MEDB 5501, Module01

2024-08-20

Topics to be covered

- What you will learn
  - About this class
  - R and RStudio
  - History of R
  - Scales of measurement
  - Tests of hypothesis
  - Confidence intervals
  - Grading rubric
  - A simple R program

Welcome to the class

- Three sections
  - 0001, Synchronous Zoom meetings on Tuesdays
  - 0002, Asynchronous
  - 0003, International students
- Introductions
  - Ricardo Moniz
  - Suman Sahil

Speaker notes

There are three sections in the class.

Section 0001 is a synchronous online class with lectures delivered via Zoom on Tuesdays from 1:00pm to 3:45pm. Students in this section are required to attend that class. If you are in this section, you must show up on Tuesday afternoons. Attendance will be taken. Two or more unexcused absences will require a discussion with the instructor and remedial actions may need to be taken.

Section 0002 is an asynchronous online class where you watch videos of the zoom session. Students in this section have the option of attending the Tuesday 1pm Zoom sessions, if their work schedule permits.

Section 0003 is a synchronous online class like section 0001 with the additional requirement of weekly sign-ins. The weekly sign-in meets an important requirement of some international students. Other than the sign-in, the class is identical to the synchronous online class.

All three sections are combined in Canvas to a single site. This makes it easier for me to make announcements, do grading, etc. Just because the Canvas site says “2024FS-MEDB-5501-0001” does not mean that you are in the wrong Canvas site and it does not mean that you are registered for the wrong class. We’re all one big family from the perspective of Canvas. Only the attendance requirements differ.

Requirements of all students

- Attend Tuesday Zoom or watch video of Tuesday Zoom
- Read book chapter
- Optional review session on Zoom on Fridays
- Complete all assignments by Monday at 11:59pm

Speaker notes

I’ll record the Tuesday 1pm meeting and make it available on Canvas by noon on Wednesday for the asynchronous students.

Everyone should also read the appropriate Chapter from the book.

Everyone is invited to an optional review session on Fridays from 10:30am to 11:30am. You can ask questions, or just sit back and listen to questions that others may have. If you have work commitments on Friday mornings, I am still happy to meet with you at any other reasonable time.

The homework assignments, quizzes, and discussion boards are due by 11:59pm on Mondays.

Attendance

- Synchronous students
  - MUST attend Tuesday Zoom sessions.
  - can also review the recordings
- Asynchronous students
  - watch the Tuesday recordings
  - or attend some of the Tuesday Zoom sessions
  - or both
- Failure to attend is a problem

Speaker notes

Attendance at the Tuesday Zoom sessions is a requirement for all synchronous students. You can also watch part or all of the recordings of the Tuesday Zoom sessions if you like.

Asynchronous students have an option that synchronous students do not.

For synchronous students to attend the Tuesday Zoom sessions is a problem. For asynchronous students, failure to either attend or watch the videos is a problem. If you have work commitments, health issues, or personal issues that prevent you from meeting the attendance requirements, you must contact us, preferably beforehand.

Assignments

Speaker notes

Assignments will include homework submissions, quizzes, and discussion boards

This class is in transition

- Taught for many years by Dr. Monica Gaddis
- My second time teaching this class
- Major changes
  - Software agnosticism
    - In class switch from SPSS to R
  - Suman Sahil will co-teach and cover programming in R
  - Discussion boards ask for feedback
  - Proposed exam questions

Speaker notes

I’m excited at the opportunity to teach this class, but I am also a bit scared. The topics in this class are relatively easy, but teaching them well is not so easy.

This class has been taught for many years by Dr. Monica Gaddis. She is an excellent teacher, but she retired this summer. I have a lot of respect for her and have kept a lot of her material. There are some important differences, however. I am not a wall flower, and I have a few ideas that might supplement an already good class.

I believe in software agnosticism. You should use whatever software you thing is best for you. Use the software that you expect to use in the real world after you graduate. It’s more work for me (and for Suman Sahil, who has to grade the various programs). Nevertheless, I want to make that effort because I should not presume to know what program is best for you and your budding career.

All the classroom examples will use R. If you use another program, expect to spend some time reading the documentation and figuring out how to convert the R code to code for your system. You can ask for assistance by email, during the help session if all the other questions are already answered, or by special appointment. I take great pride in knowing how to do things in a broad range of statistical software and should be able to help with most issues.

If you’ve never used any statistical software before, I would certainly encourage you to use R.

A second change is that I will include a brief feedback assignment after every module. I want to feedback on what you thought was most important, what you found most confusing, and what you might have liked to learn that wasn’t in the lecture.

Don’t spend a lot of time on the feedback. You are welcome to say something brief like nothing was confusing. If everything was confusing, that’s fine also, but please feel free to show up at the Friday review session or set up a special appointment if this is the case.

You are also welcome to say something like “I agree with” and mention the name of a previous student.

I might make some other small changes over time.

There’s a risk in making changes. It may create some inconsistencies between what I am asking you to do and what Dr. Gaddis covered in her videos. Use a bit of common sense, but if you are not quite sure about anything, please ask.

Why R

- Faculty consensus
  - Prepares you better for capstone/thesis work
  - More likely to be used in your future job
  - Integrates well with team based tools
- R is not as difficult as claimed
  - Work from existing templates
- Python, SAS, Stata would have also been good choices
  - You are welcome to use these if you like

Speaker notes

Last year, I taught the class using SPSS and it was an okay choice. But some of the faculty raised concerns about leaving students ill prepared if they needed to do data analysis for a capstone project or thesis. SPSS is okay for introductory statistics, but falls short for the sort of work needed in a thesis.

While some jobs out in the real world will use SPSS, far more will use R.

Most important, in my opinion, is that R integrates well with team-based tools like version control, cloud-based computing, distributed databases, and other methdologies that are being used more and more with programming teams.

You may have heard that R is a difficult language. R is not as difficult as some people may claim. It is being used successfully at a large number of universities in their Statistics and Data Science program. It is even used extensively in undergraduate classes.

Having said that, I have to admit that R is still not as easy to use as SPSS. I will try to lighten that burden by providing templates that will guide your coding throughout all of the assignments.

Now there are several good alternatives to R. Each has its advantages but also some disadvantages. It is not worth quibbling here. I've chaired a couple of panel discussion sessions at national Statistics conferences on this topic. My recommendation is to use what your boss uses.

If you want to use something other than R, go for it! You'll have to do a bit of work on your own, but I will provide support as needed. I take great pride in being able to work with a broad range of statistical software systems.

Student learning objectives, 1 of 3

- SLO1
  - The graduate will be able to use statistics to analyze and interpret data. They will understand the fundamentals of the field in the context of recognizing the effective use of data or information for the specific discipline(s). They will select and apply appropiate statistical procedures to the information. They will be able to analyze and accurately interpret of statistical result.

Speaker notes

There are five student learning objectives that our department has committed to. We need to demonstrate to the university that anyone who graduates from our department has a particular set of capabilities. This class is one of several that provides you with these capabilities.

The first student learning objective asks that you are capable of using statistics to analyze and interpret data. This is not the only class where you will develop this capability, of course. The key elements are selecting the statistical procedures, running those procedures, and writing about the statistical procedures, and the ability to run those procedures.

Student learning objectives, 2 of 3

- SLO2
  - The graduate will be able to design a testable research question or hypothesis. They will have adequate background knowledge about biological, biomedical, or population health contexts and problems including common research problems in order to generate a research question or hypothesis. They will be able to relate problems within and across levels of areas of the spectrum to bridge disciplines.

Speaker notes

The second student learning objective is the ability to design a testable research question or hypothesis.

Student learning objectives, 3 of 3

- SLO5
  - The graduate will be able to communicate scientific outcomes. This includes the ability to convey scientific methods and statistical findings, effectively field questions in an oral presentation format as well as in the preparation of thesis or capstone manuscripts.

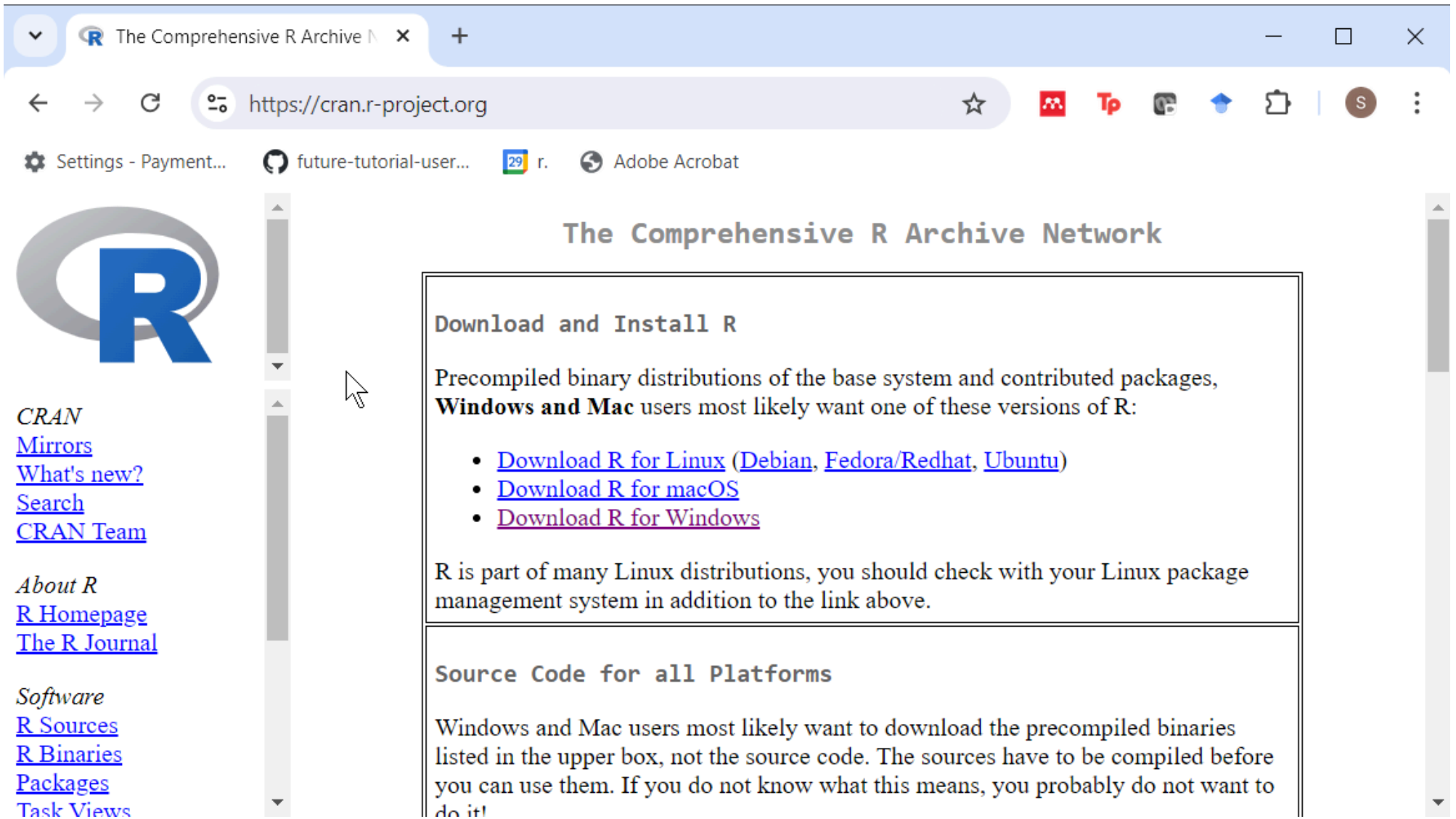
Speaker notes

The third and fourth student learning objectives are developed in other courses. The fifth (and last) learning objective is the ability to convey statistical findings and answer question in an oral presentation. This includes the ability to present your findings in a visual format.

Break #1

- What you have learned
  - About this class
- What’s coming next
  - R and RStudio

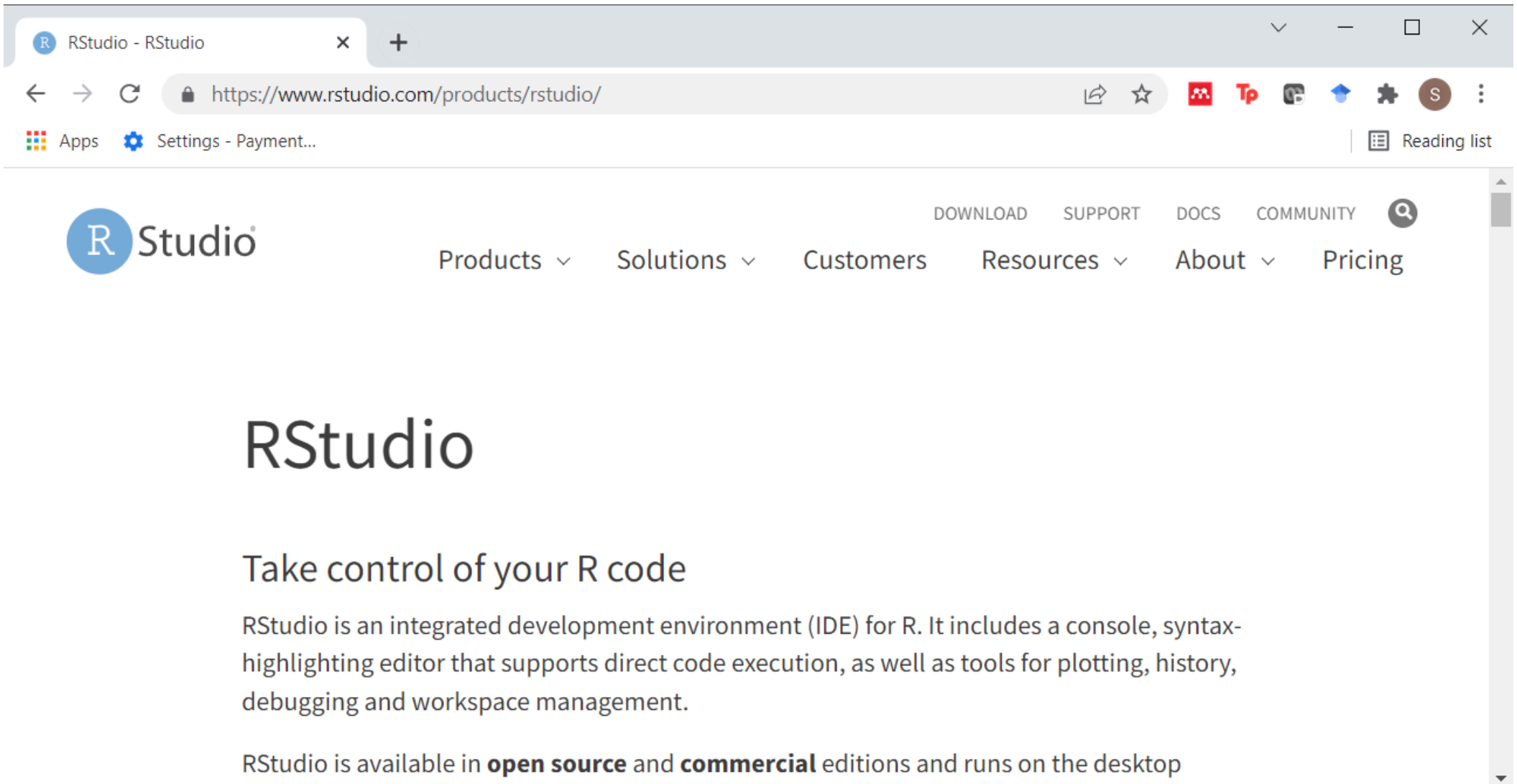
R


The Comprehensive R Archive
https://cran.r-project.org
Settings - Payment...
future-tutorial-user...
r.
Adobe Acrobat
The Comprehensive R Archive Network
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
Download R for macOS
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to
CRAN
Mirrors
What's new?
Search
CRAN Team
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Task Views

Speaker notes

This course will use R for all classroom examples and programming assignments. You can download R from CRAN, the Comprehensive R Archive Network.

RStudio


RStudio - RStudio
https://www.rstudio.com/products/rstudio/
Apps
Settings - Payment...
Reading list
RStudio
DOWNLOAD
SUPPORT
DOCS
COMMUNITY
Products
Solutions
Customers
Resources
About
Pricing
RStudio
Take control of your R code
RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.
RStudio is available in open source and commercial editions and runs on the desktop

Speaker notes

RStudio is an integrated development environment. This provides a window where you edit your code, a window to show your files, a window to display your text output, a window to display your graphs, and other helpful windows.

Other statistical software

- JMP, Python, R, SAS, SPSS, Stata
  - Use only if you are confident in your abilities
  - Avoid Microsoft Excel

Speaker notes

There are lots of other choices. I am happy to help you get started with any of these programs. I like working with software. I should warn you that the level of support that I can provide does vary. I know less about Python and Stata, for example, than I do about SPSS and SAS. But I’m pretty good with anything.

If you do decide to try one of these programs, make sure that you are confident in your own abilities and that you can figure out some stuff on your own. If you are new or relatively inexperienced, SPSS might be a better choice.

