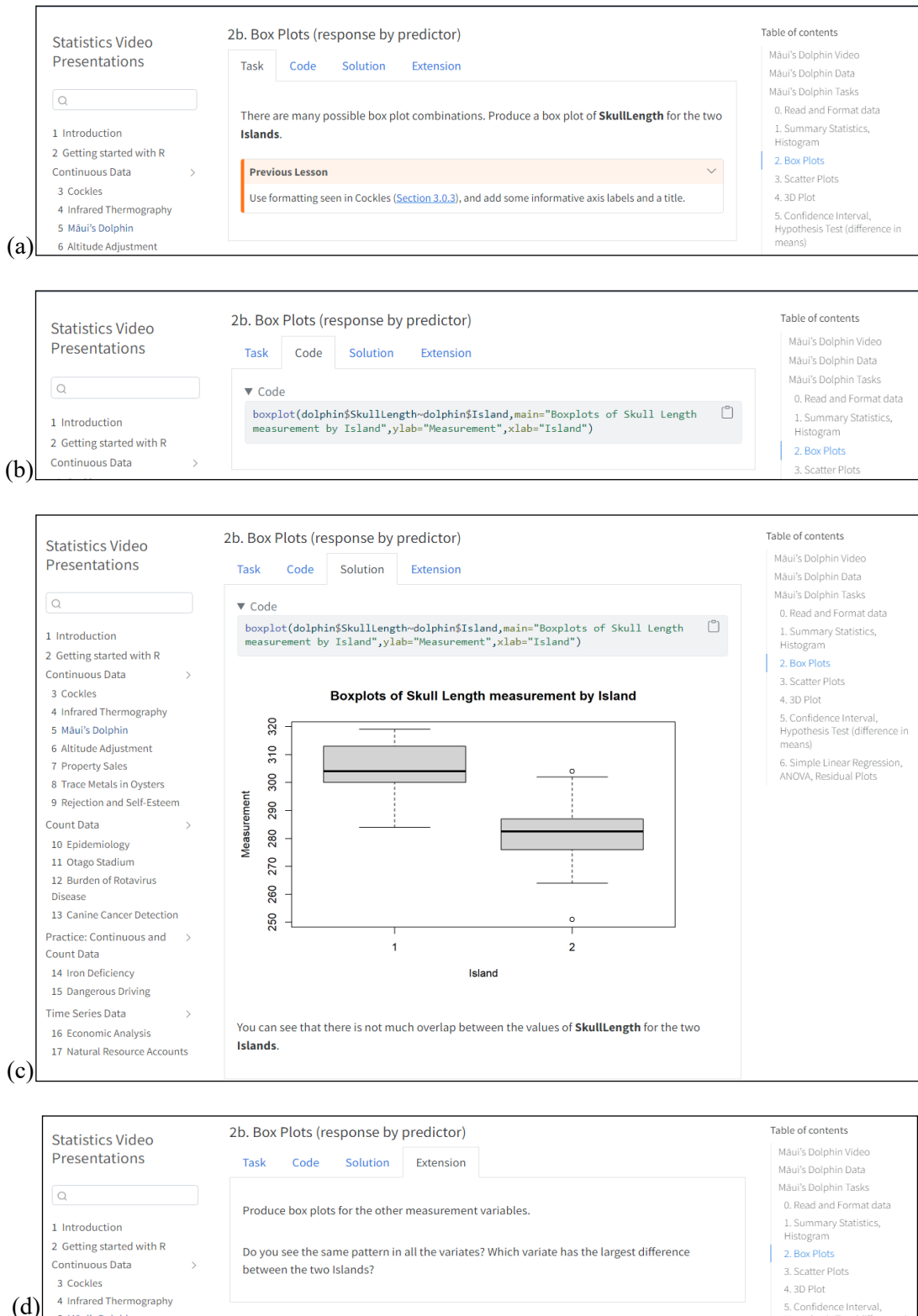


Figure 2. Screenshots showing Task 2b from Lesson 5 (Māui’s Dolphin). (a). Default tab outlining the task and including an alert box with a reminder of previous work in Lesson 3 (Cockles). (b). Code tab that shows the R code required to complete the task, with a clipboard symbol on upper right that enables code copying and a code folding arrow above the code box. (c). Solution tab that shows completed boxplot. (d). Extension tab that presents additional tasks to complete (i.e., obtain boxplots of other variables in the dataset).

