BASIC STATISTICS FOR BIOLOGISTS

UNIVERSITÄT
LEIPZIG

Erik Kusch

erik.kusch@i-solution.de

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

Course Dates & Outline I

Block I - Theory and Basics of `R`

Date	Time	Topic	Location
I.) Introduction			
Date	Time	(1) An Introduction to Basic Statistics for Biologists	Location
Date	Time	(2) Introduction to R	Location
II.) Basic statistical terminology			
Date	Time	(3) A Primer for Statistical Tests	Location
Date	Time	(4) Descriptive Statistics	Location
Date	Time	(5) Data Visualisation	Location
Date	Time	(6) Inferential Statistics, Hypotheses and our Research Project	Location

Course Dates & Outline II

Block II - Basic Statistics in R

Date	Time	Topic	Location
III.) Handling Data			
Date	Time	(7) Data Handling and Data Mining	Location
IV.) Non-parametric tests			
Date	Time	(8) Nominal Tests	Location
Date	Time	(9) Correlation Tests	Location
Date	Time	(10) Ordinal and Metric Tests for two-sample situations	Location
Date	Time	(11) Ordinal and Metric Tests for more than two-sample	Location
V.) Parametric tests			
Date	Time	(12) Simple Parametric Tests	Location
VI.) Closing			
Date	Time	(13) Summary and an Outlook on Advanced Statistics	Location

Learning Goals

1. A solid grasp of basic biostatistics
   - Have an overview of available methods
   - Be able to judge the applicability of individual methods
2. Basic proficiency in using R
   - Know base commands and how they function
   - Be able to prepare biologically relevant data sets for further analysis
   - Be able to apply basic statistical methods to biologically relevant data sets
3. Research Design
   - Understand how to formulate testable hypotheses
   - Know the importance of proper statistical approaches in research
   - Being able to critically assess statistical methods in research publications

Learning Methods

We will:

- Cover useful theory of biostatistics (lecture style)
- Run biostatistical analyses in R (seminar style)
- Work through basic biostatistical methods in a research project using simulated data
- Fully reproducible analyses (https://github.com/ErikKusch/An-Introduction-to-Biostatistics-Using-R)

We will focus heavily on actually doing the statistics!

Let Me Introduce Myself

Erik Kusch

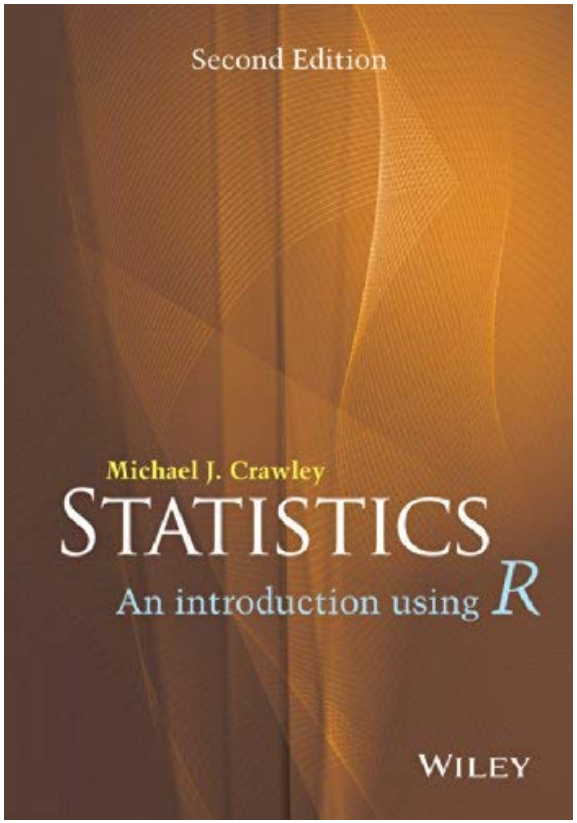
Studies:
PhD @ Aarhus University (currently enrolled)
M.Sc. @ University of Bergen
B.Sc. @ Technical University of Dresden

Experience:
Biostatistics Tutor @ University of Leipzig
Biostatistics Research Assistant @ University of Leipzig
Biostatistics Research Assistant @ University of Kyoto

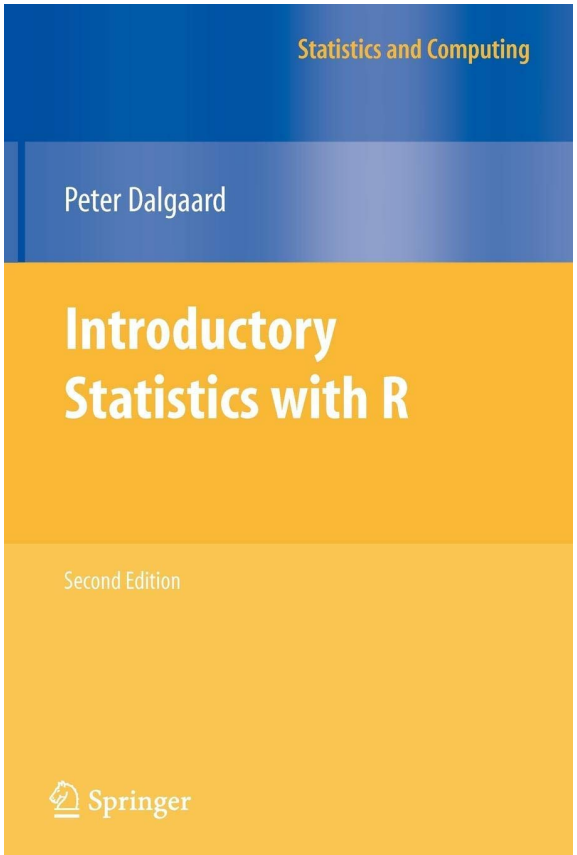
Research:
- Dryland vegetation memory analyses
- Large-scale vegetation-climate modelling
- Remote sensing approaches in macroecology
- Biostatistical approaches in behavioural ecology
- Statistical downscaling of climate reanalysis data for use in biological analyses