Break #2

- What you have learned
  - R and RStudio
- What’s coming next
  - History of R

R sprouted from S

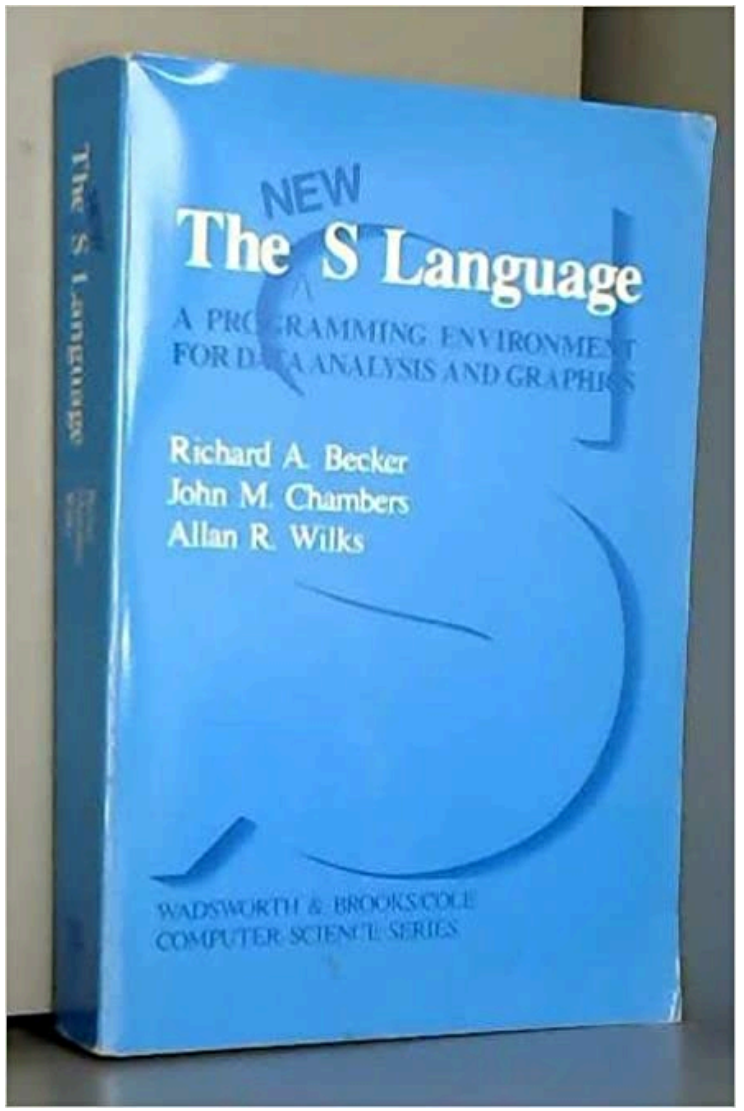


Figure 1. Book cover

Speaker notes

I’m helping to put together three separate classes, Basic data management and analysis with R, SAS, SPSS. As part of these classes, I need to discuss the history of these programs, because understanding that history will help you better understand the strengths and weaknesses of each statistical package. Here’s a brief history of R.

R has its roots in a program called S. S was developed in a time when single letters were in vogue (as in the C programming language).

Image source: Amazon

John Chambers

Figure 2. Photo of John Chambers

Speaker notes

The primary author of the S language was John Chambers. Often he gets sole credit, there were two other major contributors.

Image source: AT&T

Richard Becker

Figure 3. Photo of Richard Becker

Speaker notes

Also involved with S, is another statistician, Richard Becker.

Image source: AT&T

Allan Wilks

Figure 4. Photo of Allan Wilks

Speaker notes

A third author was Allan Wilks.

Image source: AT&T

Bell Labs

Figure 5. Aerial photograph of Bell Laboratories

Speaker notes

All three statisticians worked at Bell Labs. Bell Labs was a research division of AT&T (affectionately known as Ma Bell), back when Ma Bell held a monopoly over telephone service.

Image source: Wikipedia

Features of S.

- Intended for internal use.
- Freely available to anyone.
- Interactive
- Unique capabilities
  - Emphasis on functions
  - Object-oriented features

Speaker notes

The author of S, John Chambers, was a statistician at Bell Laboratories wrote several versions in the 1970s through the 1990s. This packages was intended for internal research use, but the code was freely available to anyone who was interested.

S was an interactive programming language, which made it quite different from other statistical software systems of the times, like SAS and SPSS.

Two unique features of the S programming language were the use of functions rather than macros for extending the language and the introduction of object oriented features (classes, objects, and methods).

S-plus

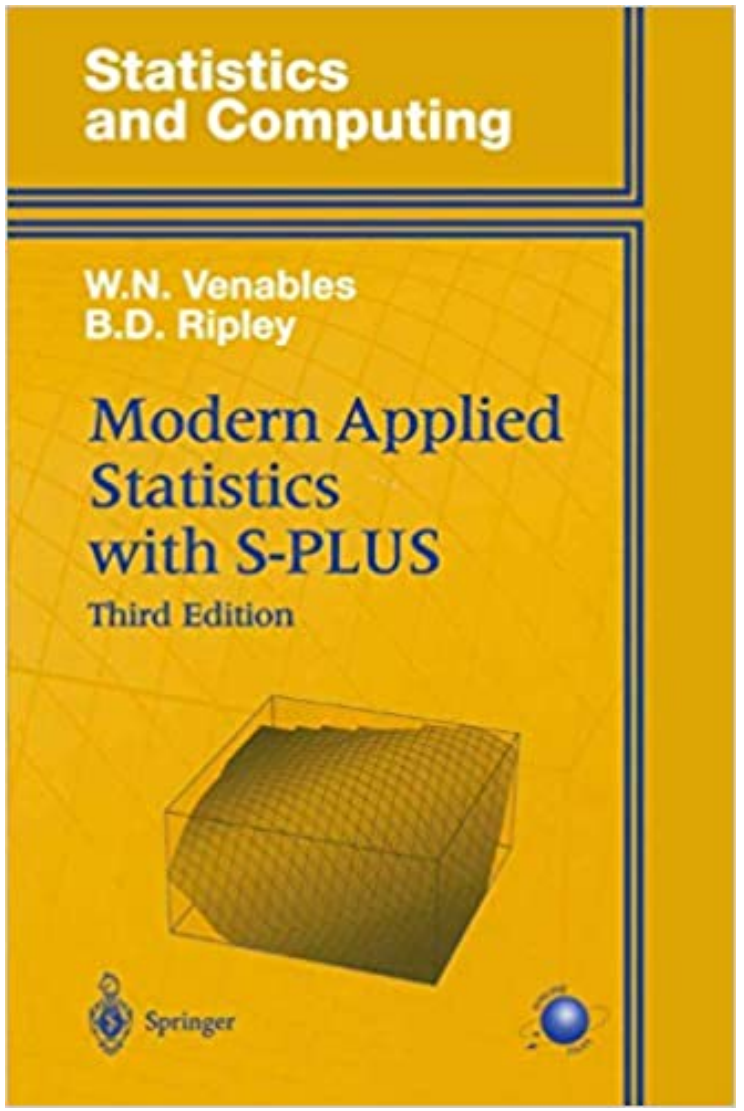


Figure 6. Venables and Ripley book cover

Speaker notes

A commercial adaptation of S was introduced by Statistical Sciences Corporation in the 1990s and became very popular. Through various mergers and buyouts, S+ has been marketed by Mathsoft, Insightful Software, and more recently Tibco Corporation.

Image source: Amazon

Beginnings of R (1/2)

R: A Language for Data Analysis and Graphics

Ross IHAKA and Robert GENTLEMAN

In this article we discuss our experience designing and implementing a statistical computing language. In developing this new language, we sought to combine what we felt were useful features from two existing computer languages We feel that the new language provides advantages in the areas of portability, computational efficiency, memory management, and scoping.

Key Words: *Computer language; Statistical computing.*

Figure 7. Excerpt from research paper

Speaker notes

About the same time, Ross Ihaka and Robert Gentleman started an effort to produce an open source and freely distributed version of S, called R. Their publication:

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299-314, 1996. Available in pdf format.

outlined the features of the R programming language.

Beginnings of R (2/2)

Figure 8. CD of release 1.0 of R

Speaker notes

The first major release of R (version 1.0.0) appeared in 2000.

Note: add image source.

Growth in popularity

Figure 9. Excerpt from New York Times article

Speaker notes

Soon R eclipsed S+ in popularity. One measure of the breadth of R's impact was a New York Times article published in 2009.

Ashlee Vance. Data Analysts Captivated by R's Power. The New York Times, 2009-01-06. Available in html format.

R Foundation

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Get Involved: Contributing
Developer Pages
R Blog

The R Foundation

The R Foundation is a not for profit organization working in the public interest. It has been founded by the members of the R Development Core Team in order to

- Provide support for the R project and other innovations in statistical computing. We believe that R has become a mature and valuable tool and we would like to ensure its continued development and the development of future innovations in software for statistical and computational research.
- Provide a reference point for individuals, instititutions or commercial enterprises that want to support or interact with the R development community.
- Hold and administer the copyright of R software and documentation.

R is an official part of the Free Software Foundation's GNU project, and the R Foundation has similar goals to other open source software foundations like the Apache Foundation or the GNOME Foundation.

Among the goals of the R Foundation are the support of continued development of R, the exploration of new methodology, teaching and training of statistical computing and the organization of meetings and conferences with a statistical computing orientation. We hope to attract sufficient funding to make these goals realities.

The **R Foundation Statutes** can be downloaded as PDF file in English or German.

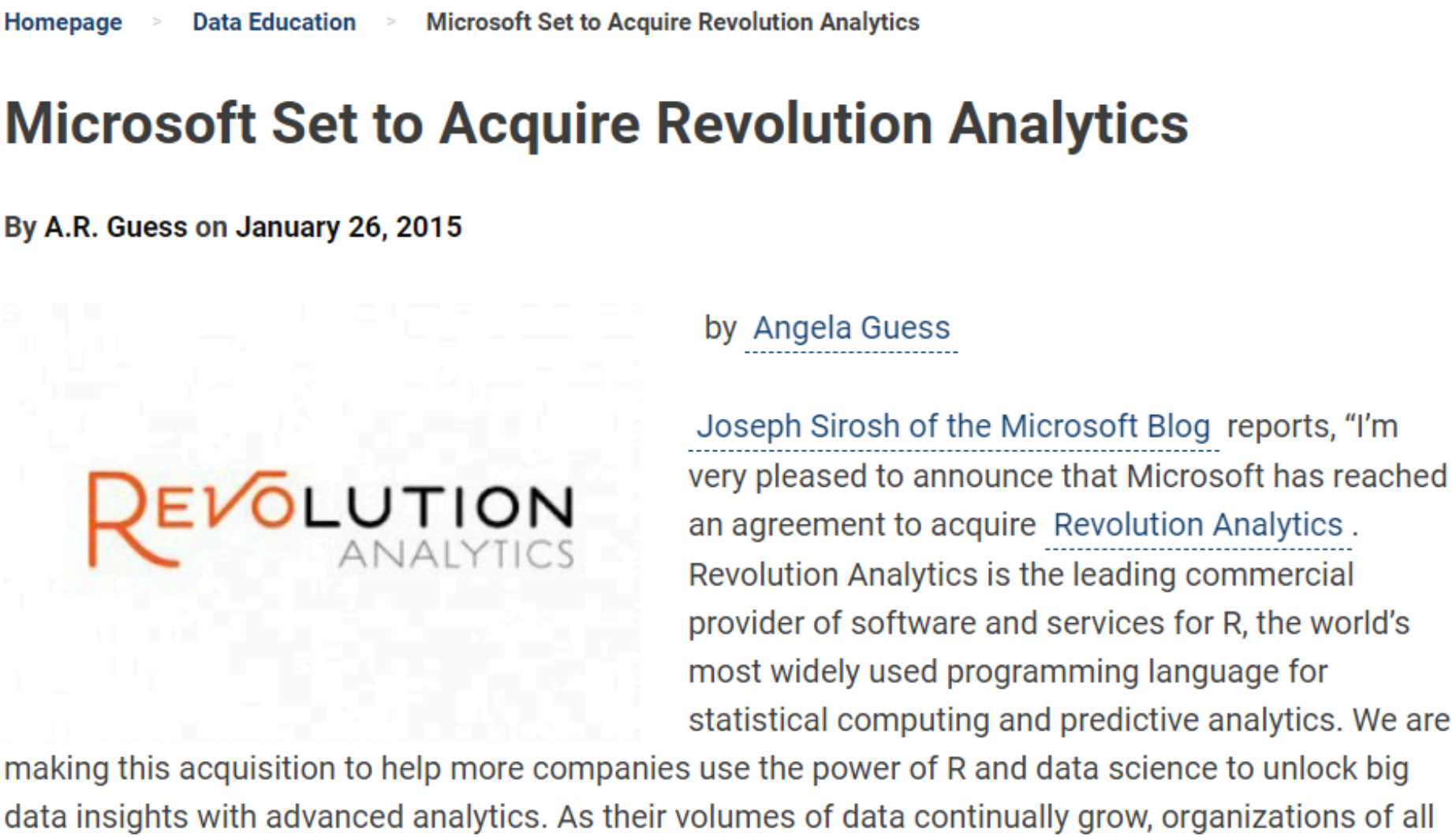
Figure 10. Excerpt from website

Speaker notes

There is a non-profit group, the R Foundation for Statistical Computing, that coordinates many of the efforts in the maintenance and development of the R programming language.

Note: add image source.

Revolution Analytics



Homepage > Data Education > Microsoft Set to Acquire Revolution Analytics

Microsoft Set to Acquire Revolution Analytics

By A.R. Guess on January 26, 2015

by Angela Guess

Joseph Sirosh of the Microsoft Blog reports, "I'm very pleased to announce that Microsoft has reached an agreement to acquire Revolution Analytics. Revolution Analytics is the leading commercial provider of software and services for R, the world's most widely used programming language for statistical computing and predictive analytics. We are making this acquisition to help more companies use the power of R and data science to unlock big data insights with advanced analytics. As their volumes of data continually grow, organizations of all

Figure 11. Excerpt from article

Speaker notes

Several commercial companies have piggybacked on R, including Revolution Analytics, which sells an enhanced version of R with capabilities for handling very large data sets.

Image source: Dataversity

https://www.dataversity.net/microsoft-set-acquire-revolution-analytics/

R packages

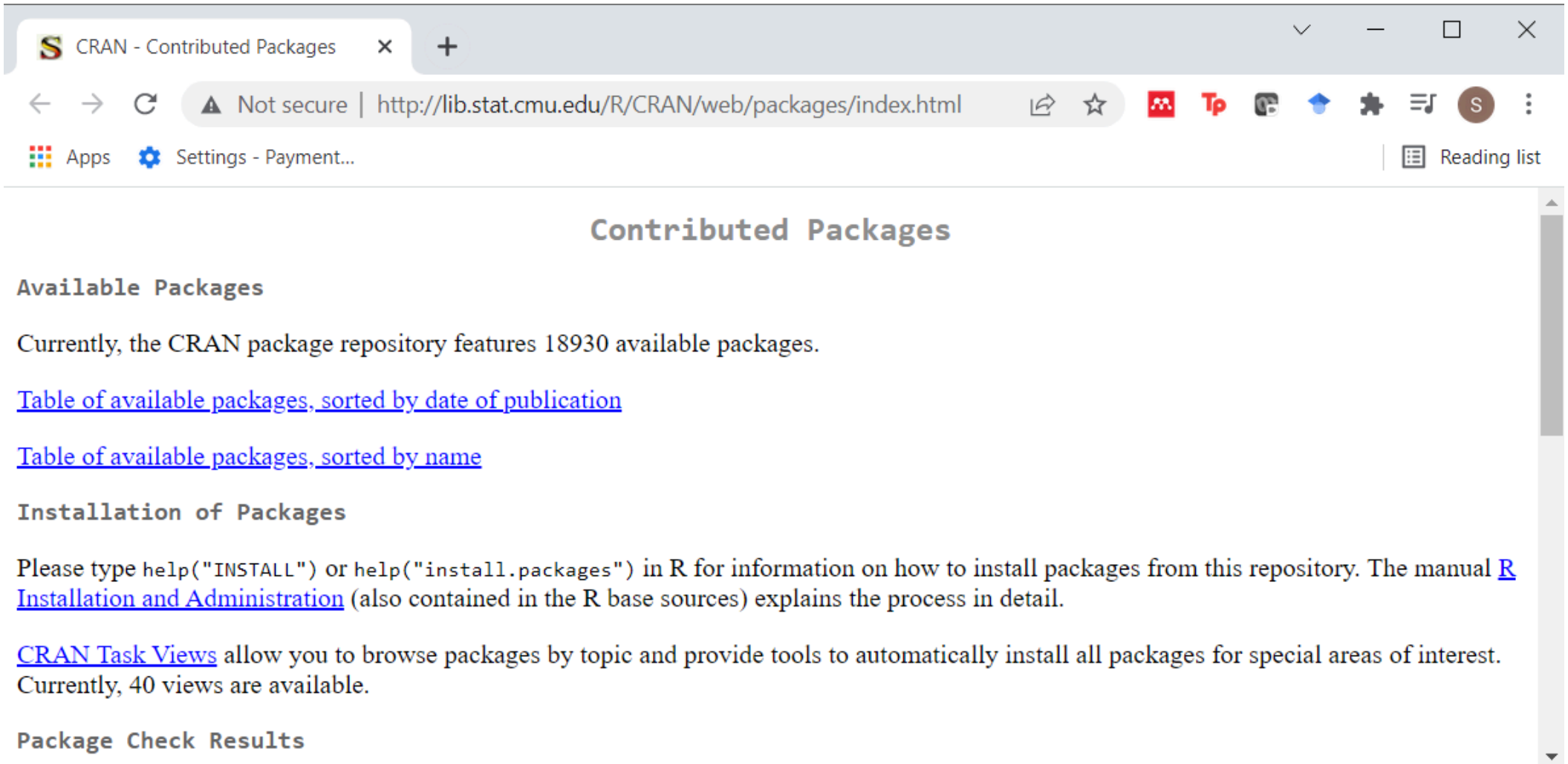


Figure 12. Excerpt from website

Speaker notes

One of the most popular features of R is the ease with which outside developers can extend the R language through libraries. Most of these libraries are available for free under and open source license at the Comprehensive R Archive Network.