DISCUSSION

The videos and accompanying lessons using R can be used as a training resource for large introductory statistics classes. The videos can motivate students in their studies, highlighting the value of statistics in a wide range of real applications. Working through the videos helps students gain the skills they need in their respective disciplines. It can aid them in understanding the quantitative results in research and statistical reports. The videos have been grouped according to the type of data being analysed, and further ordered according to the complexity of techniques that are being taught. This occurred as part of the conversion to R, with the intention of providing a more structured learning environment for students to gain confidence in applied statistical modelling and computation. The Cockles lesson introduces the construction of graphs, confidence intervals, hypothesis tests, Analysis of Variance (ANOVA), and multiple regression for continuous data sets. Subsequent lessons reinforce these techniques and present more advanced ones, including summary measures, function writing, and bootstrapping in the Infrared Thermography lesson and principal components analysis in the Trace Metals in Oysters lesson. Plotting, confidence intervals, hypothesis testing, logistic regression, ANOVA, and bootstrapping techniques are applied to categorical data in the Epidemiology and Titi lessons. These are practiced, along with higher level multiple logistic regression, stratification, and missing data strategies, in lessons such as Otago Stadium and Burden of Rotavirus Disease. The Iron Deficiency, Dangerous Driving, and Tourism applications involve a combination of data types and have many activities, potentially for use with extension and assessment. Canine Cancer Detection explores diagnostic testing. An introduction to time series visualization, decomposition, and interpretation is provided in the Economic Analysis lesson, with practice available in Natural Resource Accounts.

Use of the R language makes the videos appealing to a wider pool of students. We have taken care to ensure that there is introductory material that can support those with no previous knowledge of the R-project. Learning R from real examples is ambitious and challenging but comes with high payoffs. These payoffs include gaining familiarity with a tool, R, that is widely used in both academia and industry. For students, such knowledge can be of great benefit in their future studies, particularly if it involves a research component with data analysis. As well as this, familiarity with the R language is increasingly important in the job market.

The videos and lessons are also of benefit for those in the workforce, who may already have knowledge of R. This will be the case for the Indian Statistics Institute, who use the videos to train public service advisors for statistics work in developing countries. The same applies to Public Service workers being trained by the United Nations Institute for Training and Research (UNITAR) in Geneva through, e.g., the UNITAR e-learning course: *Understanding data and statistics better—for more effective SDG (Sustainable Development Goals) decision making*. The videos are available for both these organizations to use as they wish.

Another benefit of this development is to use the videos and R analyses for e-learning, not only in person at university, but also internationally. The R package is freely available for external students taught online. The importance of online learning has been highlighted by the COVID-19 pandemic that disrupted in-person education worldwide. Such videos could be used as a resource, not only for domestic students, but also for the education of international students. This is particularly relevant if there are limits to international travel. An example of this is New Zealand, where international students were unable to enter New Zealand between 2020 and 2022.

To retain student interest in this project, the videos need to be refreshed periodically with new case studies. This will highlight the interdisciplinary nature of statistics by showing an increasingly diverse range of applications. It will also ensure that students are being taught current and relevant statistical techniques. For example, the new study on the training of dogs to diagnose prostate, bowel, and cervical cancers in an accurate, safe, and non-invasive way has recently been made available on the website (this is the 20th video). This study was conducted by K9 Medical Detection, a health research group based in Dunedin, New Zealand (<https://www.k9md.org.nz/research>). It involves a biostatistician assessing the performance of canine detection of cancer. It involves the estimation of sensitivity, specificity, false negatives, and false positives for data collected in a double-blind testing procedure on laboratory-developed urine samples. Reports of hospital oncologists, scientists developing the urine samples initially in the laboratory for training the dogs, and reports of staff caring for the dogs are included. Proof of concept has been confirmed by the biostatistician; the dogs have now been trained to identify the cancers from laboratory developed odors for these three different types of cancer. A

clinical trial on patients from hospitals and General Practice Clinics is about to be undertaken. Similar trials are also using dogs for early detection of other diseases, including COVID-19. Diagnostic testing methods are an important part of statistical analysis. Students at all levels can learn from this case study that features the importance of a clinical trial after proof of concept.

When fully operational, we plan to get feedback on the webpage. We intend to survey users of the site. This will include students as well as schoolteachers, those in the workforce from developing countries being trained in India at the Indian Statistical Institute or online at the UNITAR in Geneva.

REFERENCES

- Al Mamun, M. A., Lawrie, G., & Wright, T. (2020). Instructional design of scaffolded online learning modules for self-directed and inquiry-based learning environments. *Computers & Education, 144*, Article 103695. <https://doi.org/10.1016/j.compedu.2019.103695>
- Allaire, J. J. (2022). *Quarto: R interface to `quarto' markdown publishing system* (Version 1.2) [Computer software]. Quarto. <https://CRAN.R-project.org/package=quarto>
- Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). *Rmarkdown: Dynamic documents for R* (Version 2.13) [Computer software]. RStudio. <https://rmarkdown.rstudio.com>
- GenStat. (2022). *GenStat data analysis software* (Version 22.1) [Computer software]. VSN International. <https://vsni.co.uk/software/genstat>
- Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere, 10*(1), Article e02567. <https://doi.org/10.1002/ecs2.2567>
- R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- RStudio Team. (2022). *RStudio: Integrated development environment for R* [Computer software]. RStudio. <http://www.rstudio.com/>