Useful Reading

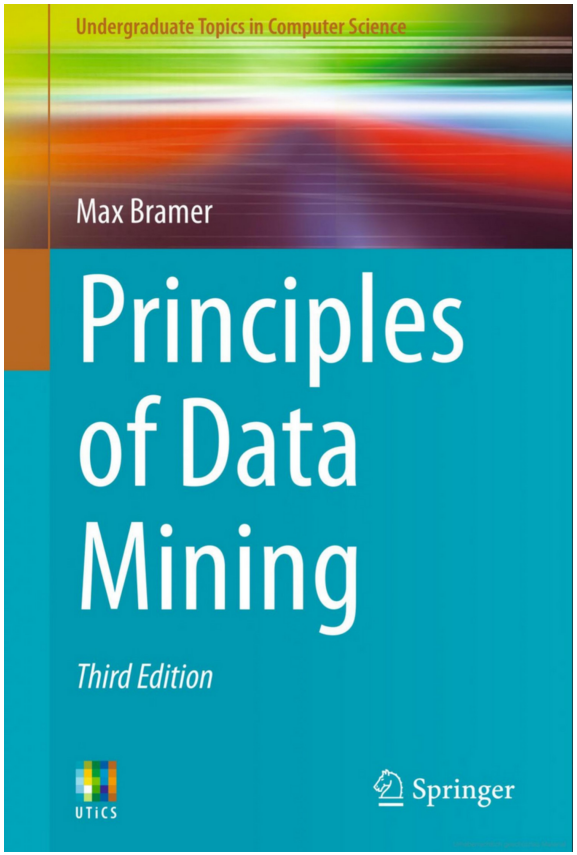
You are **NOT** required to read these!



ISBN: 978-1-118-94109-6



ISBN: 978-0-387-79053-4



ISBN: 978-1-4471-4883-8

But these books are seriously good.

When Mistakes Happen

Even the rigorous peer-review system might miss some minor flaws.

An example:

- Birkenmeyer et. al published a flawed paper in 2016.

The mistake in the data set was spotted by Dr. B. M. Weiß. in early 2017

- A corrigendum was put online
- A corrected version of the paper was uploaded

None of the results of the paper changed.

Chemical Senses, 2016, Vol 41, 525–535
doi:10.1093/chemse/bjw056
Original Article
Advance Access publication April 27, 2016

OXFORD

Original Article

Sampling the Body Odor of Primates: Cotton Swabs Sample Semivolatiles Rather Than Volatiles

Claudia S. Birkemeyer[1], Ruth Thomsen[2,3], Susann Jänig[2,4], Marlen Kücklich[2,4], Anna Slama[2], Brigitte M. Weiß[2,4] and Anja Widdig[2,4,5]

[1]Research Group of Mass Spectrometry, Institute of Analytical Chemistry, Faculty of Chemistry and Mineralogy, University of Leipzig, Linnéstr. 3, 04103 Leipzig, Germany, [2]Research Group of Behavioural Ecology, Institute of Biology, Faculty of Bioscience, Pharmacy and Psychology, University of Leipzig, Talstr. 33, 04103 Leipzig, Germany, [3]Department of Anthropology, University College London, Gower Street, London WC1E 6BT, UK, [4]Junior Research Group of Primate Kin Selection, Department of Primatology, Max-Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany and [5]German Center for Integrative Biodiversity Research (iDiv), Deutscher Platz 5E, 04103 Leipzig, Germany

Correspondence to be sent to: Claudia Birkemeyer, Institute of Analytical Chemistry, University of Leipzig, Linnéstr. 3, 04103 Leipzig, Germany. e-mail: birkemeyer@chemie.uni-leipzig.de

Accepted 22 March 2016.

Abstract

We assessed the suitability of a frequently used sampling method employing cotton swabs for collecting animal body odor for gas chromatography–mass spectrometry (GC-MS) analysis of volatile organic compounds (VOCs). Our method validation showed that both sampling material and sampling protocols affect the outcome of the analyses. Thus, among the tested protocols swabs of pure viscose baked before use and extracted with hexane had the least blank interferences in GC-MS analysis. Most critical for the recovery of VOCs was the handling time: the significant recovery losses of volatiles experienced with this sampling procedure suggest that a rapid processing of such samples is required. In a second part, we used swab sampling to sample the body odor of rhesus macaques (*Macaca mulatta*), which lack scent glands. First results after GC-MS analysis of the samples collected from these nonhuman primates emphasize that proper analytical performance is an indispensable prerequisite for successful automated data evaluation of the complex GC-MS profiles. Moreover, the retention times and the nature of the identified chemical compounds in our samples suggest that the use of swabs is generally more appropriate for collecting semivolatile rather than VOCs.

Key words: body odor, GC-MS profiling, rhesus macaques, semivolatile organic compounds (sVOCs), validation of swab sampling, volatile organic compounds (VOCs)

Introduction

In recent years, it has become increasingly evident that nonhuman primates widely use the olfactory sense, relying on the reception and recognition of chemical signals, to assess their social and ecological environment (e.g., Palagi et al. 2003; Matsumoto-Oda et al. 2007). Primates use olfactory cues, for instance, to distinguish individuals (Palagi and Dapporto 2006; Scordato and Drea 2007), sex (Drea 2015) and females' fertility (Crawford and Drea 2015) as well as group membership (Henkel et al. 2015). Hence, it is not surprising that research on olfactory cues in primates increasingly pursues the presence and function of such signals.

525

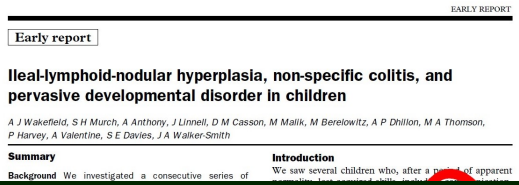
→ No big deal so long as you offer corrections to your flawed work.

Fraudulent Practices - The Case Of Andrew Wakefield

Probably one of the most reviled doctors of the 21st century

- Claimed to have found a link for vaccines and autism (Paper from 1998)
- Paper retracted by the publisher
- General Medical Council of Britain revoked his medical license

His academic career is over despite his large community of followers in the U.S., Australia and Brazil.



EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of

Introduction

We saw several children who, after a period of apparent

Fraudulent Practices - The Case Of Diederik Stapel

Former star in academia, now a laughing stock

- Manipulated data and completely fabricated entire studies
- Fired from his position as professor at Tilburg University
- 58 retracted papers
- Papers of other authors needed to be retracted as well

$\rightarrow$ Knowingly fraudulent practices can cost you your career, discredit your institution and your field of research, and even seriously impede the careers of unknowing co-workers.

Wrong/Malinformed Use

Lack of statistical knowledge

- Applying statistics to data which they aren't meant for

$\rightarrow$ *Methods can "break"*

- Flawed understanding of the methodology

$\rightarrow$ *Incorrect conclusions*

Pure biologists lack knowledge on statistics.

Uninformed Use

Lack of biological knowledge

- Delineation of nonsensical but statistically significant relationships

$\rightarrow$ *p-hacking*

- No sense of how to establish testable, feasible hypotheses

$\rightarrow$ *Waste of time*

Pure statisticians lack knowledge on biology.

Caveat

- Biologists often have preformed ideas of what to expect
  $\rightarrow$ data-tweaking to match expectations?
- Researchers also have a vested interest in uncovering extraordinary things
  $\rightarrow$ The more astounding a paper the better?

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

BY REGINA NUZZO

In 2013, five years after he co-authored a paper showing that Democratic candidates in the United States could get more votes by moving slightly to the right on economic policy[1], Andrew Gelman, a statistician at Columbia University in New York City, was chagrined to learn of an error in the data analysis. In trying to replicate the work, an undergraduate student named Yang Yang Hu had discovered that Gelman had got the sign wrong on one of the variables.

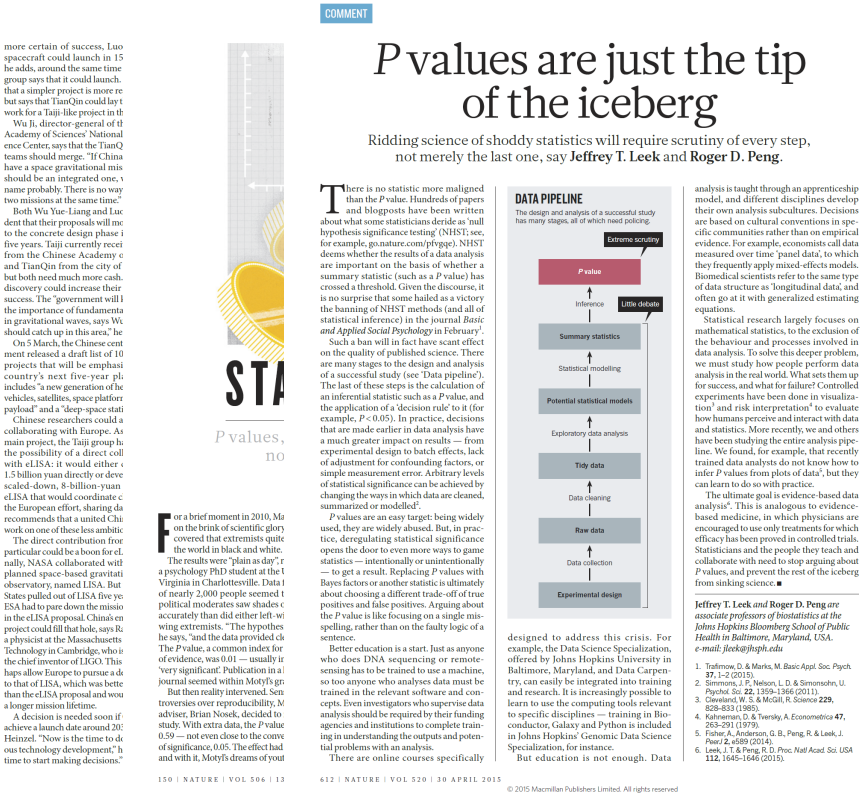
Gelman immediately published a three-sentence correction, declaring that everything in the paper's crucial section should be considered wrong until proved otherwise.

ATTENTION!

Don’t let a personal bias inform your analysis!

The Recent Debate

- p-values are a cause of concern
  → More on this in seminar 6 (Inferential Statistics and Hypotheses)
- Pre-p-value statistics and data handling increasingly subject of scrutiny
  → More on this in seminar 7 (Data Handling and Data Mining)



COMMENT

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

There is no statistic more maligned than the *P* value. Hundreds of papers and blogposts have been written about what some statisticians deride as 'null hypothesis significance testing' (NHST; see, for example, go.nature.com/pfvqqe). NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a *P* value) has crossed a threshold. Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal *Basic and Applied Social Psychology* in February[1].

Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study (see 'Data pipeline'). The last of these steps is the calculation of an inferential statistic such as a *P* value, and the application of a 'decision rule' to it (for example, $P < 0.05$). In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error. Arbitrary levels of statistical significance can be achieved by changing the ways in which data are cleaned, summarized or modelled[2].

P values are an easy target: being widely used, they are widely abused. But, in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result. Replacing *P* values with Bayes factors or another statistic is ultimately about choosing a different trade-off of true positives and false positives. Arguing about the *P* value is like focusing on a single misspelling, rather than on the faulty logic of a sentence.

Better education is a start. Just as anyone who does DNA sequencing or remote-sensing has to be trained to use a machine, so too anyone who analyses data must be trained in the relevant software and concepts. Even investigators who supervise data analysis should be required by their funding agencies and institutions to complete training in understanding the outputs and potential problems with an analysis.

There are online courses specifically designed to address this crisis. For example, the Data Science Specialization, offered by Johns Hopkins University in Baltimore, Maryland, and Data Carpentry, can easily be integrated into training and research. It is increasingly possible to learn to use the computing tools relevant to specific disciplines — training in Bioconductor, Galaxy and Python is included in Johns Hopkins' Genomic Data Science Specialization, for instance.

But education is not enough. Data analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world. What sets them up for success, and what for failure? Controlled experiments have been done in visualization[3] and risk interpretation[4] to evaluate how humans perceive and interact with data and statistics. More recently, we and others have been studying the entire analysis pipeline. We found, for example, that recently trained data analysts do not know how to infer *P* values from plots of data[5], but they can learn to do so with practice.