Image source: Comprehensive R Archive Network

Bioconductor

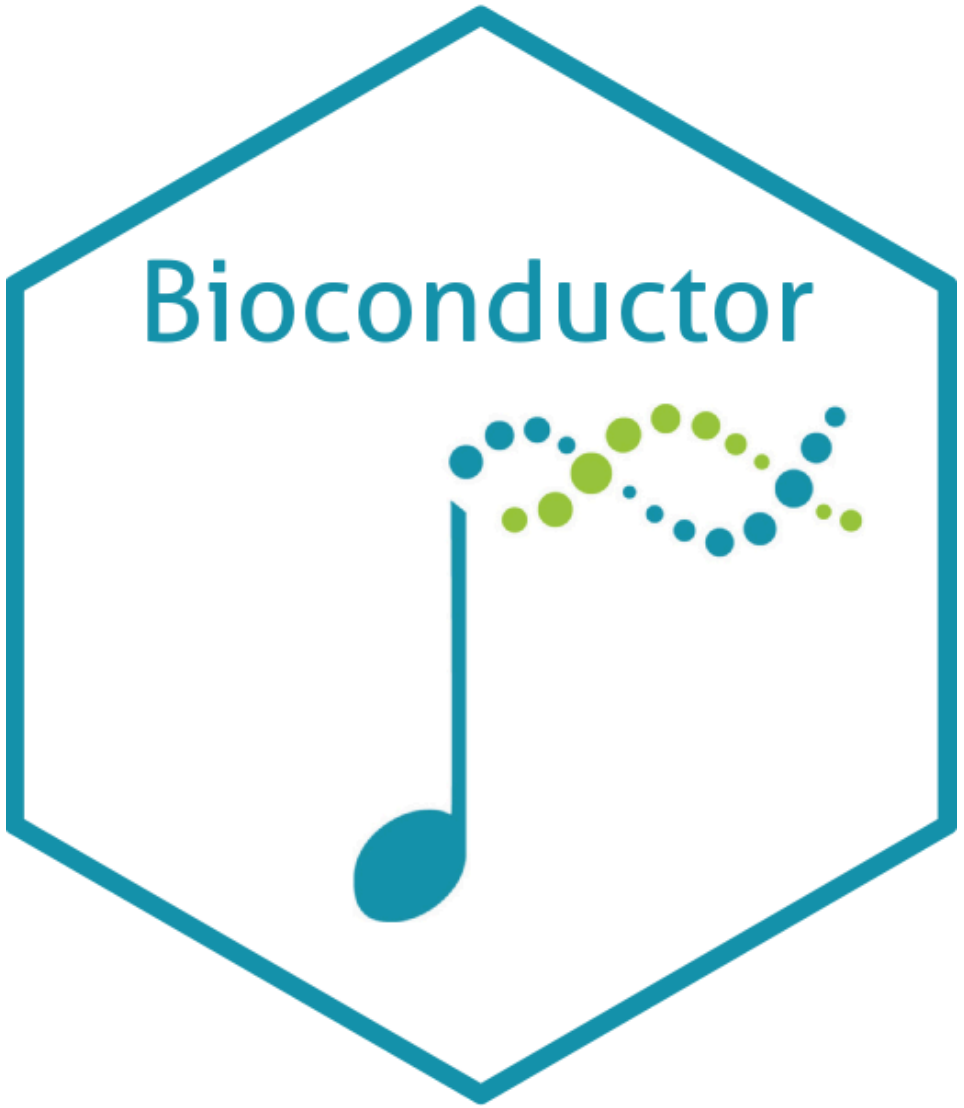


Figure 13. Excerpt from website

Speaker notes

You can also find a major effort to develop freely available libraries for statistical analysis of genetic data through the Bioconductor project.

Image source: Bioconductor

BUGS

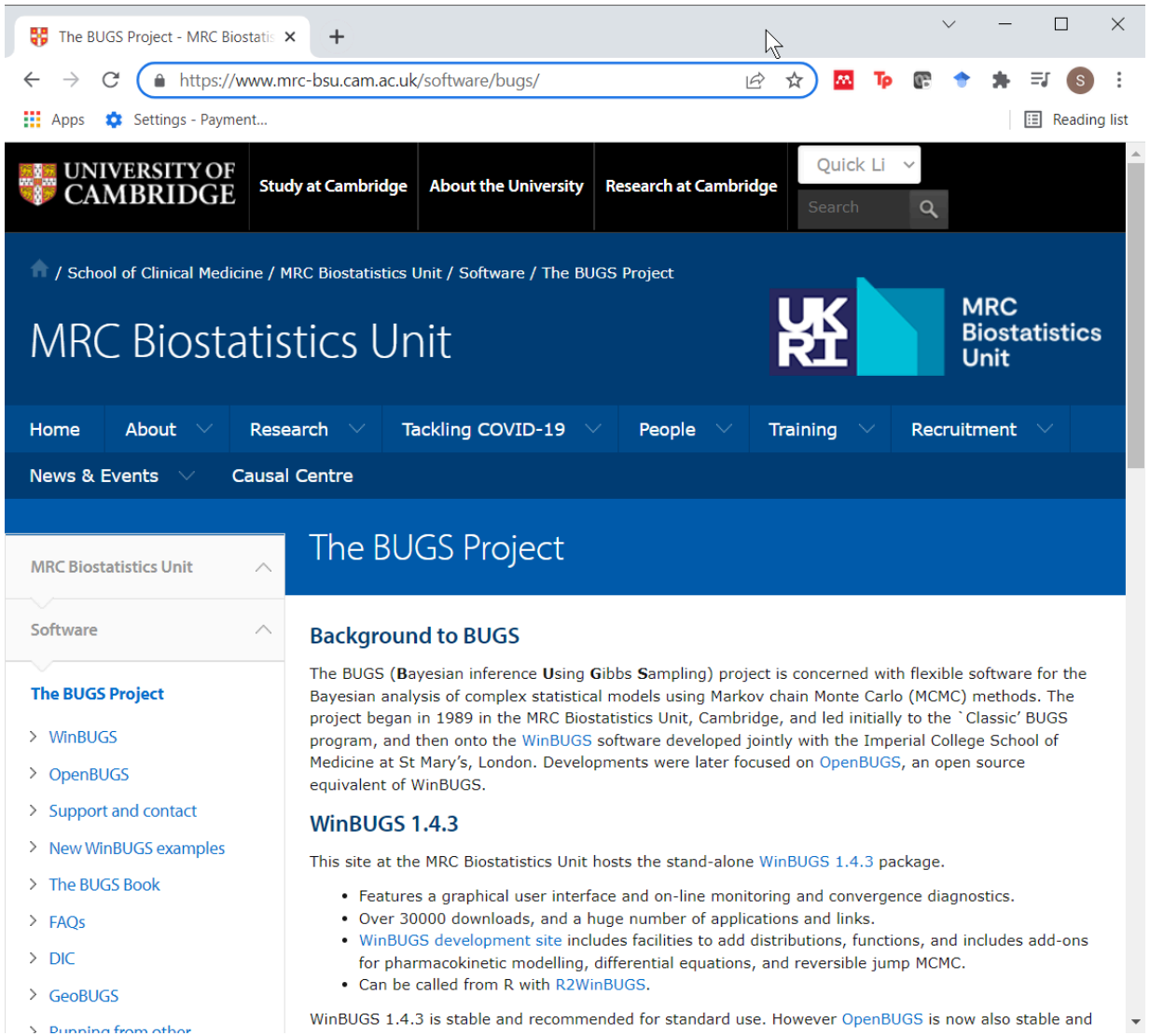


Figure 14. Excerpt from website

Speaker notes

A lot of stand-along programs have leaned heavily on R to provide an interface to run their programs and process their outputs. Notable among these are a series of programs for Bayesian analysis, starting with BUGS. BUGS is an acronym for Bayes Using Gibbs Sampling. While it can be run by itself, it is a lot easier and more convenient to run it from inside R, and most applications of BUGS appear to use R. Other packages, jags (Just Another Gibbls Sampler), and Stan (named after the famous mathematician, Stan Ulam), also rely on R. It is worth noting that these programs are also easily run from Python.

Image source: MRC Biostatistics Unit, University of Cambridge

Menu driven version of R

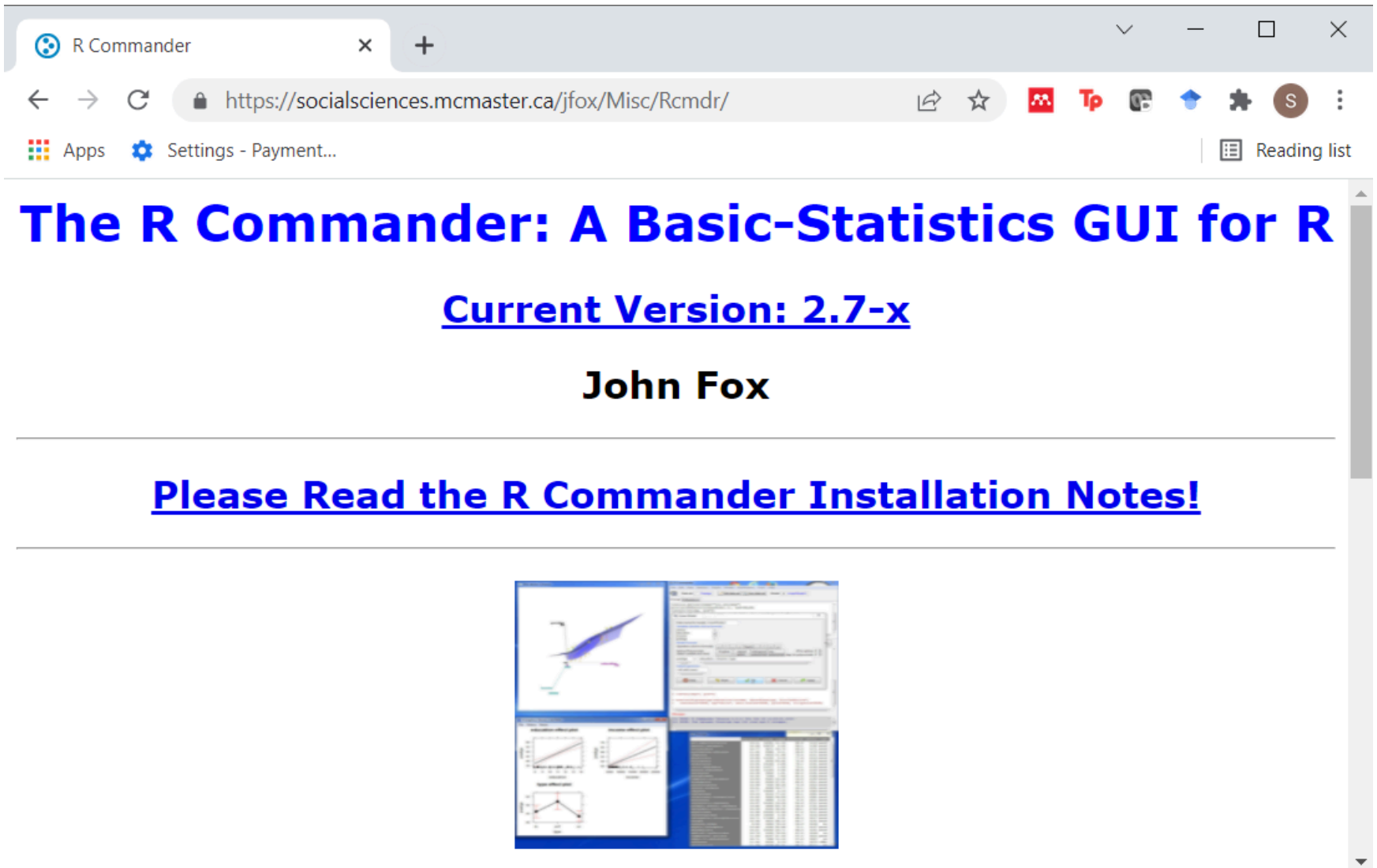


Figure 15. Excerpt from website

Speaker notes

R is an interactive programming language, but menu driven versions of R are available. The most notable of these is R Commander

Note that this link is currently broken.

Image source: McMaster University

RStudio

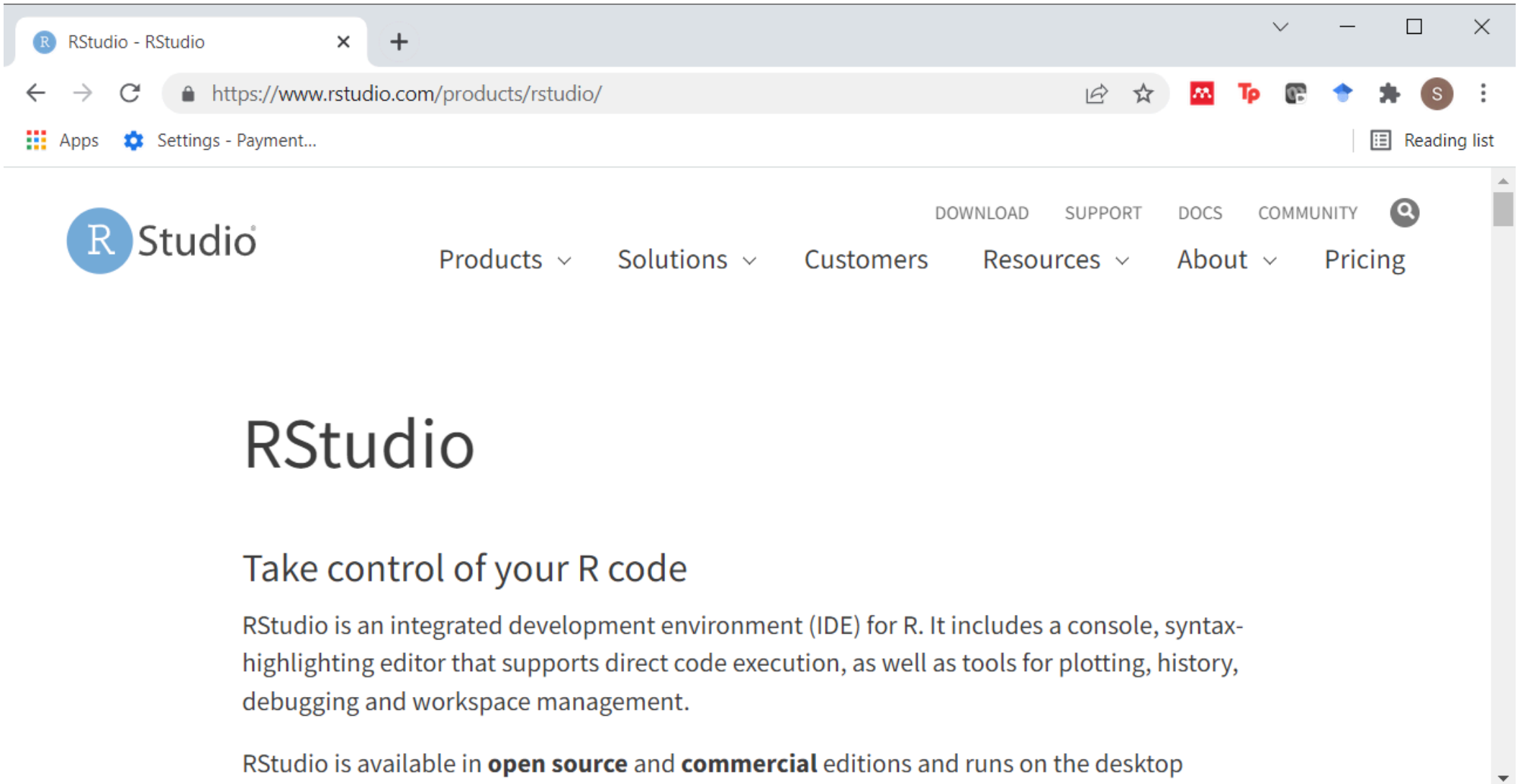


Figure 16. Excerpt from website

Speaker notes

RStudio is an integrated development environment for R, founded somewhere around 2009 to 2011.in 2009. The company that produces RStudio offers both free and commercial versions. They also employ many of the people listed below who have made major contributions to R.

