

BASIC STATISTICS FOR BIOLOGISTS



UNIVERSITÄT
LEIPZIG

Erik Kusch

erik.kusch@uni-leipzig.de

Behavioural Ecology Research Group
University of Leipzig

02/04/2019

1 What To Expect

- The Seminars
- Your Tutor

2 The Importance Of Proper Statistics

- The Consequences Of Bad Statistics
- What Are Bad Statistics?
- Statistical Concern On The Rise
- Further benefits of a statistical background

3 Terminology

- Classifying Statistics
- Basic Vocabulary

4 Introduction To R

- Why Use R?
- The R landscape
- Layouts
- Coding

Course Dates & Outline I

Block I - Theory and Basics of R

Date	Time	Topic	Location
I.) Introduction			
02/04/2019	15:00 - 16:30	(1) An Introduction to Basic Statistics for Biologists	CIP-POOL
09/04/2019	15:00 - 16:30	(2) Introduction to R	CIP-POOL
II.) Basic statistical terminology			
09/04/2019	15:00 - 16:30	(3) A Primer for Statistical Tests	CIP-POOL
14/04/2019	15:00 - 16:30	(4) Descriptive Statistics	CIP-POOL
23/04/2019	15:00 - 16:30	(5) Data Visualisation	CIP-POOL
30/04/2019	15:00 - 16:30	(6) Inferential Statistics, Hypotheses and our Research Project	CIP-POOL

Course Dates & Outline II

Block II - Basic Statistics in R

Date	Time	Topic	Location
III.) Handling Data			
07/05/2019	15:00 - 16:30	(7) Data Handling and Data Mining	CIP-POOL
IV.) Non-parametric tests			
14/05/2019	15:00 - 16:30	(8) Nominal Tests	CIP-POOL
21/05/2019	15:00 - 16:30	(9) Correlation Tests	CIP-POOL
28/05/2019	15:00 - 16:30	(10) Ordinal and Metric Tests for two-sample situations	CIP-POOL
04/06/2019	15:00 - 16:30	(11) Ordinal and Metric Tests for more than two-sample situations	CIP-POOL
V.) Parametric tests			
02/07/2019	15:00 - 16:30	(12) Simple Parametric Tests	CIP-POOL
VI.) Closing			
Handout		(13) Summary and an Outlook on Advanced Statistics	

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
(located here:
<https://github.com/ErikKusch/An-Introduction-to-Biostatistics-Using-R>)

When prof shows you how to do analysis on SPSS/R/MATLAB but first shows you by-hand theory so you can get a "conceptual understanding" first



We will focus heavily on actually doing the statistics!

Let Me Introduce Myself

Erik Kusch

Studies:

M.Sc. Studies @ Universitetet i Bergen
Research Assistant @ Universität Leipzig

Research:

- Large-scale vegetation-climate modelling
- Remote sensing approaches in landscape ecology
- Biostatistical approaches in behavioural ecology



Statistics (B.Sc.)

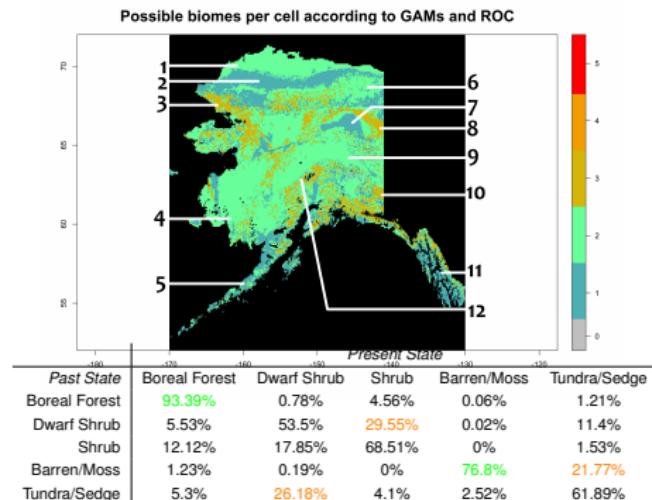
B.Sc. at Technische Universität Dresden:

- Course "Biostatistik für Biologen" by *Dr. Matthias Rudolf*
- first introduction to biostatistics

- B.Sc. thesis on landscape-scale vegetation dynamics based on remote sensing data with:

- *Dr. Alistair Seddon* (Universitetet i Bergen)
- *Prof. Dr. Klaus Reinhardt* (Technische Universität Dresden)

→ first steps in R!