The ultimate goal is evidence-based data analysis[6]. This is analogous to evidence-based medicine, in which physicians are encouraged to use only treatments for which efficacy has been proved in controlled trials. Statisticians and the people they teach and collaborate with need to stop arguing about *P* values, and prevent the rest of the iceberg from sinking science. ■

Jeffrey T. Leek *and* **Roger D. Peng** *are associate professors of biostatistics at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA.*
e-mail: jleek@jhsph.edu

1. Trafimow, D. & Marks, M. *Basic Appl. Soc. Psych.* **37**, 1–2 (2015).
2. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22**, 1359–1366 (2011).
3. Cleveland, W. S. & McGill, R. *Science* **229**, 828–833 (1985).
4. Kahneman, D. & Tversky, A. *Econometrica* **47**, 263–291 (1979).
5. Fisher, A., Anderson, G. B., Peng, R. & Leek, J. *PeerJ* **2**, e589 (2014).
6. Leek, J. T. & Peng, R. D. *Proc. Natl Acad. Sci. USA* **112**, 1645–1646 (2015).

Practices in statistics are constantly subject to change.

Why Keep Up With It?

- Journals mights enact bans on studies containing p values
  $\rightarrow$ Counter-productive according to Andrew Vickers (Memorial Sloan Kettering Cancer Center)
- Statistically robust studies hold up to scrutiny much better
  $\rightarrow$ Statistical prowess enhances your research massively
- Staying up-to-date can help advance one's understanding and career

COMMENT

EXHIBITION Adolf Fleischmann, pathology sculptor and abstract artist p.30 | INTERDISCIPLINARITY Can architecture catalyse creativity at the Crick? p.32 | CONSERVATION Legal loophole allows mango farmers to cull fruit bats in Mauritius p.33 | GEOLOGY Deep-drilling pioneers: what were they drinking? p.33

ILLUSTRATION BY DAVID PARKINS

A tragedy of errors

Mistakes in peer-reviewed papers are easy to find but hard to fix, report **David B. Allison** and colleagues.

Just how error-prone and self-correcting is science? We have spent the past 18 months getting a sense of that.

We are a group of researchers working on obesity, nutrition and energetics. In the summer of 2014, one of us (D.B.A.) read a research paper in a well-regarded journal estimating how a change in fast-food consumption would affect children's weight, and he noted that the analysis applied a mathematical model that over-estimated effects by more than tenfold. We and others submitted a letter[1] to the editor explaining the problem. Months later, we were gratified to learn that the authors had elected to retract their paper. In the face of popular articles proclaiming that science is stumbling, this episode was an affirmation that science is self-correcting.

Sadly, in our experience, the case is not representative. In the course of assembling weekly lists of articles in our field, we began noticing more peer-reviewed articles containing what we call substantial or invalidating errors. These involve factual mistakes or veer substantially from clearly accepted procedures in ways that, if corrected, might alter a paper's conclusions.

After attempting to address more than 25 of these errors with letters to authors or journals, and identifying at least a dozen more, we had to stop — the work took too much of our time. Our efforts revealed invalidating practices that occur repeatedly (see 'Three common errors') and showed how journals and authors react when faced with mistakes that need correction.

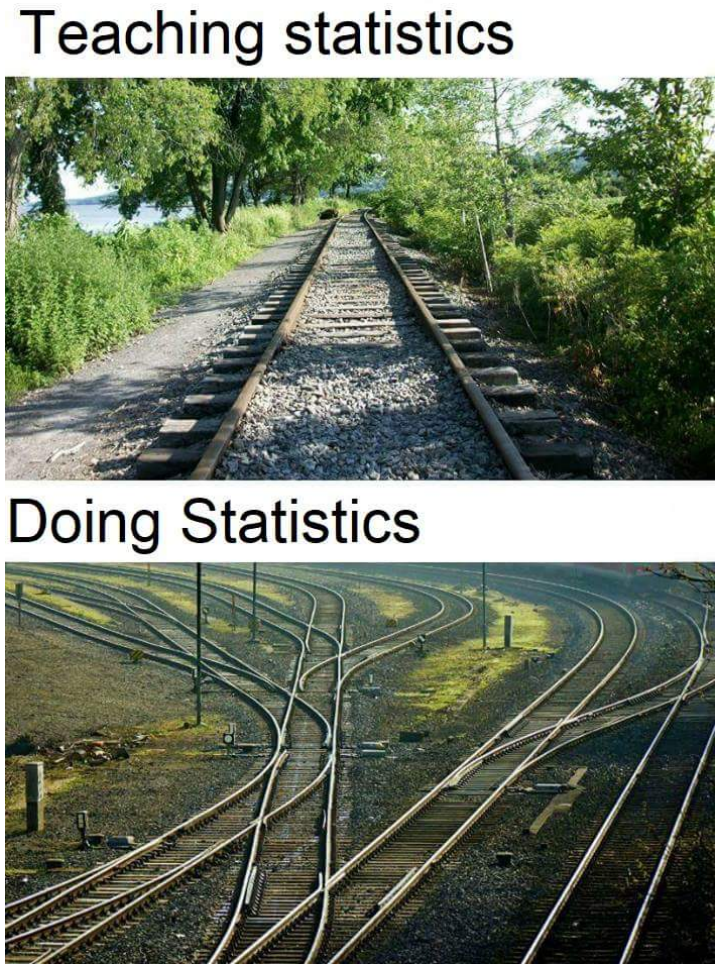
We learned that post-publication ▶

NATURE.COM
For *Nature*'s special collection on reproducibility, see: go.nature.com/huhbyr

Advancing In Statistics

"Treat statistics as a science, and not a recipe!"