Image source: R Studio (renamed in 2023 to Posit)

RMarkdown

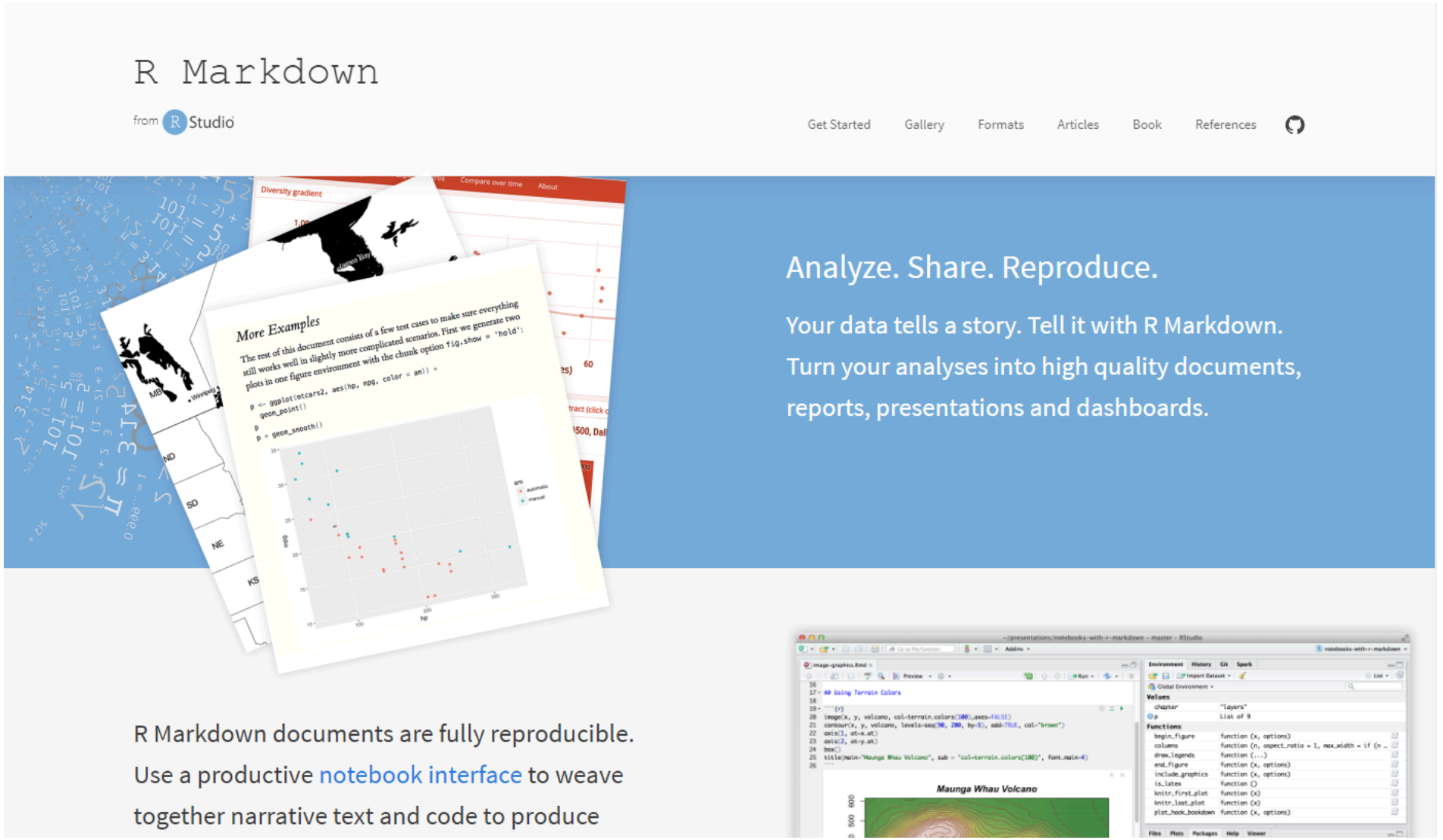


Figure 17. Excerpt from website

Speaker notes

RMarkdown is an extension of Markdown syntax. It provides a simple way to denote special formats of text (italics, bold, headings, links, images, bulleted list) and integrate it with code from R, Python, and other software. RMarkdown uses Pandoc, a program that produces output in a variety of formats: html, Latex, Microsoft Word, Microsoft PowerPoint, PDF, and others.

Image source: RStudio

Recent major contributions: Frank Harrell

Speaker notes

Frank Harrell has produced a lot of advanced statistical models for R. This includes some extremely useful spline tools. His book, Regression Modeling Strategies, a classic text, uses R code throughout.

Image source: R-bloggers

Recent major contributions: Hadley Wickham

Figure 19. Title slide from presentation

Speaker notes

Hadley Wickham has written or co-written a large number of libraries in R that have refashioned R into almost a completely new programming language. He was hired by RStudio in 2012 and has been a driving force behind almost all of the development in that company.

Image source: Data Science, College of Science, University of Notre Dame

The tidyverse library

Figure 20. Hex sticker for tidyverse

Speaker notes

Originally, these packages were referred to collectively as the “Hadleyverse.” But Hadley Wickham discouraged that in favor of the name “tidyverse.”

The tidyverse package is a collection of several different packages which provide enhancements to the R programming language. These libraries share a common programming philosophy. There are several dozen libraries in total, but only a core set of libraries are loaded with the library(tidyverse) function. Other tidyverse packages must be loaded separately.

The tidyverse is a collection of packages for the R programming language developed by Hadley Wickham and others. I single out Hadley Wickham because he has been a major force behind the programming philosophy of the tidyverse and the lead author for many of the most important packages in the tidyverse.

The tidyverse packages embrace some guiding principles described in the tidyverse manifesto. The packages in the tidyverse encourage the use of tidy data. Tidy data is related to the database concept of normalization, though it is described from a statistical perspective (which means that an idiot like me can still understand it). The general concepts behind tidy data are described in a vignette and in a 2014 publication in the Journal of Statistical Software. The tidyverse research team has published a detailed guides on coding practices and program style that are consistent with their principles.

Here are some of the libraries in core set of libraries.

Image source: RStudio github

dplyr

Figure 21. Hex sticker for dplyr

Speaker notes

dplyr provides a set of functions for data manipulation.

Image source: RStudio github

ggplot2

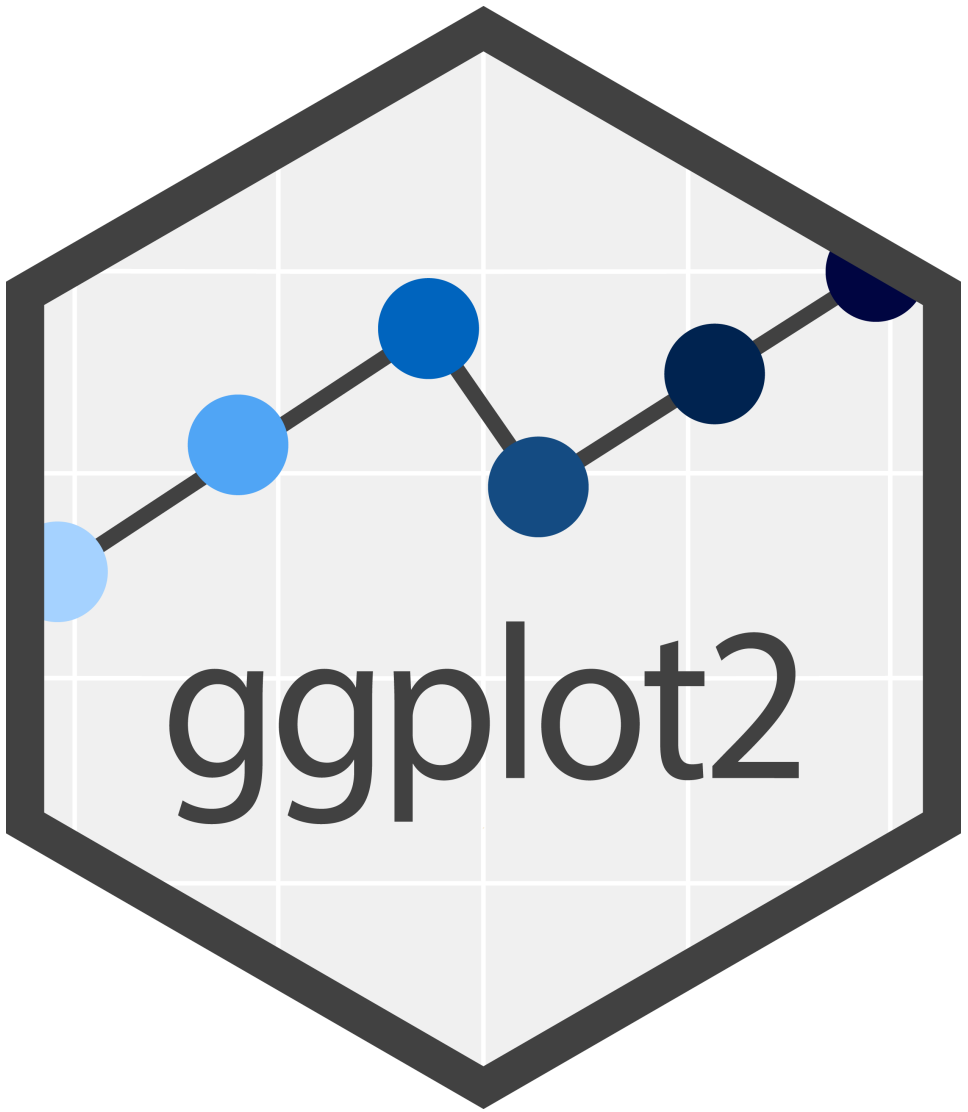


Figure 22. Hex sticker for ggplot2

Speaker notes

While R has some excellent graphics capabilities built in, they are somewhat difficult to use. The ggplot2 library simplifies the process of graphing by separating the parts of a graph into different layers. It is based on a conceptual framework developed by Leland Wilkinson in his book, The Grammar of Graphics.

Image source: RStudio github

magrittr

Figure 23. Hex sticker for magrittr

Speaker notes

magrittr provides a pipe operator. The concept of the pipe was developed first in Unix systems almost 50 years ago. The pipe operator (percent-greater than-percent) takes input from the left side of the operator and feeds it to a function listed on the right side of the operator. Pipes can be chained together. They make your code simpler and more readable.

We may or may not cover pipes in this class.

Image source: RStudio github

readr

Figure 24. Hex sticker for readr

Speaker notes

While R has many functions for reading text data, they are slow for very large files. The readr library reads text files much faster, offers some enhancements, and provides a simpler syntax.

Image source: RStudio github

stringr

Figure 25. Hex sticker for stingr

Speaker notes

stringr simplifies the manipulation of string or text data.

Image source: RStudio github

tibble

Figure 26. Hex sticker for tibble

Speaker notes

R has a variety of internal storage formats: arrays, lists, matrices, and data frames. We will focus mostly on data frames in this class. The tibble package offers an internal storage format, a tibble, that is very similar to a data frame, but it offers some extra features for convenience and simplicity.

Image source: RStudio github

tidyr

Figure 27. Hex sticker for tidyr

Speaker notes

tidyr provides a series of functions that help with data manipulation, especially for longitudinal data.

Image source: RStudio github

Other packages in the tidyverse

- In the core package
  - forcats
  - purr
- Outside the core package
  - broom
  - lubridate
  - readxl
  - many others

Speaker notes

Two other packages in the tidyverse core, forcats and purr, are for advanced applications.

Outside of the core package, some of the packages that I like are broom (which simplifies and standardizes the output from different data analysis functions) lubridate (which simplifies the manipulaton of dates), and readxl (which reads Microsoft Excel files). There are quite a few others.

Recent major contributions: Yihui Xie

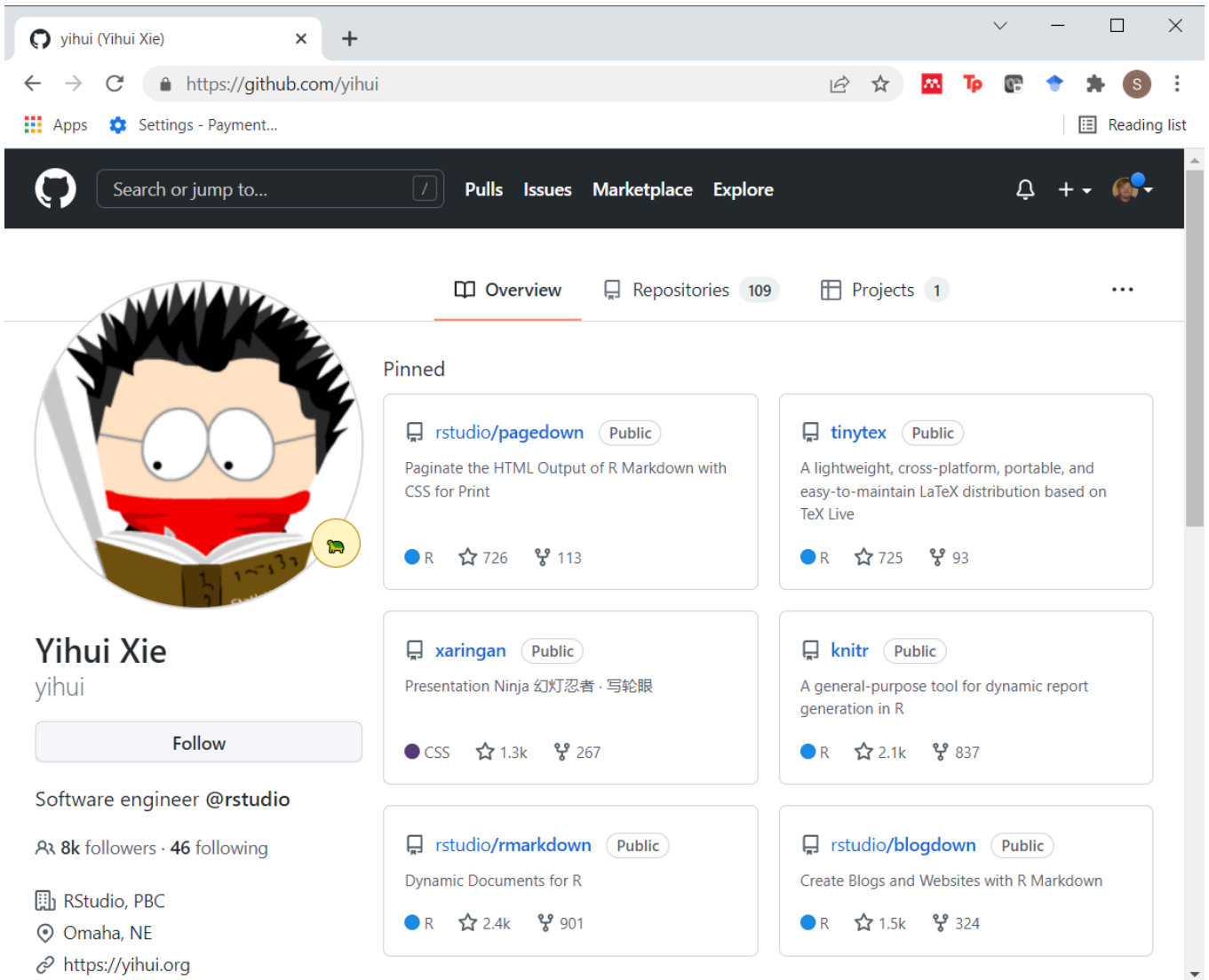


Figure 28. Exceprt from github site

Speaker notes

Another prolific contributor to R is Yihui Xie. He was an employee of Rstudio from 2013 through 2023, but was laid off as part of the transition to Posit (see below).

Image source: Yihui Xie github

knitr

Figure 29. Hex sticker for knitr

Speaker notes

He wrote the package knitr back in 2012 that has revolutionized the field of reproducible research. knitr is an improvement on the package sweave. It takes R code, runs it and creates documents in a variety of formats using Pandoc.

Image source: RStudio github

bookdown

Figure 30. Hex sticker for bookdown

Speaker notes

He wrote also wrote a package, bookdown, that has revolutionized the book publishing world. You can now write an entire book in R with the help of this package. It has publication ready graphics, tables, and formulas. It produces the table of contents, and an index. Over a thousand books have been produced using bookdown, including the definitive guide to bookdown itself, bookdown: Authoring Books and Technical Documents with R Markdown by Yihui Xie.

Image source: RStudio github

Other works by Yihui Xie

- blogdown
- tinytex
- xaringan

Speaker notes

There are a lot more works by Yihui Xie that are worth discussing. blogdown uses R Markdown code to create a blog site. It is based on an open source web development system called Hugo. I am currently trying to convert my website (over 1,800 pages) to blogdown.

tinytex is an attempt to develop a minimal package for producing LaTex documents. It has all the features that you need to work with R Markdown, but does not include some of the extra features found in other versions of LaTex, that needlessly (in his opinion) add to the complexity of using LaTeX as part of R Markdown.

xaringan is a presentation format using html that offers an alternative to beamer and slidy.

RStudio renamed as Posit

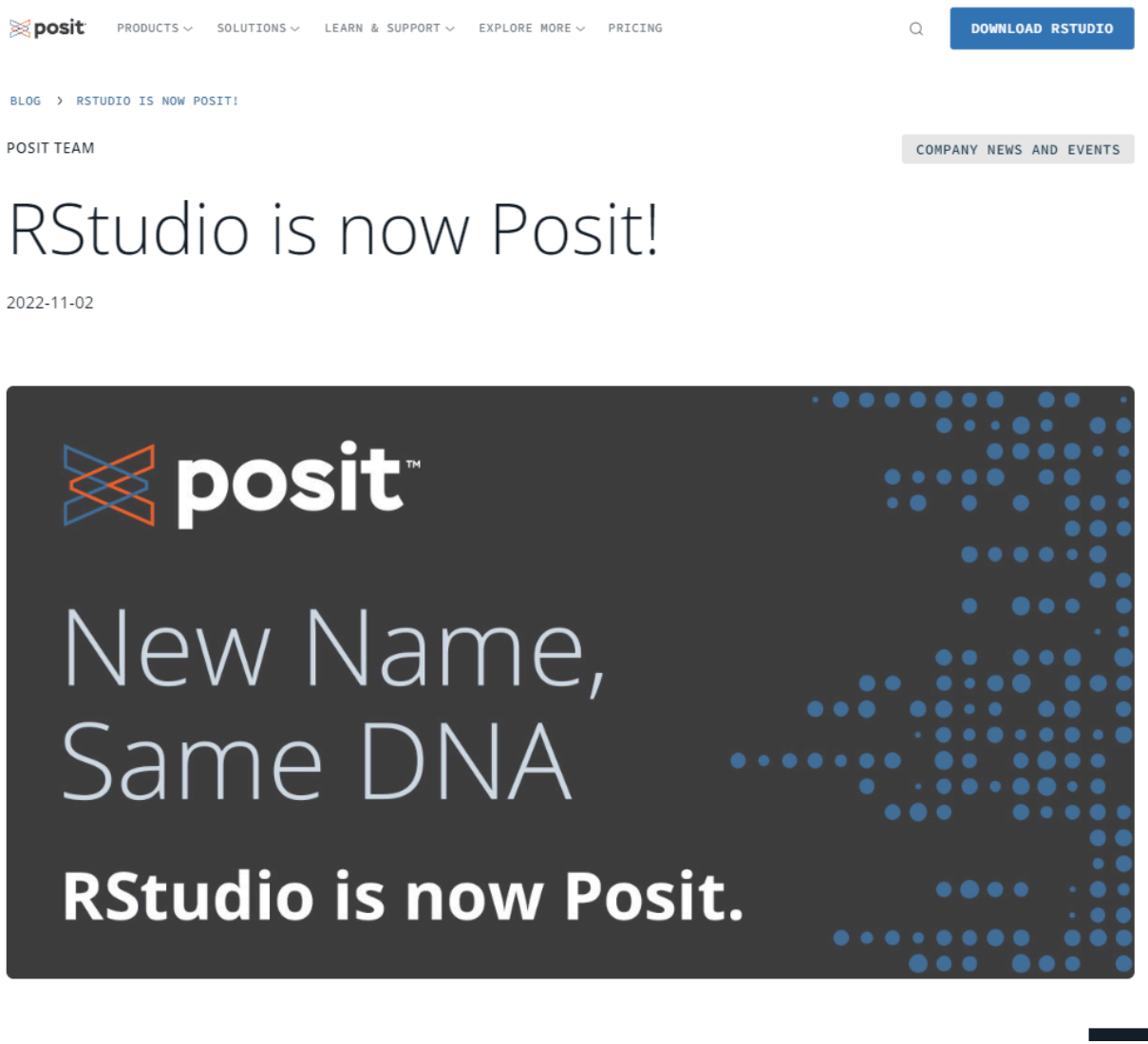


Figure 31. Excerpt from Posit blog

Speaker notes

In 2022, the company that produces RStudio renamed itself from RStudio to Posit. They felt that the name RStudio was focused just on R and they wanted to branch out with a broader focus to a larger community. In particular, they wanted to attract users of Python and users of Jupyter notebooks.

