Insights from Data Companion

Owen Petchey, Andrew Beckerman, Natalie Cooper, Dylan Childs

2020-01-02

Contents

Introduction

```
knitr::opts_chunk$set(cache = FALSE)
```

Welcome!

{{< youtube EHLLmEcqRlk >}}

0.1 Introduction

The preface[1] of *Insights* informs about features of the book, such as its aims, content, structure, intended readership, and content that it does not include. Some of the text in the Preface may come across as a sales pitch (it probably is), but it also aims to make prospective readers clear about what they will find in *Insights* and why. Read the preface and answer questions here (directs to a different website)[2], directly concerning the content of the Preface, might also help that understanding.

Here you will find the additional information mentioned in the book:

- Questions and exercises (on a different web site)[3] with instant feedback that go alongside the material in the book. Use the questions and exercises to check, consolidate, and further your understanding of the material in the book.
- One or two additional case studies, with exercises and questions .
- Links to the datasets used in the three case studies in the book.
- Details of a live data analysis demonstration we often use in our introductory undergraduate classes.
- A list of additional case studies/datasets that you might practice with (students), or add to your class (instructors).
- A list of Related Books and further reading.

[1] `hyperlink%20to%20pdf%20of%20Preface?`
[2] `insightsfromdata.org`
[3] `link%20to%20other%20web%20site`

0.2 Workflow/checklist

Refer to this post: 2019-03-14-workflow-checklist-for-data-analysis

0.3 R-project setup

Here are the ready made empty folders and Rproject files (compressed in a zip file you need to unzip – {{% staticref "files/Insights_projects.zip" "newtab" %}}Insights_projects.zip{{% /staticref %}}) mentioned in *Insights* (the book) section *R-Projects, the best thing since sliced bread.*

0.4 For instructors

Perhaps add the "Notes/ideas for instructors section from the book Preface."

0.4.1 Live data analysis demonstration

In the first class of the first week of an *Introduction to Data Analysis* course, we lead a *live data analysis demonstration*. Within one hour we go from question to answer, including collection of some data about each of the students. We believe this demonstration helps students connect with the importance and fun of the content of the course. A walkthrough of the live data demonstration is provided as an {{% staticref "files/live_data_demo.Rmarkdown" "newtab" %}}Rmarkdown file{{% /staticref %}} and as {{% staticref "files/live_data_demo.html" "newtab" %}}html{{% /staticref %}} rendered from the rmarkdown.

0.5 Datasets

Here are the datasets you will need in order to work along with the three case studies in *Insights*:

- **Case Study 1 (bat diets)** is about what bats eat and in it we work the data from the published study: *Female dietary bias towards large migratory moths in the European free-tailed bat (Tadarida teniotis)*, by Mata and colleagues in 2016 in the journal *Biology Letters* paper on Biology Letters website[4]. You must get the dataset from the online data repository in which the researchers deposited it. Get the `dataset_Mata.et.al.2016.xlsx` from the folder **Diet analysis dataset**

[4] `http://rsbl.royalsocietypublishing.org/content/12/3/20150988`

from the dryad repository[5]. Please make sure you get the file from version 2 of the dryad data publication. The dataset is stored on dryad in Excel format.

- **Case Study 2 (prey diversity)** is about the hypotheses that more pathways of energy flow into a predator (i.e. more prey species) stabilise the predator population dynamics (the article about case study 2 is published here[6]). Get the data the data file `dileptus expt data.csv` from the dryad repository[7] associated with the publication about the data (make sure you get version 2 of the dataset).
- **Case Study 3 (dietary diversty and polity)** is about whether the diversity of food available to the population of a country is associated with the political system of that country. It requires three datasets:
  - A dataset containing information about country's political system. You can acquire the full dataset from this web page of the Center for Systemic Peace[8]. Scroll down to the section *Polity IV: Regime Authority Characteristics and Transitions Datasets* and click on the *Excel Series* link to the right of the *Polity IV Annual Time-Series, 1800-2017* box.
  - A dataset of food available in countries (FAO food balance sheet data) from the FAO website[9]. The datafile could get is called `FoodBalanceSheets_E_All_Data.csv` and is 206MB. You can, as mentioned in the book, opt for the trimmed-down version of the dataset that we supply click here to start download[10].
  - A dataset that matches some country names between the political system and FAO datasets click here to start download[11].

0.6 More case studies

Here are several more datasets, again all about food, and due to the bias of the authors of the book, mostly analyses of diet composition of various organisms. As the book goes to press, and unless we have since added additional information below, we only give the link to the original publication and to the dataset, and we have not ourselves worked through these data. Its all up to you! But don't hesitate to get in touch if you do work through them, and if you find something odd, difficult, or even impossible.

[5] `https://doi.org/10.5061/dryad.m8t72.2`

[6] `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2656.2000.00446.x`

[7] `https://datadryad.org/resource/doi:10.5061/dryad.7h62c8t.2`

[8] `http://www.systemicpeace.org/inscrdata.html`

[9] `http://www.fao.org/faostat/en/#data/FBS/metadata`

[10] `www/FoodBalanceSheets_E_All_Data_reduced.Rdata`

[11] `www/data/countryname_map.csv`

0.6.1 Hungry ladybirds

An analysis of what determines how fast ladybirds eat.