$\sim$ *Andrew Vickers*



The Lack Of Biostatisticans

- Biological studies without rigorous statistical analyses are almost unpublishable
- Biostatisticians are rare
- Almost every biological research group requires at least one capable statistician

$\rightarrow$ **Biostatisticians are sought-after**

Statistics As An Apphrodisiac

Her: I'm a stats major
Me: [trying to think of something to impress her] **yea I'm bad at math too**

Frequently Used Classifications

- According to how they are done:
  - Theoretical Statistics
  - Applied Statistics
- According to topic:
  - Biostatistics
  - Economic Statistics
  - Statistical Physics
  - ...
- According to what the goal is and what kind of data is available
  - Regression
  - Classification
- According to how the analyses makes use of the data
  - Supervised
  - Unsupervised

According to the kind of information returned by the methods

- Descriptive Statistics
- Inference/Inferential Statistics

Unsupervised Approaches

Unsupervised methods are often *used to select* the *most informative* X *input variables* for *supervised approaches*.

Pre-requisites:

- Only *input variables* are observed.
- *No solution/feedback (output)* is given.

Aims:

- *Divide* the observations into relatively distinct groups.
- *Model* the underlying structure or distribution in the data.

$\rightarrow$ **"Pre-processing" before a supervised learning analysis** and **exploratory analyses**

Supervised Approaches

Supervised methods are often *informed by unsupervised approaches* and used to *gain validated information* about the data.

Pre-requisites:

- Both *predictors* $\mathbf{X}$, and *responses* $\mathbf{Y}$ are observed (there is one y_i for each x_i).
- Data is split into *Training* and *Test Data Sets*.

Aims:

- Learn a *mapping function* f from X to Y.
- *Validate* established function/model.
- Further *prediction* and *inference*.

$\rightarrow$ **Mostly inferential analyses**

Population vs. Sample

Population: describes the sum total of all *existing* values of a variable given a certain research question. This includes non-measured data.

Sample: describes the sum total of all *available* values of a variable for any given analysis. This can only include measured data.

An example:

In an experimental set-up, you rear an ant colony of exactly 10,000 individuals. You are interested in the average mandible strength of ants within the colony.

The problem: You cannot possibly take measurements of all 10,000 individuals.

The solution: Taking measurements on a **Sample** (e.g. 1,000 individuals) from within the **Population** (10,000 individuals).

Training Data vs. Test Data

This differentiation is only applicable when concerned with *modelling*, which we won't cover in these seminars.

Training Data: describes the subset of the total data which is used to *establish/train* the model.

Test Data: describes the subset of the total data which is used to *test* the performance of the model.

The problem: You have identified a way to model how mandible strength and ant size are interconnected but don't know how to assess the quality of your model (a model will always fit the data it was built on extremely well).

The solution: Split the available data into two non-overlapping subsets of data (**Training** and **Test Data**) and use these separately to build your model and assess its performance.

What Makes Data Truly Random?

Randomisation is one of the **most important** practices in biological studies.

A **sampling** procedure is **random** when any member of the *population* has an equal chance of being selected into the *sample*.

Training and *Test Data Sets* are established from the population with the same sense of randomness although there may be exceptions depending on the modelling procedure at hand.

Data collection: Number all units contained within the set-up and sample those units corresponding to random numbers.

In R: Use the `sample()` function to create truly random subsets. Remember to use `set.seed()` to make this step reproducible!

Random Sampling in R

```r
# Making it reproducible
set.seed(42)
# Establishing a population
pop <- c(1:15)
pop
```

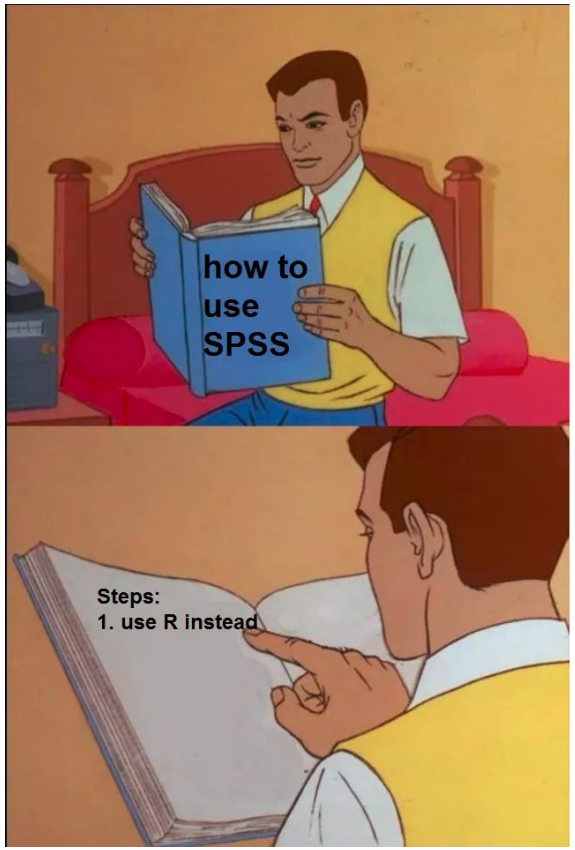
```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

```r
# Establishing a random sample
sam <- sample(pop, 5, replace = FALSE)
sam
```