One major personnel shifts during the transition was the hiring of Wes McKinney, creator of the pandas package for Python. Another was the layoff of Yihui Xie, described above.

Image source: Posit blog

Quarto

Figure 32. Hex sticker for Quarto

Speaker notes

One of the first packages announced by the new Posit company was Quarto. Quarto is an alternative to RMarkdown. While RMarkdown requires that you install R prior to use, Quarto does not. It is part of the broadening strategy of Posit and opens up the features of RMarkdown to those who prefer to use only Python on their systems. It also opens up the features of RMarkdown to those who prefer to use only Jupyter on their systems.

Posit will continue to maintain RMarkdown for the forseeable future. New features, however, will be added to Quarto and only ported over to RMarkdown if it can be done easily. So you should adopt Quarto if you want to be on the cutting edge. Any files written using RMarkdown will work with Quarto, so the transition, if you are a seasoned user of RMarkdown is easy. The reverse is not necessarily true. There are a few features in Quarto (not too many, at least in 2024) that will not work with RMarkdown.

Image source: RStudio github

Positron

README | Code of conduct | License | Security

Positron

What is Positron?

- A next-generation data science IDE built by Posit PBC
- An extensible, polyglot tool for writing code and exploring data
- A familiar environment for reproducible authoring and publishing

> **Important**
>
> Positron is an early stage project under active development and may not yet be a good fit for you. If you are interested in experimenting with it, we welcome your feedback!

Figure 33. Excerpt from README file

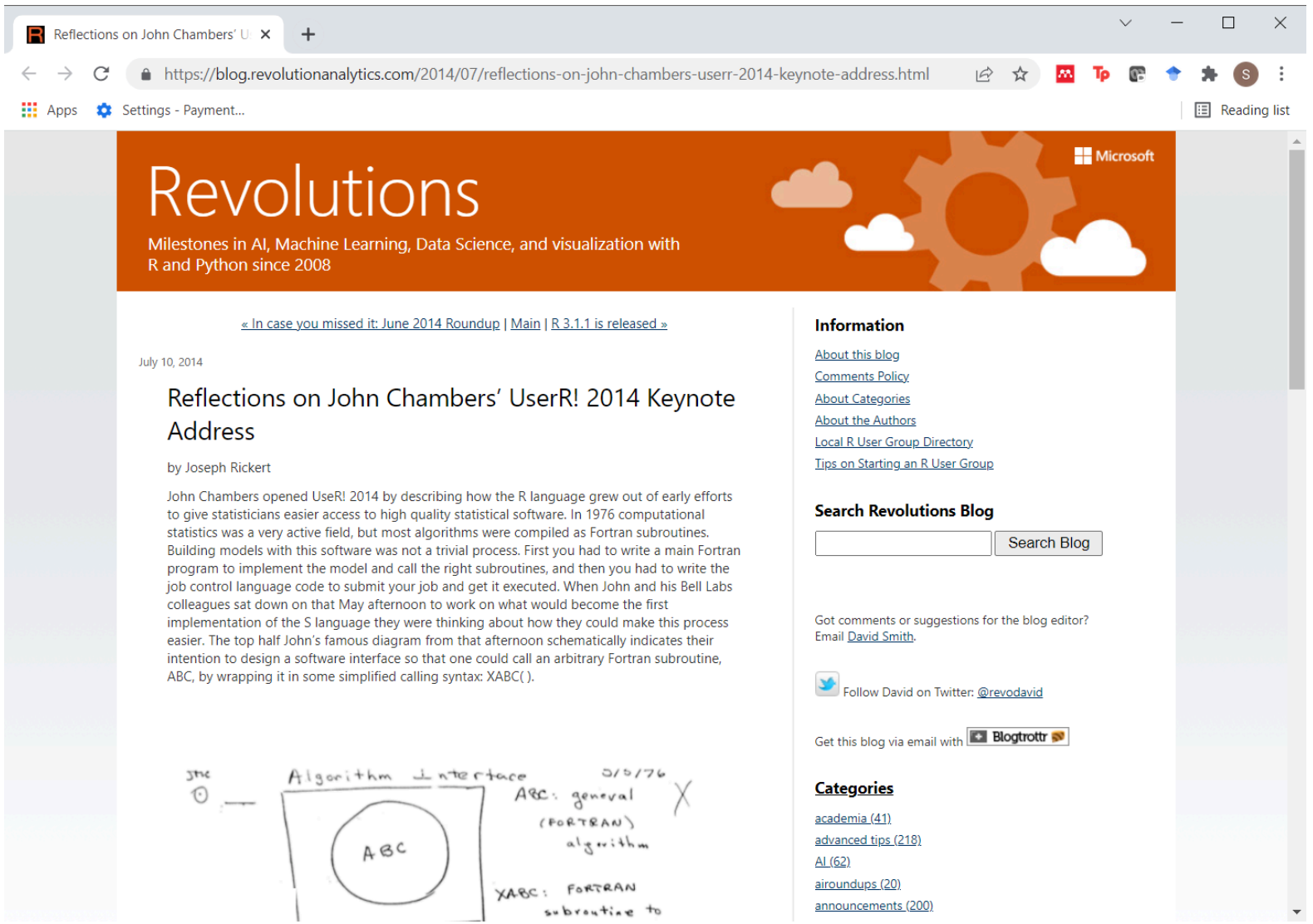
Speaker notes

A new integrated development environment, Positron, is currently (2004) in the beta testing phase. It is intended as an eventual replacement for RStudio, but I would not recommend making the transition to Positron just yet.

Positron is another part of the broadening strategy of Posit. While it allows you to run Python code, RMarkdown is built on a foundation of R. Positron, based on CODE OSS, uses the same basic integrated development environment as Microsoft Visual Studio. You can use it on a machine that only has Python installed or a machine that only has R installed. Don’t try it on a machine that has neither installed, at least not yet. A big advantage of decoupling from R is that Positron will not crash when R (or Python) crashes. it also allows you to use extensions developed for Microsoft Visual Studio.

Image source: posit-dev github

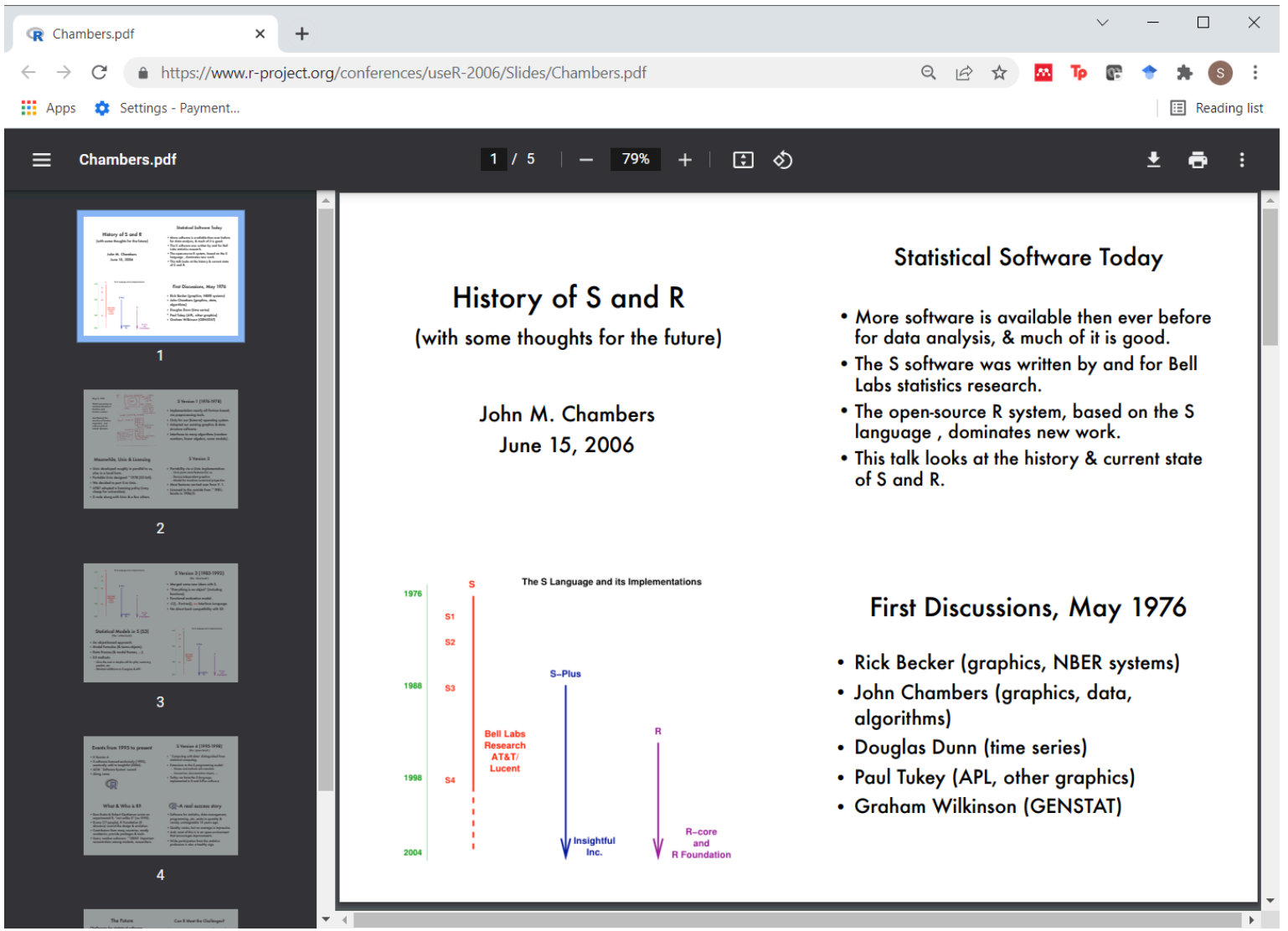
If you want to learn more: Rickert 2014



Speaker notes

The Revolutions Analytic blog posted a nice summary of a John Chambers talk on the history of S at the Use R! 2014 conference

If you want to learn more: Chambers 2006



Speaker notes

That article has links to the slides (PDF format) of a 2006 talk (again on the history of S) by John Chambers.

If you want to learn more: Hastie 2014

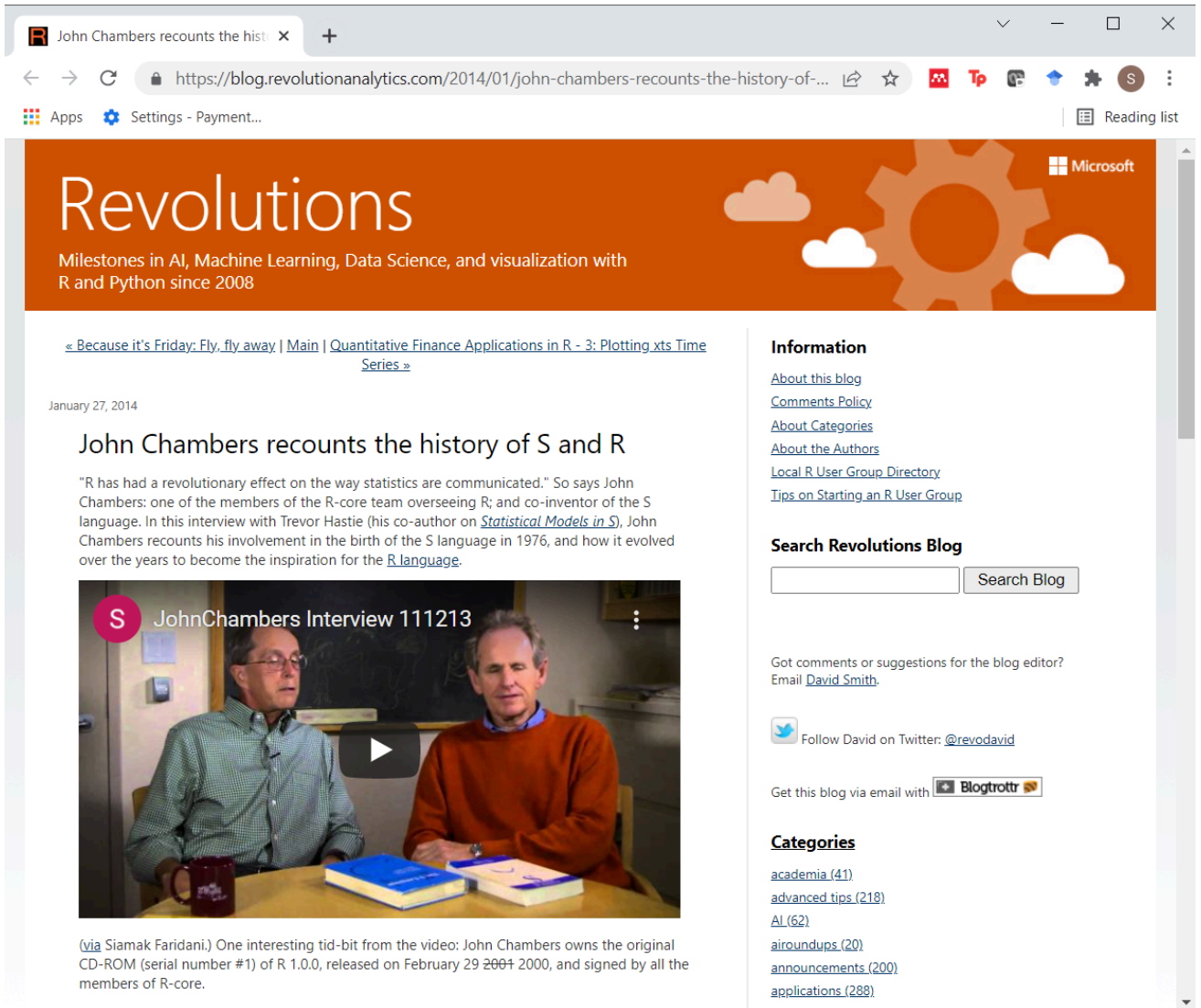


Figure 36. Excerpt from blog post

Speaker notes

as well as a video interview of John Chambers by Trevor Hastie.

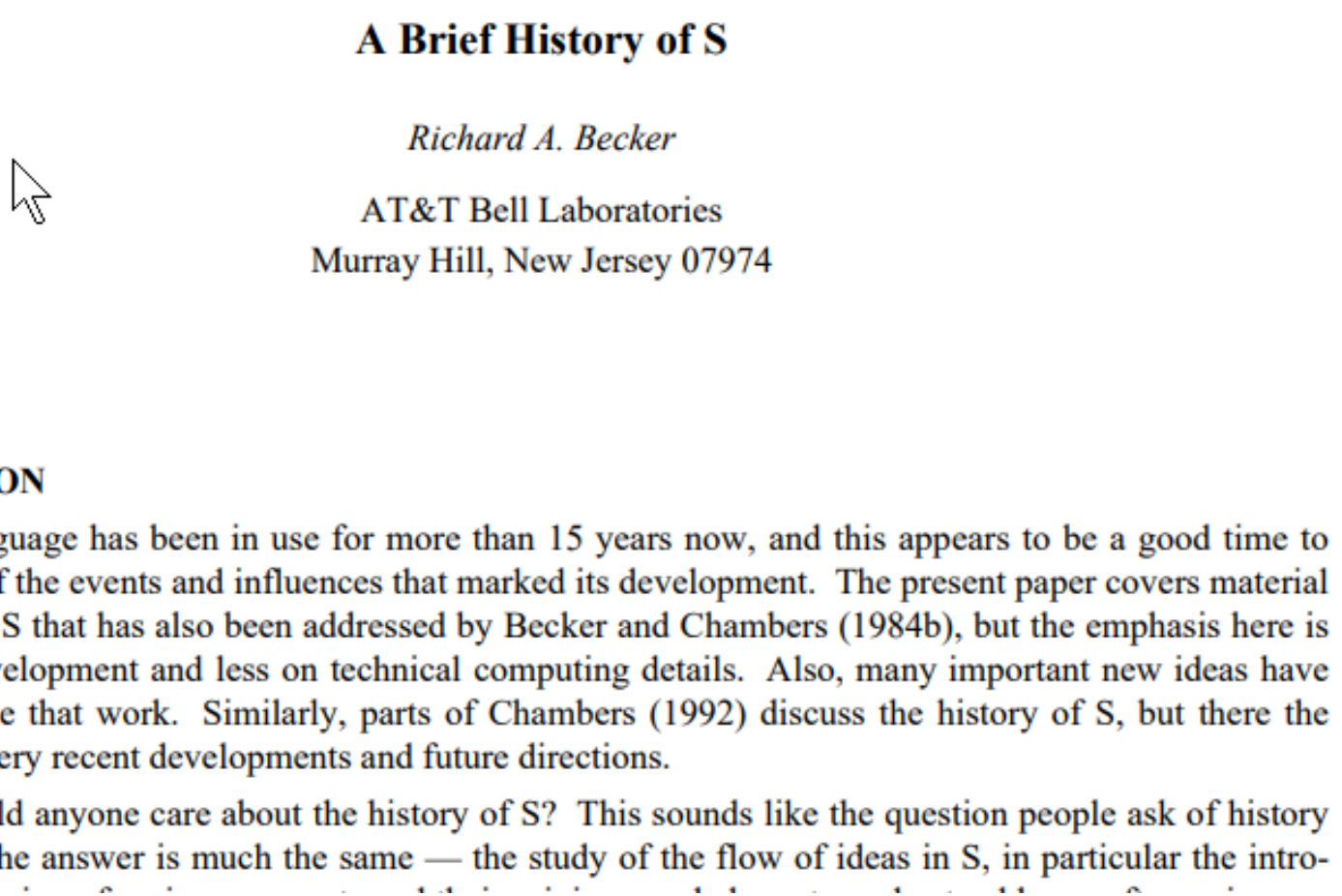
If you want to learn more: Ihaka 1998

Figure 37. Excerpt from research paper

Speaker notes

and a 1998 paper (PDF format) by Ross Ihaka on the past (!) and future of R presented at the Interface conference.