Statistics (M.Sc.)

M.Sc. at Universitetet i Bergen:

- Bayesian **Summer School** by *Dr. Joseph Chipperfield & Partners*
- **Course** "Statistical Learning" by *Dr. Yushu Li*
- **Course** "Ordination and Gradient Analysis" by *Dr. Richard Telford*
- **Course** "Biological Data Analysis II" by *Dr. Richard Telford*
- **Course** "Biostatistics" by *Dr. Knut Helge Jensen*

- **M.Sc. thesis** on vegetation memory effects on landscape-scale and the link with plant functional traits with:
 - *Dr. Alistair Seddon* (Universitetet i Bergen)
 - *Dr. Richard Davy* (Nansen Environmental and Remote Sensing Center)
- **Developing my grasp on statistical theory and practice in R**

Statistics (**Research Assistant**)

Research Assistant at University of Kyoto:

- Analysing behavioural interactions of Japanese macaques using complex social networks

Research Assistant at Universität Leipzig:

- Generating data with various implemented odour effects on which to test the methods
- Analysing primate odour profiles
- Assessing a variety of statistical methods for analytical power in identifying odour effects
- Teaching the basics of biostatistics (this is where you come in)
 - **Putting experience in statistics and R to the test**

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.

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



Original Article

Sampling t_i swabs Sam volatiles

Claudia Birkenmeyer,
Marlen Kückelhoven^a

^aInstitute of Mass Spectrometry, Institute of Applied Chemistry, Faculty of Chemistry and Biochemistry, University of Leipzig, Linienstr. 2, 0410 Leipzig, Germany; Research Group of Behavioral Ecology, Institute of Zoology, University of Leipzig, Linnéstr. 13, 0410 Leipzig, Germany

Department of Archaeology, University College London, Gower Street, London WC1E 6BT, UK; Marine Research Group of Private Mrs. Seelmann, Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Döhmenweg 12, 04323 Leipzig, Germany; and Turner Center for Integrative Biomedicine (TCIB) Döbnerstrasse 17, 04329 Leipzig, Germany

*Correspondence to: Dr. Marlen Kückelhoven, Institute of Applied Chemistry, University of Leipzig, 0410 Leipzig, Germany. E-mail: marlen.kueckelhoven@uni-leipzig.de

Accepted 23 March 2016

Abstract

The most used and safe sampling method for volatile organic compounds (VOCs) is cotton swabbing. However, the use of cotton swabs has been questioned.

Therefore, we conducted a critical analysis of the sampling method and found that it is required in a recent study.

Based on our observations, we conclude that these measurements are not reliable enough and that the use of needles is general.

Key words: body odor, GC-MS, volatile organic compounds

Introduction

In recent years, it has been primarily widely used for off-gas measurements (e.g., Palusz et al.

© The Author(s) 2016. Published by Oxford University Press, 0962-4928/16/010001-06 \$12.00 © 2016 The Authors. *Journal of Chemical Sensors* published by Oxford University Press, 0962-4928/16/010001-06 \$12.00

Original Article

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

Claudia Birkenmeyer,¹ Ruth Thomann,² Fabrice Henig,³ and Anja Widdowson^{4,5}

¹Research Group of Mass Spectrometry, Institute of Applied Chemistry, Faculty of Chemistry and Biochemistry, University of Leipzig, Linienstr. 2, 0410 Leipzig, Germany; Research Group of Behavioral Ecology, Institute of Zoology, University of Leipzig, Linnéstr. 13, 0410 Leipzig, Germany

²Department of Archaeology, University College London, Gower Street, London WC1E 6BT, UK; Marine Research Group of Private Mrs. Seelmann, Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Döhmenweg 12, 04323 Leipzig, Germany; and Turner Center for Integrative Biomedicine (TCIB) Döbnerstrasse 17, 04329 Leipzig, Germany

*Correspondence to: Dr. Marlen Kückelhoven, Institute of Applied Chemistry, University of Leipzig, 0410 Leipzig, Germany. E-mail: marlen.kueckelhoven@uni-leipzig.de

Accepted 23 March 2016

Abstract

We evaluated the suitability of a frequently used sampling method employing cotton swabs for validating animal body odor for gas chromatography-mass spectrometry (GC-MS) analysis of volatile organic compounds (VOCs). Our critical validation showed that both sampling methods and analytical methods were not suitable for this purpose. Therefore, we would like to encourage sample managers to use a more appropriate sampling method.