```
## [1]  1  5 15  9 10
```

The Power Of R

1. **R** is a powerful **statistical** and **graphical tool**
2. Available for almost every platform (Windows, Linux, Mac, FreeBSD, etc.)
3. It is **completely free**
4. **Open source**
   - It can be modified heavily to suit individual demands
   - Constant, moderated user input to widen functionality
   - Dedicated, heavily frequented forums online
   - Allows for reproducible coding



R is the rising star of statistical applications in biological sciences!

Obtaining R

R is a free statistical environment that is used by many researchers all around the globe.

How to get it?

- R is available at https://www.r-project.org/
- A host of editors is available freely on the internet. I recommend RStudio (available at https://www.rstudio.com/).

What if I need help?

Multiple dedicated forums online:

- https://stackoverflow.com/
- https://stackexchange.com/

Layouts - The Console

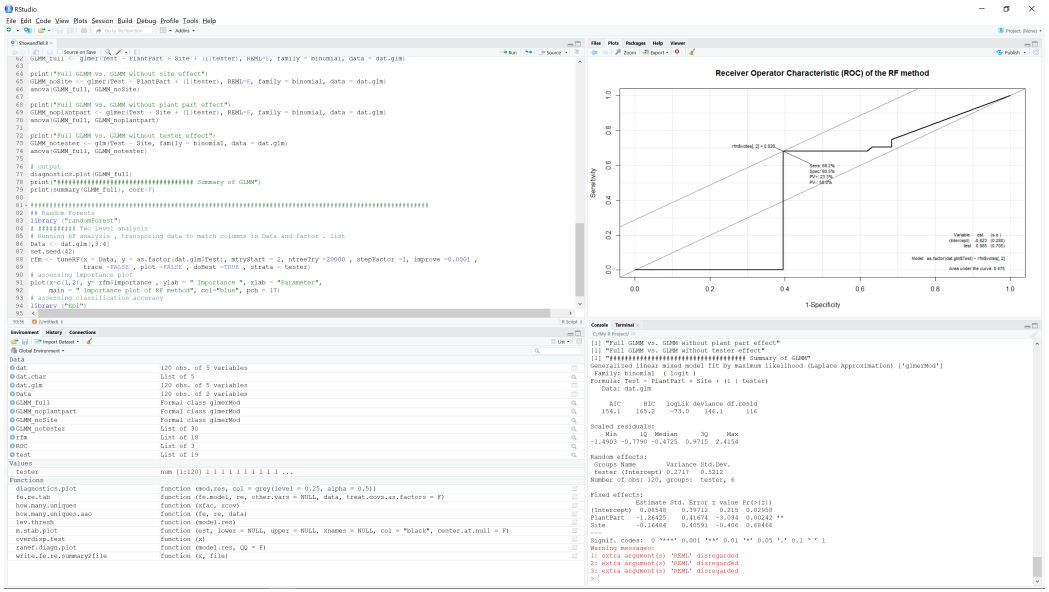
Running R through the console . . .

... is a **bad idea.**

But you will have access to it anyway as it comes with R (we will use version 3.4.2. https://cran.r-project.org/bin/windows/base/old/3.4.2/).

Layouts - The Editor

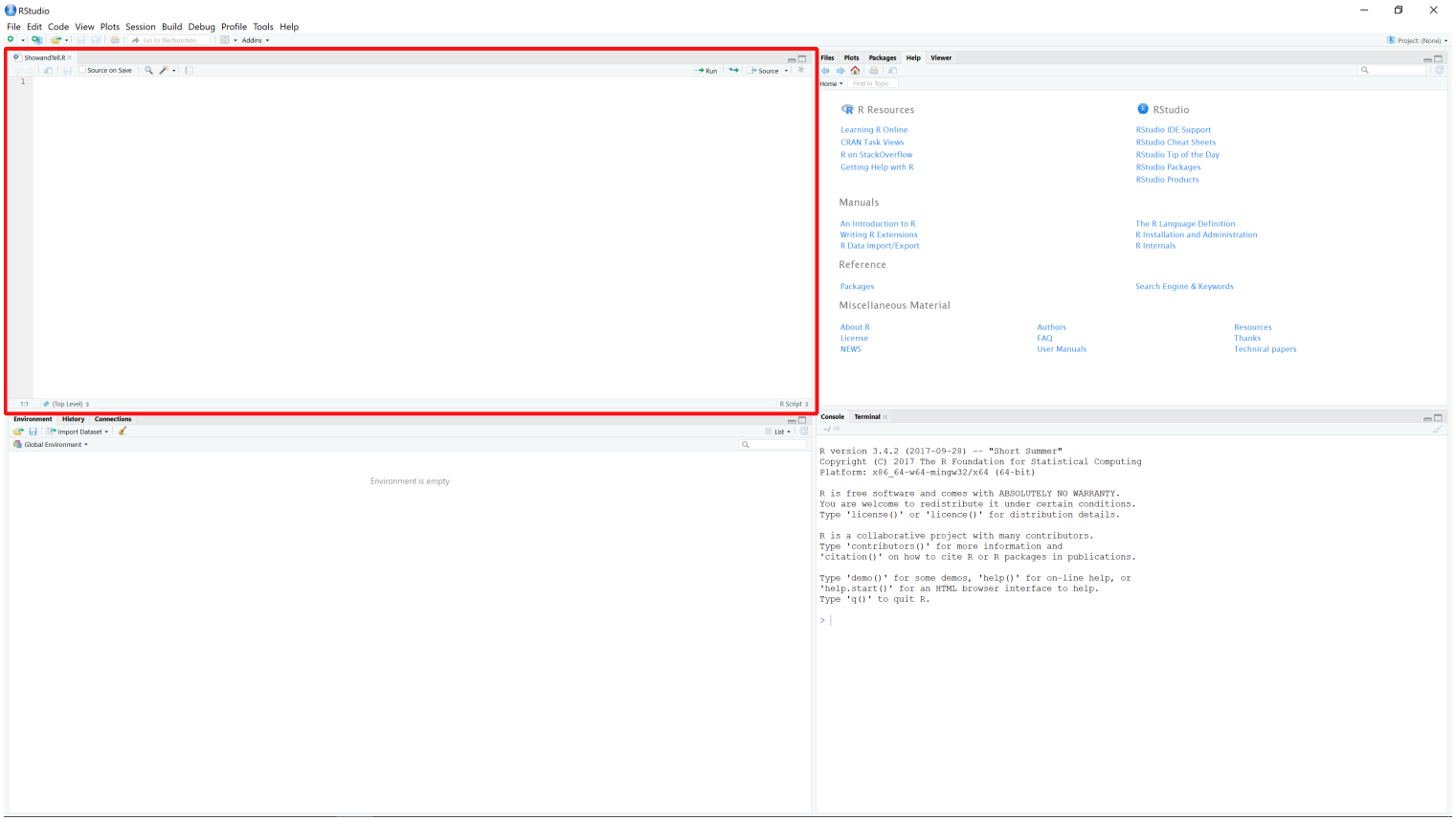
Running R through an editor. . .



... is a **much better idea!**

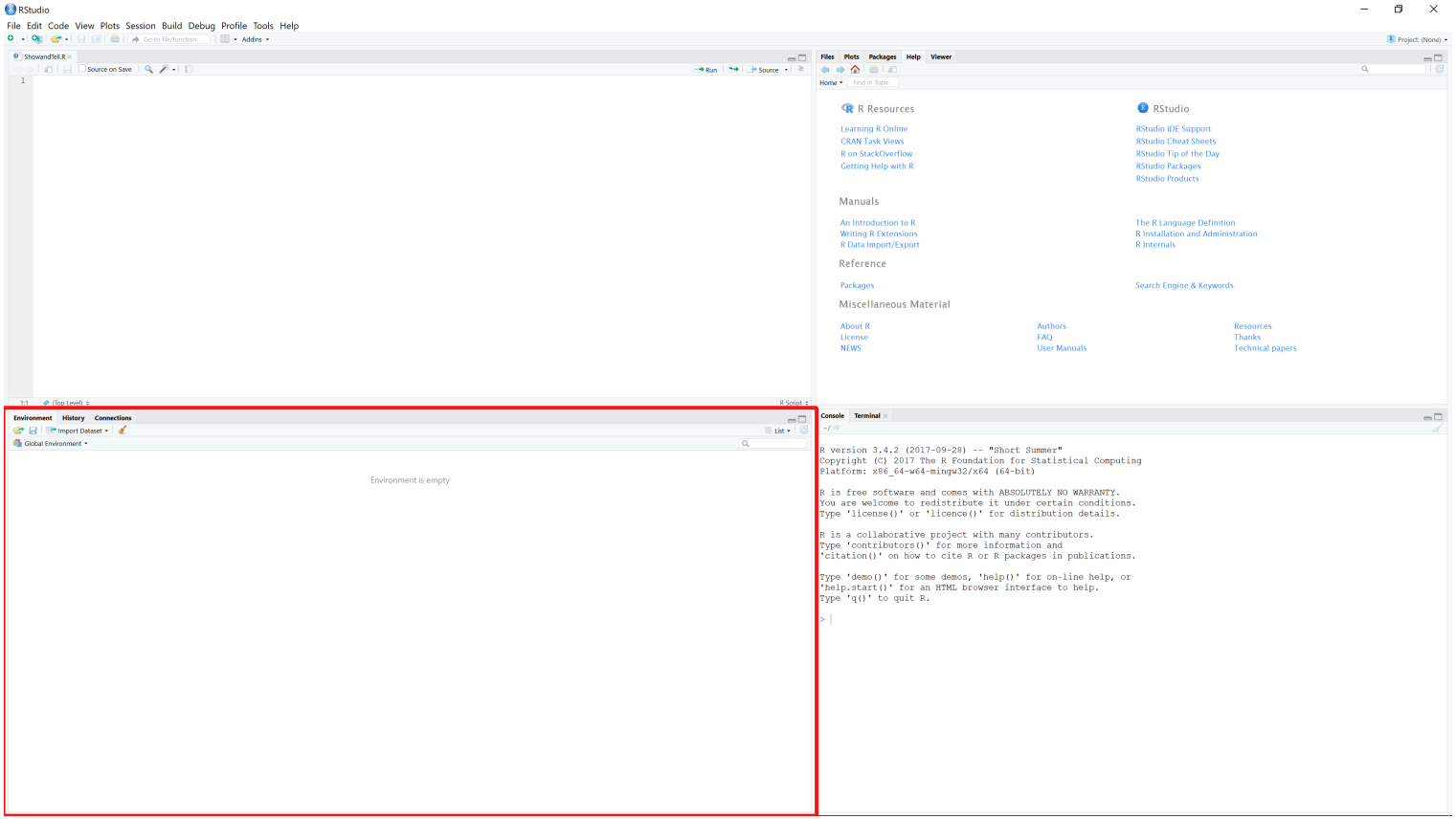
I recommend RStudio (https://www.rstudio.com/). If you use it a lot, I also recommend changing the appearance to 'Vibrant Ink' (setting located in the 'Global Options' window nested within the 'Tools' tab).

Layouts - The Editor Explained



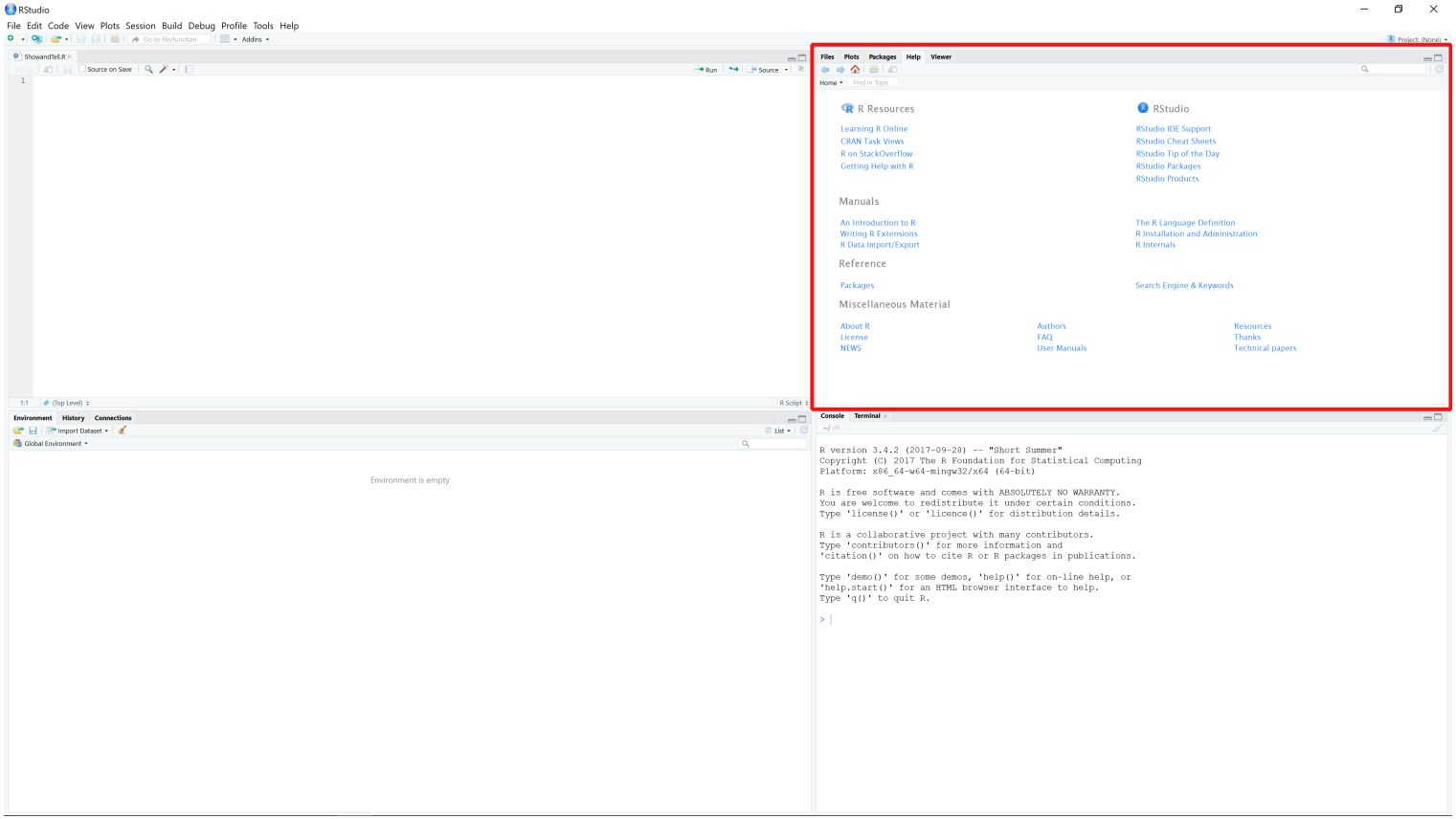