If you want to learn more: Becker (no date)



A Brief History of S

Richard A. Becker

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

INTRODUCTION

The S language has been in use for more than 15 years now, and this appears to be a good time to recollect some of the events and influences that marked its development. The present paper covers material on the design of S that has also been addressed by Becker and Chambers (1984b), but the emphasis here is on historical development and less on technical computing details. Also, many important new ideas have come about since that work. Similarly, parts of Chambers (1992) discuss the history of S, but there the emphasis is on very recent developments and future directions.

Why should anyone care about the history of S? This sounds like the question people ask of history in general, and the answer is much the same — the study of the flow of ideas in S, in particular the intro-

Figure 38. Excerpt from paper

Speaker notes

Richard Beckman. A Brief History of S. Available in pdf format

If you want to learn more: Smith 2020

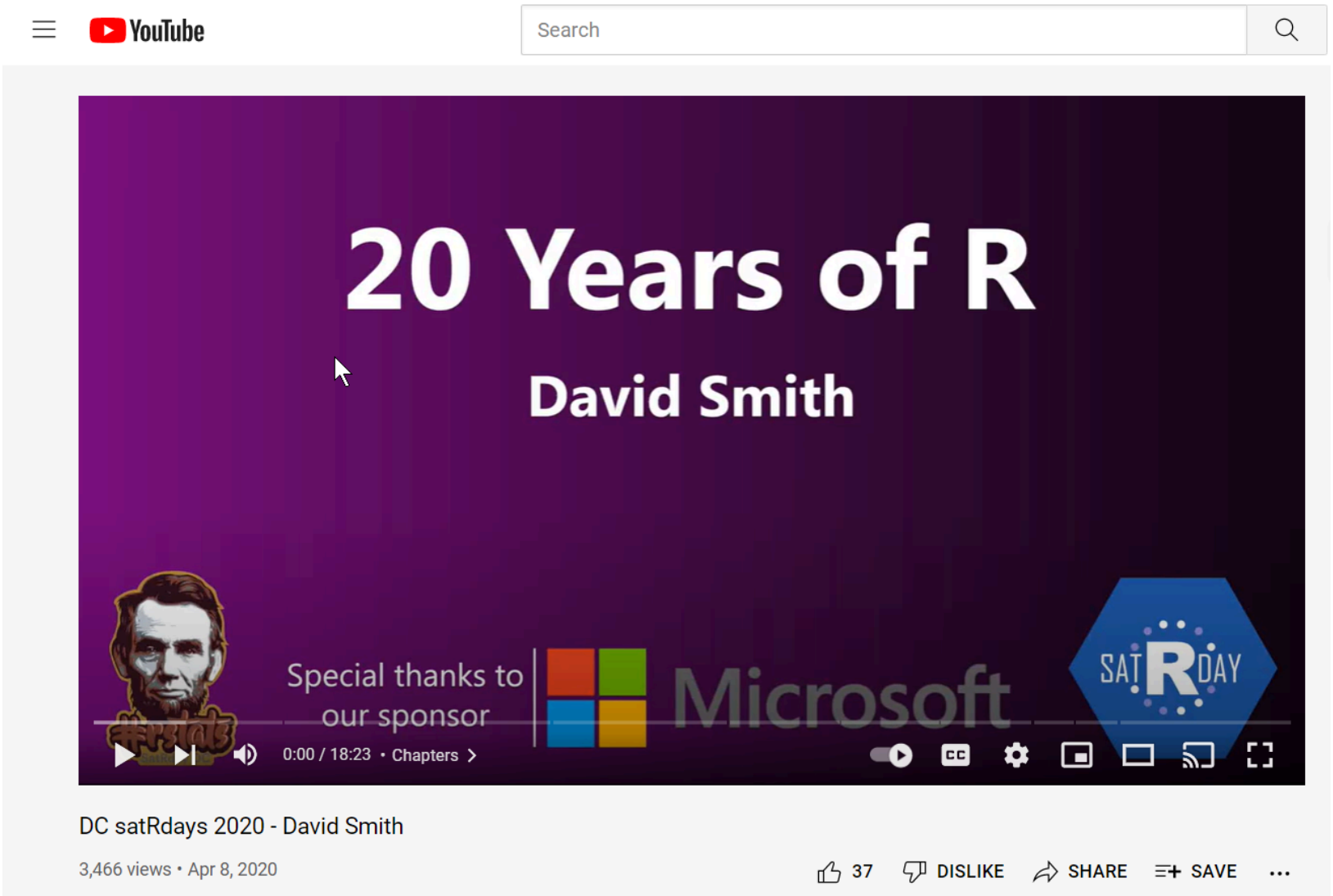


Figure 39. Excerpt from website

Speaker notes

David Smith, 20 years of R, presented at DC satRdays. Available as a YouTube video

You can find an earlier version of this page on my blog.

Break #3

- What you have learned
  - History of R
- What's coming next
  - Scales of measurement

Scales of measurement

- Dichotomy
  - Continuous
  - Categorical
- Stevens scales of measurement (controversial!)
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Addition/subtraction not allowed for ordinal data
  - Mean of ordinal data is meaningless

Speaker notes

There is a basic dichotomy of data types: categorical and continuous. Now I have said many many times before that I hate dichotomies. Nevertheless, they are useful.

A psychologist, Stanley Smith Stevens divided the entire universe of data into four categories: nominal, ordinal, interval, ratio. I won’t review the definitions for all of these, but ordinal data is categorical data where there is a natural ordering of categories. An important limitation to ordinal data, but where the spacing between successive units is not consistent.

The belief among many (but not all) researchers, is that certain statistics are inappropriate for certain measurement scales. In particular, these researchers are very fussy about ordinal data.

An example of ordinal data.

- “Do you agree or disagree with the following statements”
  - “I believe that knowledge of Statistics is important for my job.”
    - 1 = Strongly disagree,
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly agree

Speaker notes

Speaker notes

An example of ordinal data is the Likert scale. This takes various forms, but often it is used with group of questions on a questionnaire that reads something like

“Do you agree or disagree with the following statements”

You are asked to respond 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree.

Now I’m sure everyone today is going to choose 5. But assigning numbers 1, 2, 3, 4, and 5 to categories of strongly disagree, disagree, neutral, agree, and strongly agree may falsely imply that a jump from 3 (neutral) to 4 (agree) is about the same amount of improvement as a jump from 4 (agree) to 5 (strongly agree). That’s probably not the case.

You can’t really average ordinal data, some people say because that implies that two responses of “Agree” are the same as one response of “Neutral” along with a response of “Strongly agree”.

Do you want everyone to be at least somewhat on your side or do you want to have a smaller number of very enthusiastic supporters.

If you believe that two 4’s are not the same as a 3 and a 5, then you can’t average.

Now I beg to disagree here, but I am part of a minority opinion. I think that if at the start of this class, your average rating was 3.2 and after I finish the lecture, your average rating climbs to 4.4, that I have done my job well.

If it only jumps to 3.6, then I have still done well, but not as much as that jump to 4.4.

Another example of ordinal data, course grades

- A = 4
- B = 3
- C = 2
- D = 1
- F = 0

Speaker notes

Speaker notes

Another example of ordinal data is grades assigned to students. Now everyone in this class is getting an A, but in other classes I teach I might assign different grades. You can attach a number to each of these grades, 4 for A, 3 for B, 2 for C, 1 for D, and 0 for F.

These numbers seem to imply that a student with two B's is as smart as a student with an A and a C.

It raises an interesting story. A colleague of mine told me that he would never hire anyone with a single F on their transcript. An F is a red flag, he felt. So he would not want to assign a value of 0 to F, because that implies that the difference between an F and a D is equivalent to the difference between a B and and A. He's want to assign a value like negative one million to an F so that the average would be pulled way down for a single F, no matter what the other grades would be.

Now I would never be so harsh, but there is really nothing wrong with his perspective. And I would certainly treat a student with three A's and one F differently from a student with two A's and two C's even though mathematically, both average out to 3.0.

Now, in spite of all the obvious problems with equivalence between different grades, most of us still accept a grade point average as a meaningful indicator of how well a student did in school.

Break #4

- What you have learned
  - Scales of measurement
- What’s coming next
  - Tests of hypothesis

What is a population?

- Population: a group that you wish to generalize your research results to. It is defined in terms of
  - Demography,
  - Geography,
  - Occupation,
  - Time,
  - Care requirements,
  - Diagnosis,
  - Or some combination of the above.

Speaker notes

A population is a group that you have an interest in. You want to get a better understanding of this group, so you conduct a research study and wish to generalize the results of that study to the population.

In clinical research, a population is almost always a group of people. There are a few exceptions. Sometimes you want to characterize inanimate objects, such as a group of hospitals or a group of medical devices. But let’s keep the focus on people for now.

A population of people is defined in terms of certain characteristics. Usually it is a combination of these characteristics.

Example of a population

All infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

Speaker notes

Speaker notes

Here is an example of a population. It has many of the characteristics described on the previous slide: demography (infants), geography (born in Missouri), time (born in calendar year 1995, during first year of life) and care requirements (one or more ER visits).

Most times the population is so large that it is difficult to get data on all the individuals of that population.

Here, we actually did have access to the data on all 29,637 infants, but most times you would not be so fortunate.

What is a sample?

- Sample: subset of a population.
- Random sample: every person has the same probability of being in the sample.
- Biased sample: Some people have a decreased probability of being in the sample.
  - Always ask “who was left out?”

Speaker notes

Speaker notes

A sample is a subset of a population. Because that population of infants was so large, you decided to collect data on a smaller group, a sample of 100 infants, say.

Statistics, according to one definition is the use of data from samples to make inferences about populations. That may be a bit too narrow a definition, but it does characterize quite a bit of what we statisticians do.

A random sample is a special type of sample. It is chosen in a way to insure that every person in the sample has the same probability of being in the sample.

In contrast a biased sample is one where some people in the population have a decreased chance of being in the sample. Often in a biased sample some people in the population are totally excluded.

An example of a biased sample

- A researcher wants to characterize **illicit drug use in teenagers.** She distributes a questionnaire to students attending a local public high school
- (in the U.S. high school is grades 9-12, which is mostly students from ages 14 to 18.)
- Explain how this sample is biased.
- Who has a decreased or even zero probability of being selected.

Type your ideas in the chat box.

Speaker notes

Speaker notes

Here is a scenario where a researcher selects a biased sample. I should note here that this is an example specific to the United States. In Italy, you might talk about a survey distributed to the scuola secondaria di secondo grado.

STOP AND GET STUDENT RESPONSES

There are a variety of responses here. The sample does not include home schooled students, students in private schools, students with chronic diseases that force frequent school absences, and students who have dropped out.

Fixing a biased sample

- Redfine your population
  - Not all teenagers,
    - but those attending public high schools.

What is a parameter?

- A parameter is a number computed from a population.
  - Examples
    - Average health care cost associated with the 29,637 children
    - Proportion of these 29,637 children who died in their first year of life.
    - Correlation between gestational age and number of ER visits of these 29,637 children.
  - Designated by Greek letters (μ, π, ρ)

What is a statistic?

- A statistic is a number computed from a sample
  - Examples
    - Average health care cost associated with 100 children.
    - Proportion of these 100 children who died in their first year of life.
    - Correlation between gestational age and number of ER visits of these 100 children.
  - Designated by non-Greek letters ($\bar{X}$, $\hat{p}$, r).

What is Statistics?

- Statistics
  - The use of information from a sample (a statistic) to make inferences about a population (a parameter)
    - Often a comparison of two populations

What is the null hypothesis?

- The null hypothesis (H_0) is a statement about a parameter.
- It implies no difference, no change, or no relationship.
  - Examples
    - $H_0: \ \mu_1 - \mu_2 = 0$
    - $H_0: \ \pi_1 - \pi_2 = 0$
    - $H_0: \ \rho = 0$

What is the alternative hypothesis?

- The alternative hypothesis (H_1 or H_a) implies a difference, change, or relationship.
  - Examples
    - $H_1: \ \mu_1 - \mu_2 \neq 0$
    - $H_1: \ \pi_1 - \pi_2 \neq 0$
    - $H_1: \ \rho \neq 0$

Hypothesis in English instead of Greek

- Only statisticians like Greek letters
  - Translate to simple text
  - For two group comparisons
    - Safer, more effective
  - For regression models
    - Trend, association

Speaker notes

As a researcher, you should always think about your hypothesis in terms of population parameters, but your writing should use text. Translate the Greek letters to English.

If you have a hypothesis that compares two groups, look for comparative words like “safer” or “more effective”. If your hypothesis involves some type of regression model, you should consider terms like “trend” or “association”.

Use PICO

- P = patient population
- I = intervention
- C = control
- O = outcome

Example of text hypotheses, 1 of 2

- "… the objective of this 78-week randomised, placebo-controlled study was to determine whether treatment with nilvadipine sustained-release 8 mg, once a day, was effective and safe in slowing the rate of cognitive decline in patients with mild to moderate Alzheimer disease."
  - Lawlor B, Segurado R, Kennelly S, et al. Nilvadipine in mild to moderate Alzheimer disease: A randomised controlled trial. PLoS Med. 2018; 15(9): e1002660. DOI: 10.1371/journal.pmed.1002660

Speaker notes

Here’s an example of a two group comparison. One group gets nilvadipine and the other group gets a placebo. Safety was measured as the proportion of patients who experienced an adverse event. The researchers also measured the proportion of patients who experienced a serious adverse event. So the Greek hypothesis would involve pi’s.

Effectiveness was measured using the Alzheimer’s Disease Assessment Scale Cognitive Subscale-12 and the Clinical Dementia Rating Scale sum of boxes. Both of these outcome measurements are continuous, so the Greek hypothesis would involve mu’s.

PICO for this study

- P = patients with mild to moderate Alzheimer disease
- I = Nilvadine
- C = placebo
- O = cognitive function

Example of text hypotheses, 2 of 2

- "… we investigated trends in BCC incidence over a span of 20 years and the associations between incident BCC and risk factors in a total population of 140,171 participants from 2 large US-based cohort studies: women in the Nurses' Health Study (NHS; 1986–2006) and men in the Health Professionals' Follow-up Study (HPFS; 1988–2006)."
  - Wu S, Han J, Li WQ, Li T, Qureshi AA. Basal-cell carcinoma incidence and associated risk factors in U.S. women and men. Am J Epidemiol. 2013; 178(6): 890–897. DOI: 10.1093/aje/kwt073

Speaker notes

This study used a regression model, a Cox regression model, to study trends and associations, so the Greek hypotheses would involve beta’s.

PICO for this study

- P = female nurses/male health professionals
- I = various risk factors
- C = absence of various risk factors
- O = presence/absence of BCC

One-sided alternatives

- Examples
  - $H_1: \ \mu_1 - \mu_2 > 0$
  - $H_1: \ \pi_1 - \pi_2 > 0$
  - $H_1: \ \rho > 0$
- Changes in only one direction expected
- Changes in opposite direction uninteresting

Passive smoking controversy

- EPA meta-analysis of passive smoking
  - Criticized for using a one-sided hypothesis
  - Samet JM, Burke TA. Turning science into junk: the tobacco industry and passive smoking. Am J Public Health. 2001;91(11):1742–1744.

Speaker notes

Available in html format or PDF format.

Consider a study of the effects of second-hand smoke. These studies always use directional alternatives. From what we know about active cigarette smoking is that it increases the risk of cancer and cardiovascular disease. So there is no reason to expect that passive smoke exposure should be any different than active smoking. Maybe it is less toxic, because of dilution and because the smoking coming off a cigarette from one end is different than the smoke coming off the cigarette from the other end. Fair enough, but there is not reason to believe that things are so different that all of a sudden the smoke becomes protective.

Since there is no scientific basis for a protective effect of passive smoking, it makes sense to test that passive smoking has no effect versus it having an increase in bad outcomes compared to the control group. So your null hypothesis is “not harmful” and your alternative is “harmful”. The beneficial hypothesis is lumped into the null hypothesis, but no one would dare claim that passive smoking was protective.

Actually, the tobacco companies did complain that the use of a directional alternative violated the norms of science. They won in a court battle in North Carolina, but lost on appeal.

As another aside, I was involved with prayer study. We planned this study using a one-sided hypothesis (remote prayer has a positive effect on health). The Institutional Review Board suggested changing this to a two-sided hypothesis (remote prayer has either a positive or a negative effect on health). Thankfully, we did not observe an outcome in the opposite tail as that would have been very difficult to explain.

What is a decision rule? 1 of 3

- Example
  - $H_0: \ \mu_1 - \mu_2 = 0$
  - $H_1: \ \mu_1 - \mu_2 \neq 0$
  - t = $(\bar{X}_1 - \bar{X}_2)$ / se
  - Accept H_0 if t is close to zero.