“Please Janusz! In March include a note in the manuscript that we are not living in an era of difference but in an era of sameness.”

First we used the first sample manager’s example sample manager 1 (SM1) and pure viscous ketone (ketone).

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

Introduction

In recent years, it has become increasingly evident that substances present in the animal body odors are not the exception, rather than the norm. This is particularly true for primates. For example, the smell of primate stool before use and extracted with benzene had the least basal interference in GC-MS analyses (Birkenmeyer et al., 2016). Therefore, we conclude that the use of cotton swabs to provide evidence of body odor samples is not appropriate. The use of needles to obtain samples is recommended.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

“Normalized” equalized test sample was an aqueous solution of 100% benzene and 100% hexane.

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.



DAILY REPORT

Early report

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

A.J. Wakefield, B. H. Murch, A. Anthony, J. Linton, D. M. Caspary, M. Scott, M. Everett, A. P. Dhaliwal, M. A. Thomson, P. Harvey, A. Valentejo, E. E. Davies, J. A. Walker-Smith

Summary

Background We identified a cluster of cases of ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.

Introduction

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



→ 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/Mal informed Use

Lack of statistical knowledge

- Applying statistics to data which they aren't meant for
→ *Methods can “break”*
- Flawed understanding of the methodology
→ *Incorrect conclusions*

Pure biologists lack knowledge on statistics.

Uninformed Use

Lack of biological knowledge

- Delineation of nonsensical but statistically significant relationships
→ *p-hacking*
- No sense of how to establish testable, feasible hypotheses
→ *Waste of time*

Pure statisticians lack knowledge on biology.

Caveat

- Biologists often have preformed ideas of what to expect
→ data-tweaking to match expectations?
 - Researchers also have a vested interest in uncovering extraordinary things
→ 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

第31章

ATTENTION!

Don't let a personal bias inform your analysis!

Why Keep Up With It?

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



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, the summer of 2014, one of us (D.B.A) read a research paper in a well-known

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 overestimated effects by more than tenfold. Van and others submitted a letter to the editor explaining the problem. Months later, the

were gratified to learn that the authors had elected to retract their paper. In the face of popular articles proclaiming that science

Sadly, in our experience, the case is not representative. In the course of our

representative. In the course of assembling weekly lists of articles in our field, I 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

does 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

We learned that post-publication correction.

Advancing In Statistics

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

~ *Andrew Vickers*

Teaching statistics



Doing Statistics



The Lack Of Biostatisticians

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

→ "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 X*, and *responses 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*.

→ **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

```
# Making it reproducible
set.seed(42)

# Establishing a population
pop <- c(1:15)
pop

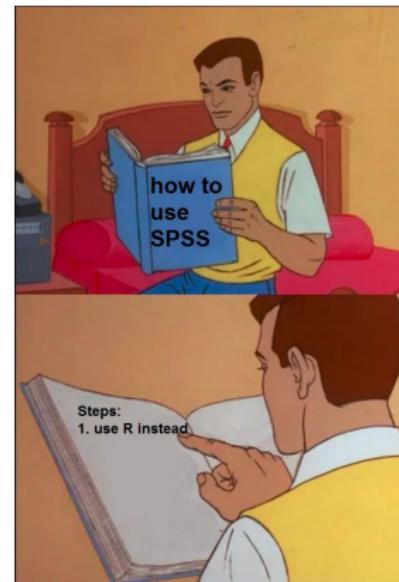
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

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

## [1] 14 15 4 10 8
```

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

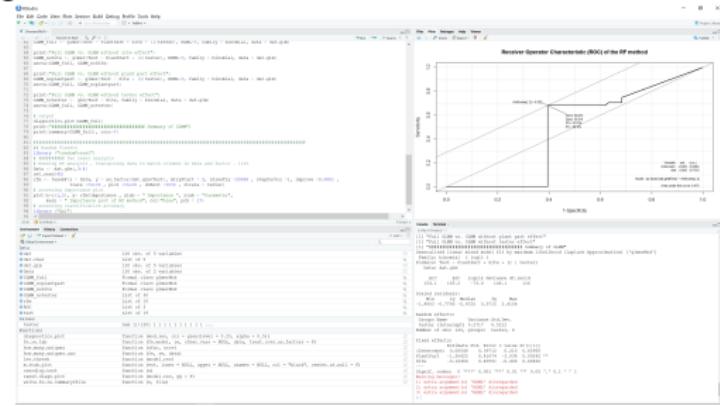
Running R through the console . . .