The **Source** is where you load scripts and write most of your coding document.

Layouts - The Editor Explained



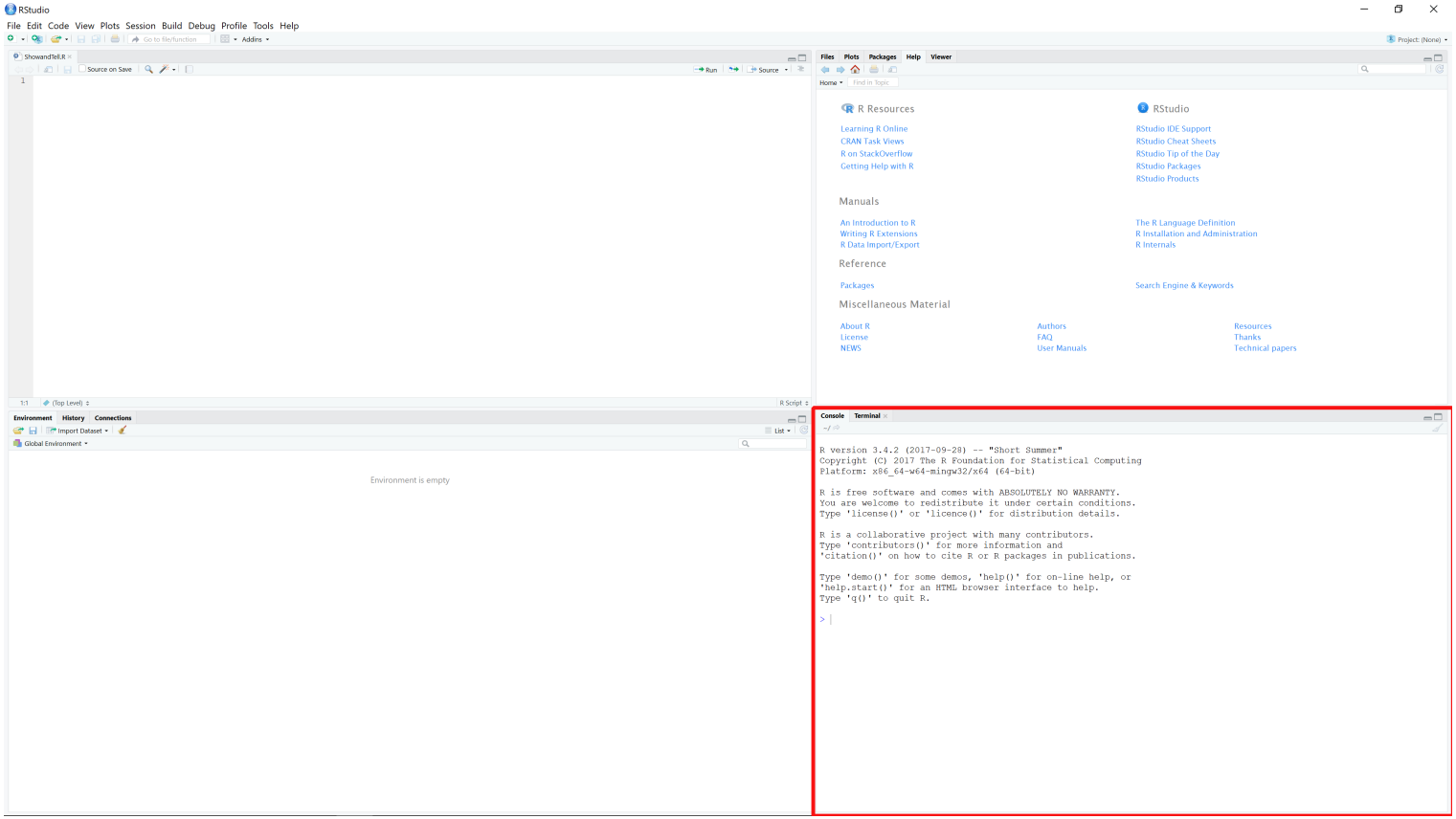
The **Environment, History, Connections** is where you will be able to quickly access all objects of your current R session.

Layouts - The Editor Explained



Files, Plots, Packages, Help Viewer are especially useful for document navigation, data visualisation and to get information on certain functions in R.

Layouts - The Editor Explained



The **Console** is where you execute short commands, and warning and error messages are displayed.

The Evolution Of Code

- Your code and coding practices evolve
- Comment **every line** of code
- Elegant code makes an analysis easier to reproduce
- **Avoid hard-coding!**

"If it looks stupid but it works, it isn't stupid."