- Original publications: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13159
- Data: https://datadryad.org/resource/doi:10.5061/dryad.gq224h3

0.6.2 Seal suppers

An analysis of what seals eat.

- Original publication: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4474?af=R
- Data: https://doi.org/10.5061/dryad.g23j32s

0.6.3 More bat poop

Another analysis of bat diets. This one of data from 1'252 faecal pellets of five species of bat.

- Original publication: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4559
- Data: https://doi.org/10.5061/dryad.6880rf1.

0.6.4 Marten isotopes

Analysis of marten diets by stable isotope analysis

- Original publication: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4559
- Data: https://doi.org/10.5061/dryad.6880rf1.

0.6.5 Snake diets

- Original publication: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.12972
- Data: https://datadryad.org/resource/doi:10.5061/dryad.8kt4675

0.6.6 Desert bat diets

- Original publication: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.4896
- Data: https://datadryad.org/resource/doi:10.5061/dryad.7j0c8dm

0.6.7 Desert bat diets

- Original publication: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.4896
- Data: https://datadryad.org/resource/doi:10.5061/dryad.7j0c8dm

0.6.8 Birds eating insects

- Original publication: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.4787
- Data: https://datadryad.org/resource/doi:10.5061/dryad.4f1n785

0.6.9 Diets of predatory fish

- Original publication: https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.4857
- Data: https://datadryad.org/resource/doi:10.5061/dryad.0jm1dt2

0.6.10 Cervical spine compression and MRI (not food related)

- Original publication: https://doi.org/10.5061/dryad.kk653rs
- Data: https://datadryad.org/resource/doi:10.5061/dryad.kk653rs

0.6.11 Lots of other datasets here:

https://www.nature.com/articles/s41559-017-0458-2/tables/1

and here:

https://cran.r-project.org/web/packages/TH.data

0.7 Related reading

Refer to this article: 2019-03-14-some-useful-r-help-tutorial-reference-websites

- Grafen & Hails (2002) Modern Statistics for the Life Sciences. 368 pages. Focuses on and thoroughly covers statistics, using general linear models. Works with Minitab, SAS, SPSS.
- Crawley (2005) Statistics - An Introduction Using R. 327 pages. A concise introduction focused on statistical analyses using R. Crawley (2012) The R Book. 1076 pages. A comparatively encyclopedic account of R; "extensive and comprehensive". Hothorn & Everitt (2014) A Handbook of Statistical Analysis Using R. 456 pages. Focuses on statistical analyses; probably more graduate level.
- Whitlock & Schluter (2015) The Analysis of Biological Data. 818 pages. Contains practice & assignment problems. Focused on statistics, covers data management/visualization in passing.
- Maindonald & Braun (2010) Data Analysis and Graphics Using R. 549 pages. Assumes some existing knowledge of statistics and data analysis. For final year undergraduate / graduate level. Reaches to Bayesian methods, GLMMs, and random forests.
- Hector (2015) The New Statistics with R. 199 pages. Focused on statistics, specifically linear models. "New" refers to new methods that are included, and focusing on effect sizes rather than p-values.
- Field, Miles, & Field (2012) Discovering Statistics using R. 957 pages. Focused on statistics, though covers data management and visualization. Goes up to multilevel linear models. Classic R and R Commander (no RStudio). Written with humour, has "characters", associated website with datasets, scripts, webcasts, self-assessment question, additional material, answers, powerpoint slides, links, and cyberworms of knowledge.
- Field (2016) An Adventure in Statistics. 768 pages. At first (and perhaps later) sight quite inspirational. Starts with a chapter on why we need science (maybe to get insights?) followed by one on reporting findings. As such, has similar approach to Insights, to start with motivation and with the end in mind. Continues with a thorough account of data analysis and statistics suitable for undergraduates.
- Bolker (2008) Ecological Models and Data in R. 396 pages. Page 3 states "I assume that you've had the equivalent of a one-semester undergraduate statistics course..." and on page 4 "If you have used R already, you'll have a big head start." Venables, Smith, et al (2009) An Introduction to R. Reference book for the R Language (classic R). Very concise. Contains a 15-page chapter on statistics, including linear and non-linear models.
- Grolemund & Wickham (2017) R for Data Science. 492 pages. Focus on "Data Science", "an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge." Book organized broadly by the workflow: Explore, Wrangle, Program, Model, Communicate. Quite comprehensive in coverage of the "tidyverse" approach to using R.