... 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

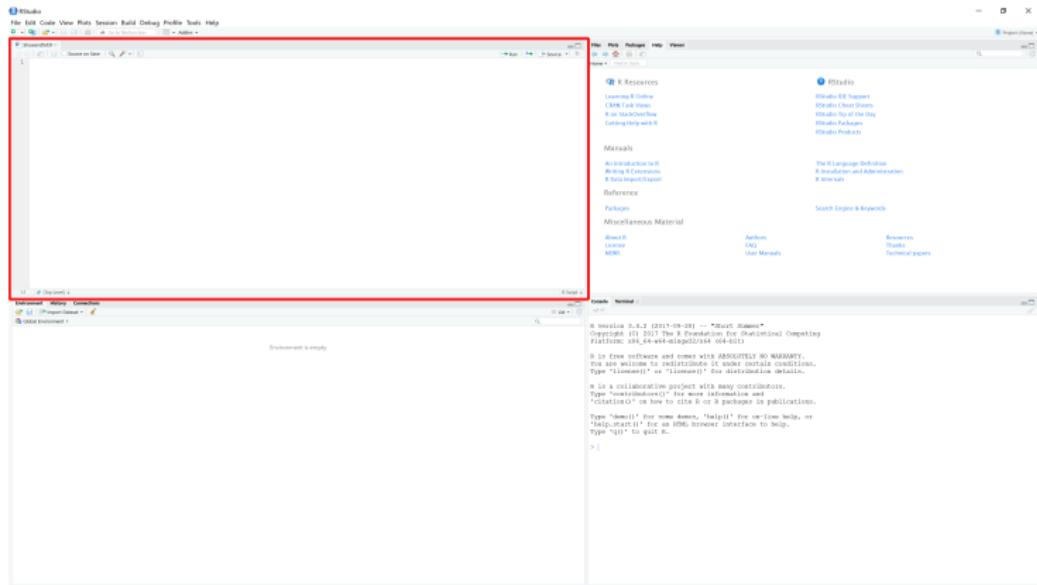
Running R through an editor...



... is a much better idea!

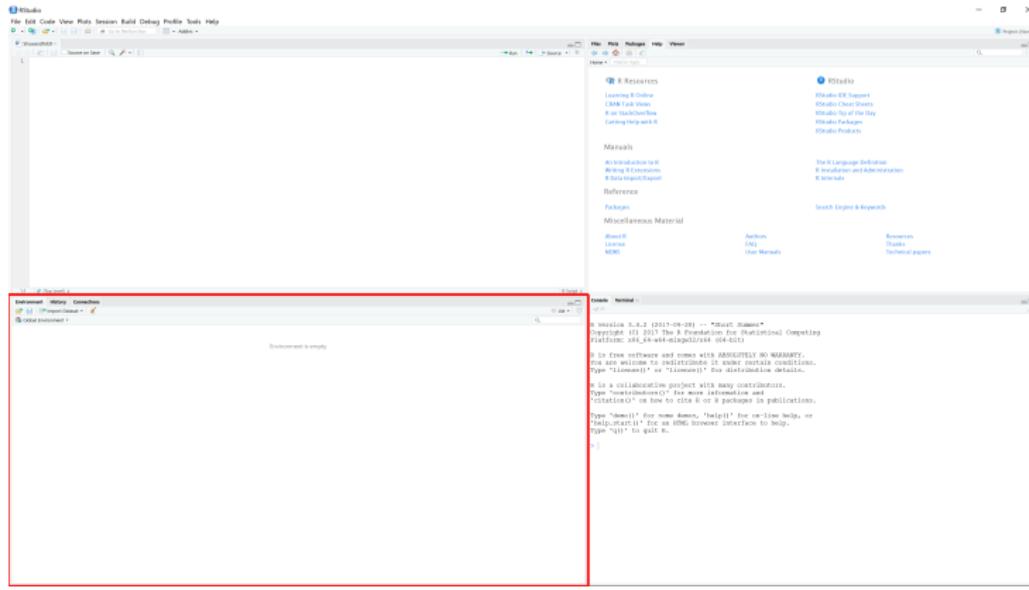
I recommend RStudio (<https://www.rstudio.com/>). If you use it a lot, I also recommend changing the appearance to ‘Vibrant Ink’ (setting located in the ‘Global Options’ window nested within the ‘Tools’ tab).