What is a decision rule? 2 of 3

- Example
  - $H_0: \pi_1 - \pi_2 = 0$
  - $H_1: \pi_1 - \pi_2 \neq 0$
  - t = $(\hat{p}_1 - \hat{p}_2)$ / se
  - Accept H_0 if t is close to zero.

What is a decision rule? 3 of 3

- Example
  - $H_0: \ \rho = 0$
  - $H_1: \ \rho \neq 0$
  - t = r / se
  - Accept H_0 if t is close to zero.

What is a Type I error?

- A Type I error is rejecting the null hypothesis when the null hypothesis is true
  - False positive
  - Example involving drug approval: a Type I error is allowing an ineffective drug onto the market.
- α = P[Type I error]

Speaker notes

Speaker notes

In your research, you specify a null hypothesis (typically labeled H0) and an alternative hypothesis (typically labeled Ha, or sometimes H1). By tradition, the null hypothesis corresponds to no change. When you are using Statistics to decide between these two hypothesis, you have to allow for the possibility of error. Actually, if you are using any other procedure, you should still allow for the possibility of error, but we statisticians are the only ones honest enough to admit this.

A Type I error is rejecting the null hypothesis when the null hypothesis is true.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type I error would be allowing an ineffective drug onto the market.

Remember that the hypotheses involve population parameters. Population parameters are impossible to compute. So you can only talk about Type I errors in an abstract sense. You will never know for certain if you have made a Type I error.

Alpha is the probability of a Type I error, and alpha is a value that you can compute. In most studies, researchers work hard to keep the probability of a Type I error low, typically at 5%.

What is a Type II error?

- A Type II error is accepting the null hypothesis when the null hypothesis is false.
  - False negative result
  - Usually computed at MCD
  - An example involving drug approval: a Type II error is keeping an effective drug off of the market.
- β = P[Type II error]
- Power = $1 - \beta$

Speaker notes

Speaker notes

A Type II error is accepting the null hypothesis when the null hypothesis is false. You should always remember that it is impossible to prove a negative. Some statisticians will emphasize this fact by using the phrase “fail to reject the null hypothesis” in place of “accept the null hypothesis.” The former phrase always strikes me as semantic overkill.

Many studies have small sample sizes that make it difficult to reject the null hypothesis, even when there is a big change in the data. In these situations, a Type II error might be a possible explanation for the negative study results.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type II error would be keeping an effective drug off the market.

It bears repeating that population parameters are impossible to compute. So you will never know for certain if you have made a Type I error.

Beta is the probability of a Type II error. Beta is a known quantity. Typically researchers try to keep beta small. 10% is a typical value, though in some settings, a Type II error rate as large as 20% could be tolerated.

Power is defined as 1-beta. I will talk more about power in a little bit.

What is a p-value?

- Let t =
  - $(\bar{X}_1 - \bar{X}_2)$ / se, or
  - $(\hat{p}_1 - \hat{p}_2)$ / se, or
  - r / se
- p-value = Prob of sample result, t, or a result more extreme,
  - **assuming the null hypothesis is true**
- Small p-value, reject H_0
- Large p-value, accept H_0

Speaker notes

Speaker notes

A p-value is a measure of how much evidence we have against the null hypothesis.

The smaller the p-value, the more evidence we have against H0.

The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true.

The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

Alternate interpretations

- Consistency between the data and the null
  - Small value, inconsistent
  - Large value, consistent
- Evidence against the null
  - Small, lots of evidence against the null
  - Large, little evidence against the null

Speaker notes

Speaker notes

There are two interpretations that I feel are more practical. You can think of the p-value as a measure of consistency between the data and the null hypothesis. A small value implies inconsistency. It is very unlikely that you will get a value like you've seen in your sample or a value more extreme under the assumption that the null hypothesis is true. So you should reject that assumption.

On the other hand if the sample results or anything more extreme has a high probability under the assumption that the null hypothesis is true, then you should feel comfortable accepting that assumption.

I have argued that the p-value is a measure of evidence. Some have called it a poor measure of evidence, but I stand by my interpretation.

If the p-value is small, you have lots of evidence against the null hypothesis. If the p-value is large, you have little or no evidence against the null hypothesis.

What the p-value is not, 1 of 2

- A p-value is NOT the probability that the null hypothesis is true.
  - P[t or more extreme | null] is different than
  - P[null | t or more extreme]
    - P[null] is nonsensical
    - μ, π, or ρ are unknown constants (no sampling error)

Speaker notes

Speaker notes

The p-value is a conditional probability, and you always need to be careful about conditional probabilities. It is a probability about a sample result given an assumption about the population result. It is not a probability about a population result given the sample result. There are two reasons for this.

First, you can’t reorder a conditional probability. The probability of A given B is almost never the same as the probability of B given A. The example I give for this is the probability of being happy given that you are rich. That’s a pretty high number, I hope you’ll agree. There are a few rich people who lead miserable lives, but from everything I’ve seen, most rich people are pretty darn happy. The reverse of this is the probability of being rich given that you are happy. That number is much smaller. Because although I believe that money can buy happiness, a lot of other things can also buy happiness just as well. It’s not quite as easy to find happiness if you’re poor, but somehow, a lot of poor people find a way to be happy anyway.

A second reason that you can’t reverse the order is that you cannot make a probability statement about population parameters. They are numbers computed from the entire population, and are fixed values. You cannot make a probability statement about something that has no sampling error.

Only numbers computed from a sample (i.e., statistics) have sampling error.

What the p-value is not, 2 of 2

- Not a measure FOR either hypothesis
  - Little evidence **against** the null $\neq$ lots of evidence **for** the null
- Not very informative if it is large
  - Need a power calculation, or
  - Narrow confidence interval
- Not very helpful for huge data sets

Speaker notes

The p-value is not a measure for either hypothesis. It is always a measure against a particular hypothesis. Now when the p-value is small, you can make a strong statement. We have lots of evidence against the null hypothesis. That translates into lots of evidence in favor of the alternative hypothesis.

When the p-value is large, however, you are in a quandary. Little or no evidence against the null hypothesis is not the same as lots of evidence for the null hypothesis.

It’s possible to have little or no evidence against the null and also have little or no evidence against the alternative. This happens whenever you have a really small sample size combined with a lot of noise.

You can’t prove a negative, so the saying goes. Well, you can prove a negative, but you have to work harder at it. A large p-value by itself is not persuasive, but if you combine it with a power calculation done prior to data collection, that’s pretty good evidence in support of the null hypothesis.

You could also combine a large p-value with a narrow confidence interval to support the null hypothesis. I’ll talk about that more in just a bit.

In general, the p-value is not very helpful for large samples. We’re seeing this more and more. Just about everything pops up as statistically significant with these huge data sets, and you can’t use the p-value to separate the important stuff from the trivial stuff. You need to look instead at the magnitude of the sample estimates and calculate how much uncertainty you can remove in your future predictions.

A bad test question

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence **for** the null
2. Strong evidence **for** the alternative
3. Little or no evidence **for** the null
4. Little or no evidence **for** the alternative
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Here’s that pop quiz again. Take a look at it quickly. Note that the p-value is of evidence against the null hypothesis. So each of the first four responses is wrong.

I wrote this question quickly, so shame, shame on me. But I’ve reproduced the example because it illustrates an important point.

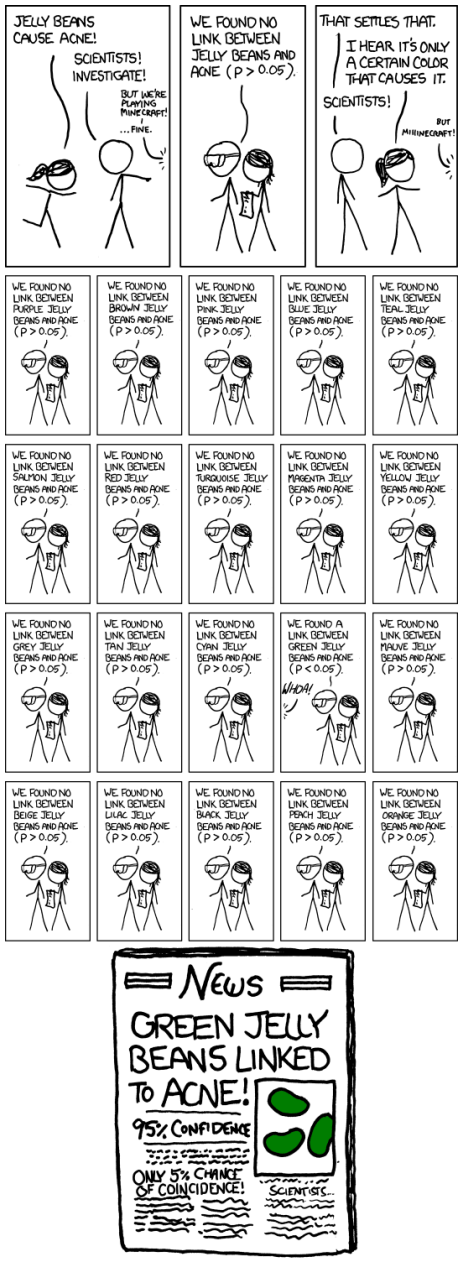


Figure 1: xkcd cartoon about jelly beans and cancer

Speaker notes

Speaker notes

This cartoon is impossible to read, but you can find it on the Canvas site or in the readings. Here's a brief run down.

In the first panel, a woman runs up to a man and shouts: Jelly beans cause acne!

The man replies : Scientists! Investigate!

In the second panel, one scientist, holding a clipboard announces: We found no link between jelly beans and acne ($p > 0.05$).

In the third panel, the woman says: I hear it's only a certain color that causes it.

In a bunch of small panels, the scientist with a clipboard reports: We found no link between purple jelly beans and acne ($p > 0.05$).

We found no link between brown jelly beans and acne ($p > 0.05$).

We found no link between pink jelly beans and acne ($p > 0.05$).

The same for blue, teal, salmon, red, and so forth. And then…

We found a link between green jelly beans and acne ($p < 0.05$). An off-screen voice goes: Whoa!

The next six panels show

We found no link between mauve jelly beans and acne ($p > 0.05$).

We found no link between beige jelly beans and acne ($p > 0.05$).

We found no link between lilac jelly beans and acne ($p > 0.05$).

We found no link between black jelly beans and acne ($p > 0.05$).

We found no link between peach jelly beans and acne ($p > 0.05$).

We found no link between orange jelly beans and acne ($p > 0.05$).

At the bottom is a newspaper with the headline: Green Jelly Beans Linked To Acne! 95% Confidence. Only 5% chance of coincidence!

If you are interested in a transcript and a detailed explanation, https://www.explainxkcd.com/wiki/index.php/882:_Significant

What is p-hacking?

- Abuse of the hypothesis testing framework.
  - Run multiple tests on the same outcome
  - Test multiple outcome measures
  - Remove outliers and retest
- Defenses against p-hacking
  - Bonferroni
  - Primary versus secondary
  - Published protocol

Speaker notes

This is an example of p-hacking. You change the testing process to increase the probability of a Type I error (Rejecting the null hypothesis when the null hypothesis is true). This increases the chance of getting a positive result, which you may find desirable, but only by increasing the probability of a false positive result.

Some examples of p-hacking. Run multiple tests on the same outcome measure. Start with the regular t-test, include the t-test that allows for unequal variances, and run two different non-parametric tests, the Wilcoxon-Mann-Whitney test and the sign test. Choose the test with the smallest p-value.

You also might consider multiple outcome measures. Compare the mortality rate, the relapse rate, and the re-hospitalization rate. If any of the three is statistically significant, claim victory.

You could also do this with longitudinal data. Compare pain relief at one hour and at four hours. If you see a difference at one hour, claim that your new medication is faster acting. If you see a difference at four hours, claim that your medication is longer lasting.

You might run a test with the full data set and then with an outlier or two removed. Report for the data set that has the smaller p-value and pretend that this was your original choice all along.

These are only a few of the choices. I don’t want to say more because I feel like I’m the devil tempting you.

There are two defenses against p-hacking. Well three if you count being honest. But what I mean is there are two things that you can do that will satisfy others that you are playing fairly.

First, you can adjust your decision rule by using a Bonferroni correction. Bonferroni divides alpha by the number of tests. If you are using three different outcome measures, compare your p-value of 0.0133 instead of 0.05.

Second, you can designate one of your outcome measures as primary. If you achieve statistical significance on your primary outcome, great. The remaining outcome measures are secondary. If you achieve statistical significance on a secondary outcome meausure only, report the results as provisional and requiring independent replication.

You should publish a detailed protocol, either through a clinical trial registry, or now there are journals which accept publications of the research protocols before any data are collected. It’s a paper with literature review and methods section, but no results and no discussion section.

Now p-hacking has happened because some people have a skewed view of research. They are interested in using research to promote their own agenda rather than using research to uncover the truth. Perfectly understandable if you are a drug company, but you as an independent researcher should never try to skew the data. It hurts you and it hurts your patients. You need to adopt a disinterested posture in that you are glad when the research points in one direction and you are glad when it points in the opposite direction, because either way, you know more than you did before and you can treat your patients better because of this knowledge.

Break #5

- What you have learned
  - Tests of hypothesis
- What’s coming next
  - Confidence intervals

What is a confidence interval?

- Range of plausible values
  - Tries to quantify uncertainty associated with the sampling process.

Speaker notes

Speaker notes

We statisticians have a habit of hedging our bets. We always insert qualifiers into our reports, warn about all sorts of assumptions, and never admit to anything more extreme than probable. There's a famous saying: "Statistics means never having to say you're certain."

We qualify our statements, of course, because we are always dealing with imperfect information. In particular, we are often asked to make statements about a population (a large group of subjects) using information from a sample (a small, but carefully selected subset of this population). No matter how carefully this sample is selected to be a fair and unbiased representation of the population, relying on information from a sample will always lead to some level of uncertainty.

A confidence interval is a range of values that tries to quantify uncertainty associated with the sampling process.

Consider it as a range of plausible values.

There is a confidence level associated with any confidence interval, usually 95%, but sometimes 90% or 99%.

The confidence level is related to the alpha level (probability of a Type I error).

It also has a long range sampling interpretation.

If you repeatedly sampled from the same population, then 95% (or 90% or 99%) of the confidence intervals produced would contain the true value in the population.

Example of a confidence interval

- Homeopathic treatment of swelling after oral surgery
  - 95% CI: -5.5 to 7.5 mm
  - Lokken P, Straumsheim PA, Tveiten D, Skjelbred P, Borchgrevink CF. Effect of homoeopathy on pain and other events after acute trauma: placebo controlled trial with bilateral oral surgery BMJ. 1995;310(6992):1439-1442.

Speaker notes

http://www.bmj.com/content/310/6992/1439.full

Always look for confidence intervals that are wide enough to drive a truck through. They are very good indicators of small sample sizes.

Consider a recent study of homoeopathic treatment of pain and swelling after oral surgery (Lokken 1995). When examining swelling 3 days after the operation, they showed that homoeopathy led to 1 mm less swelling on average. The 95% confidence interval, however, ranged from -5.5 to 7.5 mm. From what little I know about oral surgery, this appears to be a very wide interval. This interval implies that neither a large improvement due to homoeopathy nor a large decrement could be ruled out.

Now, you can’t drive a truck through an interval that goes from -5.5 to 7.5 mm, but from the perspective of a human mouth, this interval is huge. Generally when a confidence interval is very wide like this one, it is an indication of an inadequate sample size, an issue that the authors mention in the discussion section of this paper.

Confidence interval interpretation (1 of 7)

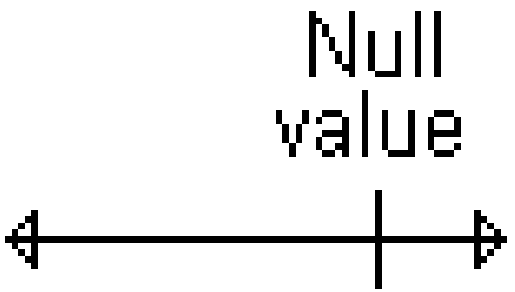


Figure 2: Interval that contains the null value

Speaker notes

When you see a confidence interval in a published medical report, you should look for two things. First, does the interval contain a value that implies no change or no effect? For example, with a confidence interval for a difference look to see whether that interval includes zero. With a confidence interval for a ratio, look to see whether that interval contains one.

Here’s an example of a confidence interval that contains the null value. This interval implies no statistically significant change.

Confidence interval interpretation (2 of 7)

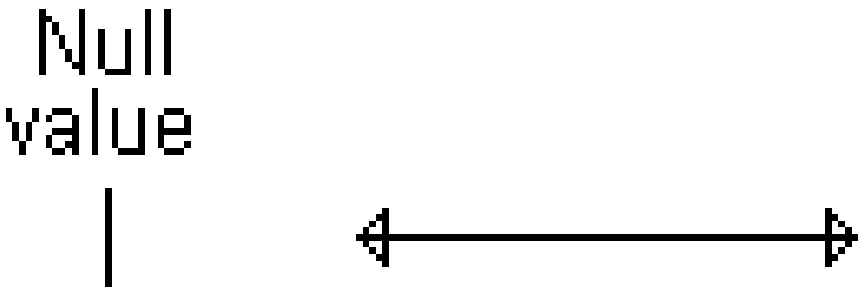


Figure 3: Interval entirely above the null value

Speaker notes

Speaker notes

Here’s an example of a confidence interval that excludes the null value. If we assume that larger implies better, then the interval would imply a statistically significant improvement.