- McKillup (2012) Statistics Explained. An Introductory Guide for Life Scientists. 400 pages. Quite well rounded, including experimental design, collecting and displaying data, doing science, ethics. Majority walks through statistical tests… linear models, non-parametric tests, multivariate.
- Dytham (2010) Choosing and Using Statistics: A Biologist's Guide. 320 pages. Focused on statistics, as the title suggests.
- Adler (2012) R in a Nutshell. A Desktop Quick Reference. 611 pages. A great reference book.
- Dalgaard (2008) Introductory Statistics with R. 364 pages. A concise introduction focused on statistical analyses using R. S* pector (2008) Data Manipulation with R. 154 pages. Covers importing data, working with databases, character manipulation, dealing with dates, using loops, conversion to data frames.
- Ellis (2010) The Essential Guide to Effect Sizes. 188 pages. Focuses on interpreting the practical everyday importance of research results, power, and synthesizing disparate results. Does this via effect sizes. Based on a course for honed on "smart graduate students".
- Gotelli & Ellison (2012) A Primer of Ecological Statistics. 614 pages. Upper-undergraduate to graduate level. Probability and statistical thinking, distributions, central tendency and spread, p-values, etc. Then experimental design; then specific analyses. Finishes by covering estimates of diversity and occurrence.
- Gonick & Smith (1993) The Cartoon Guide to Statistics. 230 pages. Covers summary and display of data, probability, central limit theorem, confidence interval estimation, etc.
- McKillup (2011) Statistics Explained. An Introductory Guide for Life Scientists. 416 pages. Begins by explaining about doing science, collecting and displaying data, experimental design, and responsibility and ethics. Then works through a good list of statistical methods for beginning to upper-level undergraduates.
- Sokal & Rohlf (1995) Biometry. The Principles and Practices of Statistics in Biological Research. 880 pages. Thorough, comprehensive, and often quite technical title focused on statistics.
- Zar (2010) Biostatistical Analysis. 960 pages. Thorough and comprehensive coverage of "statistics analysis methods used by researchers to collect, summarise, analyse and draw conclusions from biological research. Suitable for beginners to advanced users.
- McElreath (2016) Statistical Rethinking. 469 pages. Brilliant. What should be taught to undergraduates, if only the world would then be ready for them.
- Healy (2017) Data Visualisation for Social Science. A practical introduction with R and ggplot2. Focuses on appropriate visualization for getting knowledge from data. Covers principles and practices of looking and presenting data.
- Zumel & Mount (2019) Practical Data Science with R.

Insights Workflow

Before touching R/RStudio

- Question
- Hypothesis(es)
- Study methods, materials, design.
- Response variable(s)
- Explanatory variable(s)
- Prediction(s)
- Secure resources
- Perform study, including data collection.

If you don't know, e.g. because you did not conduct the study, *before* import into R inspect the datafiles in a spreadsheet program (so long as they're not too big) and note the following:

- if multiple datafiles are used, which contains what
- what variable names are used in the datafiles, and what these mean (i.e. which are response variables, which are explanatory, and what are others)
- number of rows and columns in the datafiles
- arrangement of the data in the datafile, e.g. tidy or not tidy
- any obvious things to deal with (e.g. how missing values are coded, date/time information, codes that need expanding, variable/column names that will need changing)

After we read the datafile(s) into R:

- number of variables/columns
- number of rows
- variable types
- appropriate reprentation of missing values
- tidy the data (at some point)
- check for innapropriate duplicates
- fix dates
- replace any codes with informative words
- check for appropriate variable entries, e.g. levels of characters, ranges of numerics
- numbers of "things", number of experimental units, treatments, treatment combinations, temporal samples
- calculate response and/or explanatory variable(s) (if required)

After data tidying and cleaning

- Shapes of variables (i.e. inspect the histograms of explanatory and response variables).
- Relationships among explanatory variables.
- Relationships relevant to hypotheses/predictions.
- Assess confidence in revealed patterns.

Additional R

To be completed.

Additional concepts

To be completed.

Are diets more diverse in more democratic countries?

0.8 About this case study

This case study is primarily meant to supplement the material in the book *Insights from Data* by Petchey, Beckerman, Childs, and Cooper. If you don't understand something in the case study, have a look at that book (perhaps again), or search for help, or get in touch with us.

0.9 Introduction to the study and data

In this case study we will look at the diversity of food available to people living in countries around the world. And we will explore one potential influence of this diversity—the political system (e.g. democratic) of the country. Why this? Why not? There are likely lots of factors governing the diversity of food available to the population of a country, and political system is just one. We could also look at gross domestic product, location, population size, and so on. But let's keep things relatively simple and straightforward for the moment. You can make things more complex if you choose to look further into the question, and it may be essential to do so. Just wait until after you've mastered the simpler though perhaps still challenging material in this case study.

To be clear, we are asking the question of whether the diversity of food available to the population of a country is associated with the political system of that country. We hypothesise that more freedom will result in more diversity of food availability (figure XXa – illustrate the hypothesis). We will restrict ourselves to two variables: diversity of food available and political system. Let's use a measure diversity known as the Shannon index, which is greater if there are more food types and if the amount of each is more evenly distributed among them (figure XXb – illustrate shannon in 2x2). As far as we know, the analyses are novel. Any insights will be new! Though, since the findings are not published /