Layouts - The Editor Explained



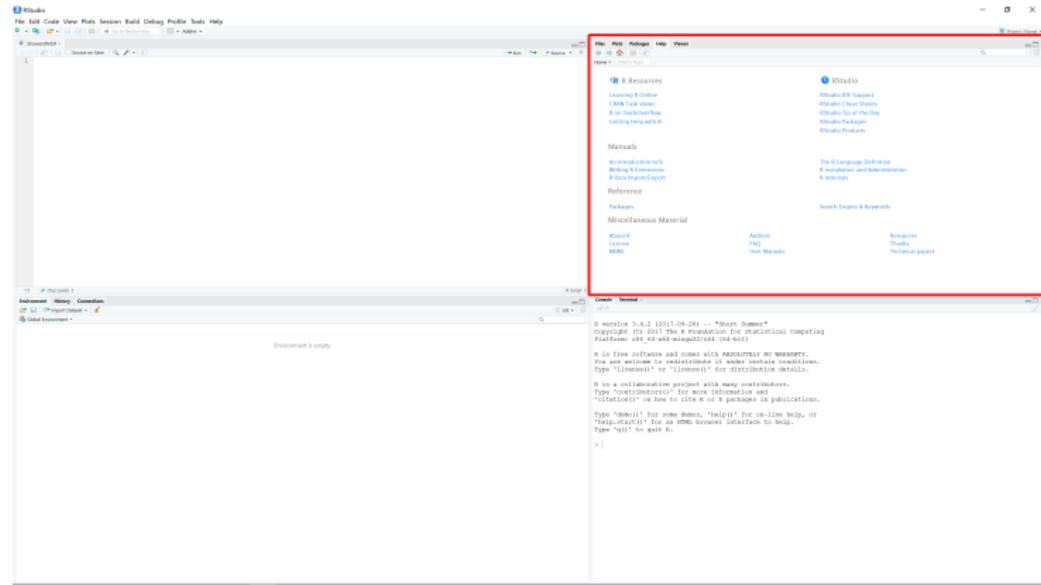
The **Source** is where you load scripts and write most of your coding document.

Layouts - The Editor Explained



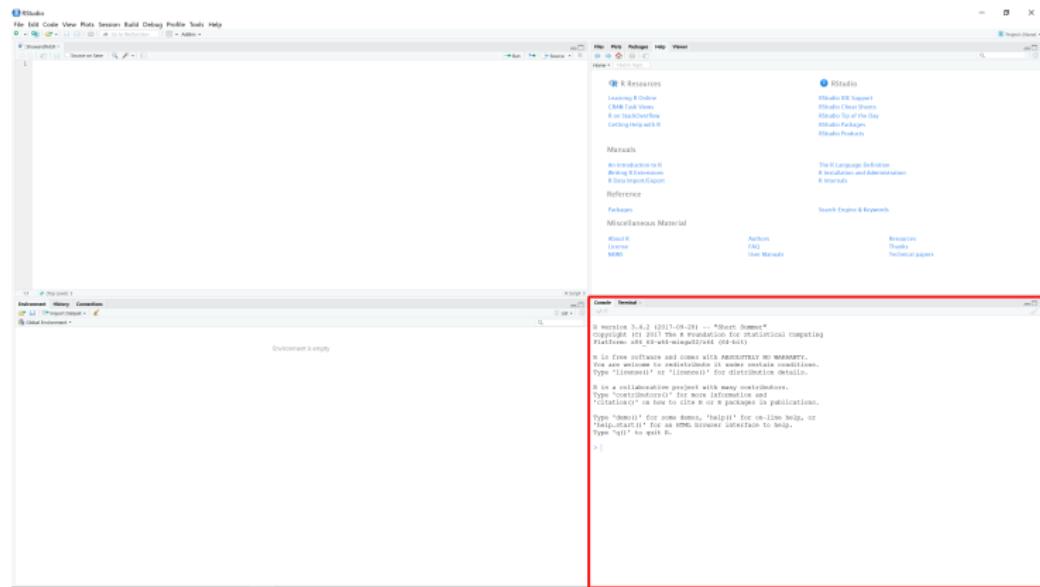
The **Environment, History, Connections** is where you will be able to quickly access all objects of your current R session.

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."