Confidence interval interpretation (3 of 7)

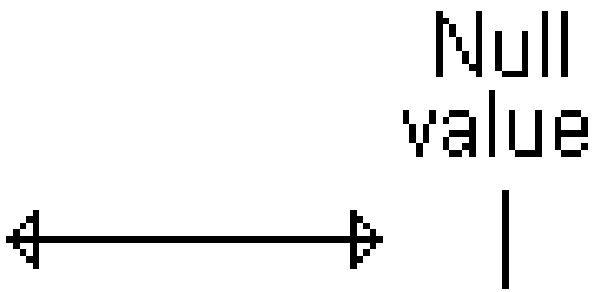


Figure 4: Interval entirely below the null value

Speaker notes

Speaker notes

Here’s a different example of a confidence interval that excludes the null value. This interval implies a statistically significant decline.

Confidence interval interpretation (4 of 7)

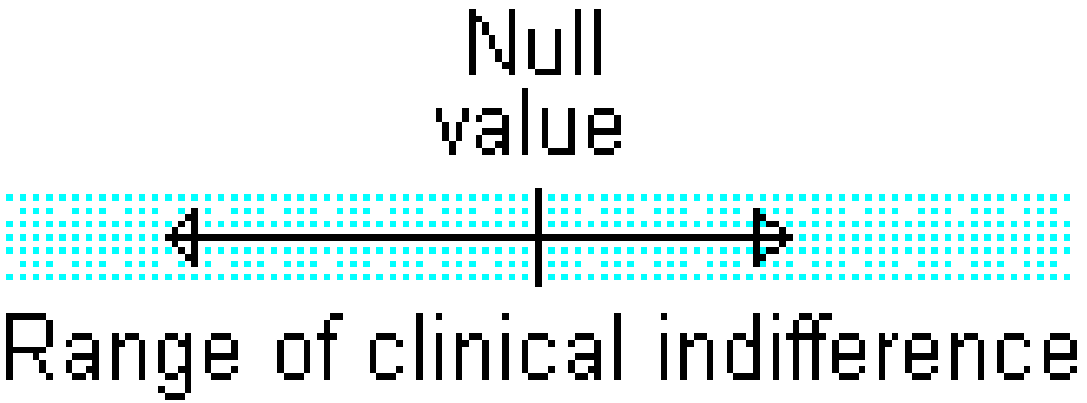


Figure 5: Interval entirely inside the range of clinical indifference

Speaker notes

You should also see whether the confidence interval lies partly or entirely within a range of clinical indifference. Clinical indifference represents values of such a trivial size that you would not want to change your current practice. For example, you would not recommend a special diet that showed a one year weight loss of only five pounds. You would not order a diagnostic test that had a predictive value of less than 50%.

Clinical indifference is a medical judgment, and not a statistical judgment. It depends on your knowledge of the range of possible treatments, their costs, and their side effects. As statistician, I can only speculate on what a range of clinical indifference is. I do want to emphasize, however, that if a confidence interval is contained entirely within your range of clinical indifference, then you have clear and convincing evidence to keep doing things the same way.

Confidence interval interpretation (5 of 7)

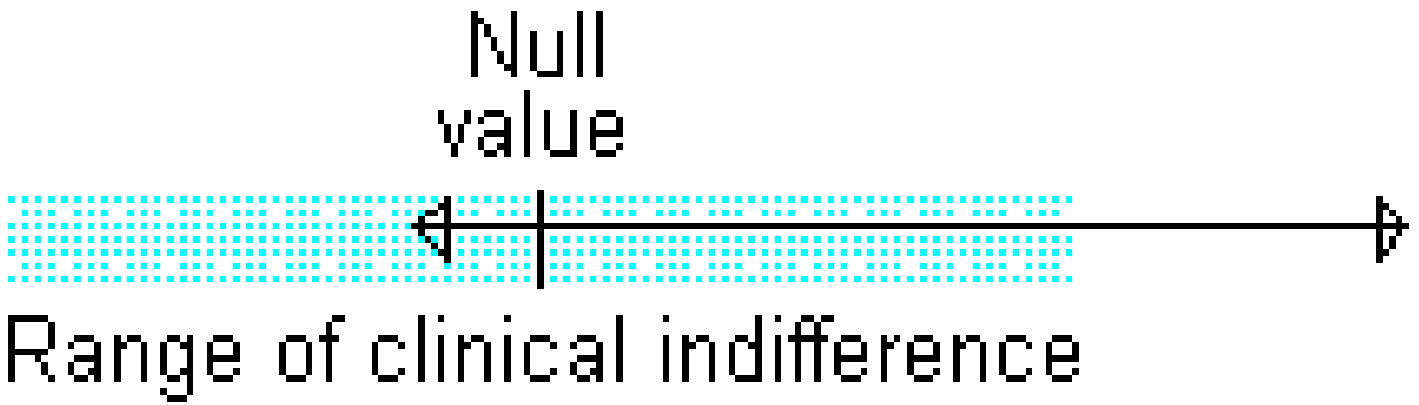


Figure 6: Interval partly inside/outside range of clinical indifference

Speaker notes

One the other hand, if part of the confidence interval lies outside the range of clinical indifference, then you should consider the possibility that the sample size is too small.

The interval contains zero, so it is plausible to behave as if the difference in population means or proportions is zero. But the interval also contains values that are clinically important. So it is plausible to behave as if there is a clinically important difference in means. How can you have two such different interpretations being plausible at the same time? That’s the definition of ambiguity. If you don’t like it, get used to it. Statistics will often identify areas of ambiguity, which is a good thing, because it tells us to not act prematurely, but instead demand more data before you make a definitive decision.

Quiz question, revisited

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. This confidence interval tells that the result is

1. statistically significant and clinically important.
2. not statistically significant, but is clinically important.
3. statistically significant, but not clinically important.
4. not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

Speaker notes

Let’s go back to that question I posed earlier.

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

Well, it tells you that a relative risk of 1 (equal risks) is plausible, but that a relative risk of 2 (a doubling of risk) is also plausible. A tripling of risk is plausible. Good grief! This is an ambiguous result.

Doesn’t this bother you? It should. Someone ran a terrible experiment. An experiment so poorly designed that it can’t distinguish between no change in risk, or a tripling of risk.

It’s a terrible thing, but it happens all the time and it doesn’t seem to bother anyone but me. This is wretched. You got a hundred patients to let you poke and prod them. They took some bitter pills or maybe placebos. They are sacrificing their bodies in the name of science. And the best you can do is a confidence interval that goes from 0.82 to 3.94. Hang your head in shame!

Confidence interval interpretation (6 of 7)

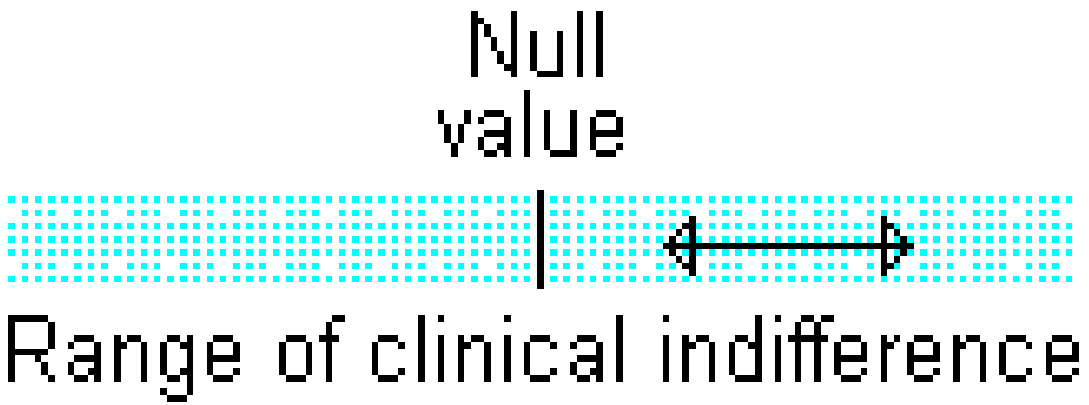


Figure 7: Confidence interval entirely inside the range of clinical indifference

Speaker notes

Some studies have sample sizes that are so large that even trivial differences are declared statistically significant, especially in this era of big data. If your confidence interval excludes the null value but still lies entirely within the range of clinical indifference, then you have a result with statistical significance, but no practical significance.

Confidence interval interpretation (7 of 7)

Figure 8: Confidence interval entirely outside the range of clinical indifference

Speaker notes

Speaker notes

Finally, if your confidence interval excludes the null value and lies outside the range of clinical indifference, then you have both statistical and practical significance.

Let’s talk about one more case. I don’t have a picture, but imagine a confidence interval that is mostly in the white region, the region of clinical importance, but the lower limit stretches into the range of clinical indifference. It doesn’t quite include the null value, but it comes within kissing distance. That’s a result that achieves statistical significance, but it does not provide definitive proof of clinical importance. No one ever talks about this case, but they should. Your confidence interval indicates statistical significance, but just barely. So don’t pretend that your results are the final word. You should not stop researching until you get a confidence interval that lies entirely inside or entirely outside the range of clinical indifference.

Why you might prefer a confidence interval

- Provides same information as p-value,
  - Clinical importance
  - Distinguish between
    - definitive negative result, or
    - more research is needed

Break #6

- What you have learned
  - Confidence intervals
- What’s coming next
  - Grading rubric

Grading rubric: general requirements.

- Documentation
- Graphs
- Tables
- Readability
- Interpretation

Speaker notes

Some of the general requirements listed below are not easy to meet. You cannot always rely on the default options that your program produces. Pay attention in class because you will see examples of how to meet these general requirements.

Documentation is required!

- Documentation should include
  - the name of the author (you!),
  - the creation date,
  - the purpose of your program, and
  - any restrictions on use.

Speaker notes

You must include a documentation header at or near the top of every programming assignment. Your documentation header must include

You can choose to place your program in the public domain (no restrictions) or you can specify restrictions on how others can use your program (e.g., no commercial use). This is entirely your discretion.

Graphs cannot rely on default choices

- Modify your graphs
  - Include your name and date on the title of any graph
    - “Steve Simon produced this graph on 2023-09-19.”
  - Avoid the display of unnecessary decimal places on the axes
  - Use comma separators for large numbers
  - Replace category codes with descriptive labels

Speaker notes

If the assignment requires graphs, you must modify the graphs to publication quality. This means you change how numbers and categories are displayed on the graph.

First, please be sure to include your name and the date the graph was first created on the title.

Look at what appears on the graph axes. Avoid axes that go 10.00, 20.00, 30.00, etc. If the numbers are large, use comma separators (anything 1,000 or larger). Use descriptive labels rather than gender=1 and 2.

- More modifications
  - Replace short variable names with longer descriptors
    - Include units of measurement, if needed
  - Avoiding the gratuitous use of color
    - Unless needed to distinguish between groups
    - Fill boxes and points with white/transparent colors

Speaker notes

Don’t use the default variable names if they are short and use abbreviations (change “bw” to “Birth weight”). Make sure that the units of measure (meters, kilograms, etc.) are listed. Some variable, of course, may not have a measurement unit.

Avoid the gratuitous use of color. If you need color to distinguish between groups, that’s fine. Otherwise, fill boxes and points with either white or transparent colors.

Tables also need modification

- Round to two or three significant figures
- Use comma separators if numbers are >= 1,000
- Avoid scientific notation (e.g., 1.23E-04)
- Avoid small p-values (e.g., p=0.000)
  - Change to p<0.001
- Suppress the printing of unneeded tables
  - Sometimes difficult
- Unmodified table acceptable with a detailed text description

Speaker notes

If the assignment requires the production of tabular data (other than a listing of the data itself), you must ensure that the table is easily readable. This means rounding values and using comma separators either in the table itself or in your interpretation of the table. Avoid the use of scientific notation.

ny p-values smaller than 0.001 should be rounded up to “0.001” or replaced with “<0.001”. A p-value of “0.000” is not acceptable.

Some programs (SPSS in particular) make it difficult for you to modify the tables. You are allowed to print a table without modification, but only if it is followed by a detailed text description that follows these rules.

Your code must be easy to read

- Make liberal use of
  - blank lines
  - line breaks
  - indenting
  - vertical lists

Speaker notes

Your code should be easy to read. You should include a lot of white space. Put a blank line between sections of your code. Break long lines into several lines with indents. Use vertical lists.

If you are using a system like SPSS that is largely menu driven, please include the syntax in a separate file and modify the syntax to make it more readable.

Always include an interpretation

- Use simple evaluative words
  - Young/Elderly
  - Less than half/more than half
  - Almost all/almost none
  - Substantial improvement/roughly comparable
- Depends on context
  - No penalty for subjective judgements

Speaker notes

For anything more complicated than a simple listing of the data, you should provide an interpretation. If you report an average age of 70.2 and a range of 65 to 85, it would be reasonable to use the word “elderly” in your description. A proportion of 0.45 could be described as “less than half” and a proportion of 0.89 as “almost all”. A difference between two groups might include evaluative words like “substantial improvement” or “roughly comparable”. Characterize a confidence interval as either wide or narrow and note whether it includes the null value. State the practical implications if you accept the null hypothesis.

The actual interpretation depends largely on the context of the problem and you will not be penalized for subjective assessments. You can call a small difference as large or a large difference as small. You will lose points, however, if you fail to include any description.

Depending on the software you use and the question you are addressing, you might present your interpretation as comments in your code, as a title on your graph or table, or as a separate block of text.

Grading rubric for a ten point assignment

- Four points for general requirements
  - Documentation
  - Graphs
  - Tables
  - Readability
  - Interpretation)
- Six points for Accuracy

Speaker notes

This grading rubric is for an assignment worth 10 points. Change these numbers proportionately for an assignment worth a different number of points (e.g, triple all the point values for an assignment worth 30 points).

- How well did you follow the general requirements specified above?
  - Poor (0 points) Two or more major problems with documentation, graphs, tables, readability, or interpretation.
  - Fair (2 points) One major or two minor problems with documentation, graphs, tables, readability, or interpretation.
  - Good (4 points) Includes all required elements

Speaker notes

Try as best you can to meet all of the general requirements shown above. For some assignments you cannot be evaluated on every element. For example, your assignment may not include any graphs.

It is the number of problems and whether the problems are major or minor that determine your score on this element.

- How accurate was your work?
  - Poor (0 points): Three or more major errors or omissions
  - Fair (2 points): Two major or four minor errors or omissions
  - Good (4 points): One major or two minor errors or omissions
  - Excellent (6 points): Complete and accurate answers to all questions

Speaker notes

Accuracy means providing a complete and correct answer to each question. it is the number of errors and whether they are major or minor that will determine your points.

File history

This file was written by Steve Simon on 2023-08-15 with the last major revision on 2024-08-20. It is in the public domain and you can use it any way you please.

Break #7

- What you have learned
  - Grading rubric
- What’s coming next
  - A simple R program

albuquerque-housing, 1 of 6

```
data_dictionary: albuquerque-housing

format:
  txt: tab-delimited
  csv: comma-delimited
  sas7bdat: proprietary SAS
  sav: proprietary SPSS

varnames: first row of data

missing_value_code: '.'
```

albuquerque-housing, 2 of 6

```
description: |
  From the original source (no longer available)  A random sample of records
of resales of homes from Feb 15 to Apr 30, 1993 from the files  maintained by
the Albuquerque Board of Realtors. This type of data is collected by multiple
listing agencies in many cities and is used by  realtors as an information
base.

download_url:
https://raw.githubusercontent.com/pmean/datasets/master/albuquerque-
housing.csv
```

albuquerque-housing, 3 of 6

```
source: |
  DASL (Data and Story Library), a repository for various data sets useful for
teaching. This file was lost in the transition of DASL from statlib to
datadescription.

copyright: |
  Unknown. You should be able to use this data for individual educational
purposes under the Fair Use guidelines of U.S. copyright law.

size:
  rows: 117
  columns: 7
```

albuquerque-housing, 4 of 6

```
price:
  label: Sales price of house
  scale: ratio
  unit: dollars
sqft:
  label: Square footage of house
  scale: ratio
  unit: square feet
age:
  label: Age of house
  scale: ratio
  unit: years
```

albuquerque-housing, 5 of 6

```
features:
  label: Number of features of house
  scale: ratio
  range: 0 to 13
northeast:
  label: Is house located in Northeast Albuquerque?
  scale: nominal
  value: yes/no
```

albuquerque-housing, 6 of 6

```
custom_build:
  label: Is the house custom built?
  scale: nominal
  value: yes/no
corner_lot:
  label: Is the house on a corner lot?
  scale: nominal
  value: yes/no
```

simon-5501-01-template.qmd, 1 of 5

```
---
title: "Template for 5501-01 programming assignment"
author: "Steve Simon"
format:
  html:
    embed-resources: true
date: 2024-08-18
---

This program reads data on housing prices in Albuquerque, New Mexico in 1993.
Find more information in the [data dictionary][dd].

[dd]: https://github.com/pmean/datasets/blob/master/albuquerque-housing.yaml

This code is placed in the public domain.
```

Speaker notes

The first few lines are the documentation header

simon-5501-01-template.qmd, 2 of 5

````
## Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages
and warnings are suppressed.

```{r setup}
#| message: false
#| warning: false
library(tidyverse)
```
````

simon-5501-01-template.qmd, 3 of 5

````
## Read the data and view a brief summary

Use the read_csv function to read the data. The glimpse function will produce
a brief summary.

```{r read}
alb <- read_csv(
  file="../data/albuquerque-housing.csv",
  col_names=TRUE,
  col_types="nnnnccc",
  na=".")
glimpse(alb)
```
````

simon-5501-01-template.qmd, 4 of 5

````
## Calculate overall means

The summarize_if function produces means, but only for numeric data. You
wouldn't want to compute means for data with values "yes" and "no".

```{r means}
alb |>
  summarise_if(is.numeric, mean, na.rm = TRUE)
```
````

simon-5501-01-template.qmd, 5 of 5

```
## Summarize price

The average price of a home, 106 thousand dollars, is quite low because the
data comes from 1993.

## Summarize sqft

## Summarize age

## Summarize features
```

Summary

- What you have learned
  - About this class
  - R and RStudio
  - History of R
  - Scales of measurement
  - Tests of hypothesis
  - Confidence intervals
  - Grading rubric
  - A simple R program