Introduction

Outline for today:

- About the course
- Course objectives
- About the instructor
- Why we use R
- Organizing data for analysis in R
- Review of some basic concepts in statistics
- First Discussion paper

Course components:

- Presentations, discussions of papers: Tuesdays 1 - 2 pm (Zoom, video on, not recorded)
- Lectures: Tuesdays 2 - 3 pm (Zoom for now, recorded)
- R workshops: Thursday afternoon 1 - 3 pm (Zoom for now)
- Assignments: Three in total.
- No exams, but course will extend one week past official end of classes
- No textbook, but various sources, many available online. See www.zoology.ubc.ca/~bio501/R/
- First presentation and discussion is next Tuesday –
  I need two volunteers to give presentation and two to moderate!
- First workshop is this Thursday – *Introduction to R*.

Lecture days:

- All students will have read the paper.
- Presentation of the paper: two students, 20-25 min:
  - Analyze the topic, using the reading as a starting point.
  - Explain key points. Include additional opinions.
- Discussion 30 min. Whole class participation, aided by two moderators:
  - Pick up where the presenters left off (coordinate with them).
  - Have several questions ready to start discussion or keep it going if it lags.
  - Moderate the discussion that emerges.
- 5-10 minute break.
- Lecture on the current topic.

Workshop days:

- First workshop is this Thursday! (Same Zoom link as lecture.)
- Send me your Zoom email address right now! (Top right corner.)
- Use own computers. Have latest R version installed (4.1.2).
- Work through problems on the web site during the workshop.
- The “R tips” web pages contain most clues needed to carry out workshops.
- Try solving by yourself. If it is not working, ask your neighbor or breakout room member for help. Then ask me. Admit if you are stuck. You are not alone.
- 2 hours is allocated for each workshop. Workshops take longer than 2 hours. There is no specified portion of workshops that must be completed. The further you go, the more you will learn. You control that.
- Workshops are online. Later you can go back and get more practice as needed.

Grading based on:

- Assignments (50%) (Try *R Markdown*, but not required)
- Paper presentation to class (20%)
- Discussion moderating (20%)
- Back bench participation in discussion (10%)
- Presenting and moderating: I will give guidance as to what is expected of you.
- I expect you to rise to the occasion. Less guidance is given than you might have experienced before now.

Web site

- www.zoology.ubc.ca/~bio501/R/
- Updated regularly – hit your browser refresh button.
- Lecture overheads will be placed there in pdf format before lecture time.
- Discussion papers and assignments are placed there.
- Recommended stats and R books.
- The “R tips” help pages.
- The *Calculate* and *Data* R tips pages will be useful for this week’s workshop

Tentative list of lecture topics

1. Introduction
2. Graphics
3. Experimental design
4. Linear models
5. Mixed-effects models
6. Likelihood
7. Generalized linear models
8. Model selection
9. Bayesian methods
10. Computer intensive
11. Meta analysis
12. Multivariate methods
13. Species as data points

The course

- Was developed in response to needs identified by grad students in the BRC.
- Help me to improve it.
- This is a “second” course in data analysis, to take you beyond the most basic, introductory level, which I’m assuming you have already (up to ANOVA and linear regression).
- Those who eavesdrop on Zoom are welcome.

Textbook

- No required textbook.
- Course web site lists useful books, many available online.
- Use *Whitlock and Schluter (3rd ed. 2020)* or alternative as a basic stats reference.
- I like *Quinn and Keough (2002) Experimental design and data analysis for biologists* as a more advanced stats reference book.

About the instructor

- I’m not a statistician, but have basic training (and learned from many mistakes).
- I can answer most of your stats questions immediately, but some answers might require me to return to you later.
- I started using R because my students used it. I used S and S-Plus before there was R, but S-Plus was expensive and ran only on the PC. (Now extinct. R won.)
- I am not a complete R expert. I have a head start on most of you. Maybe I will hold this edge until the end of term.
- You will discover things that I don’t know about. You will tell me these things.
- I have used R on Windows and Mac, they are the almost same.
- My office hours: Tuesdays 3 - 5 pm (Zoom link in email).

Course aims:

- To help prepare you for research by reviewing the basic principles for designing good studies, gathering and organizing data, and properly analyzing those data.
- To increase your understanding of concepts in data analysis, and appreciation for magnitudes (effect sizes) over *P*-values.
- To introduce you to innovative approaches increasingly used in biology to analyze data.
- To show you an amazing statistical environment in which to analyze data: R.
- Broad coverage of current methods, rather than deep foundation on few topics.
- It is expected that in your research you will need to delve more deeply into those particular methods that turn out to be most appropriate.
- “Linear models” will be our framework, and we will start there and generalize.

What is R?

R is a language and environment for statistical computing and graphics.

It is a GNU Project, free under the terms of the Free Software Foundation's GNU General Public License.

R was inspired by the S environment, developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.

R was initially written by **R**obert Gentleman (Canadian! UBC grad!) and **R**oss Ihaka.

The current R is the result of a collaborative effort involving contributors from all over the world.

Good things about R

1. Powerful and flexible.
2. Free!
3. Runs on all computer platforms.
4. Many extensions (contributed packages). New ones always coming online.
   emmeans – magnitudes for linear model fits
   ggplot2 – graphics tools
   popbio – analyzing matrix population models
   qtl – QTL analysis
   shapes – geometric morphometrics
   vegan – ordination in community ecology
   visreg – visualize linear model fits

Good things about R

5. Superb data management and manipulation capabilities
6. Superb ability to carry out operations on arrays of numbers.
7. Superb graphics capabilities.
    - Visualize data and model fits
    - Produces vectorized graphs (pdf or eps format), permitting editing in a graphics package (e.g., Inkscape)
8. Reproducibility of analyses. R uses scripts to execute commands rather than menus.

Good things about R

9. It can be easy to do things that are difficult or impossible to do in other packages.
10. You can write your own functions to speed up your specific needs.
11. It is also a great programming tool.
12. It has a large community of users.
13. Someone has already solved your problem. Google is the encyclopedia of everything R.

Bad things about R

1. R uses scripts to execute commands rather than menus and a mouse.
2. It can sometimes be difficult to do otherwise simple things.
3. It is not a great spreadsheet.
    - Use a dedicated spreadsheet program for text files (e.g., .csv).
4. There are several kinds of data objects to remember.
    - Vectors and data frames are most common.
    - You will learn others more gradually (list and matrix).
5. Some variation in command syntax, e.g., `plot()` vs `ggplot()`.
6. Quality control concerns? Core programs are well-tested, but newest add-ons need checking. Many people out there doing the checking too, and writing about problems.

Some data recommendations

Now is a good time to think about your own strategy for data entry, storage, and organization.

Modified from: Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. Bulletin of the Ecological Society of America 90: 205-214.

Some data recommendations

1. *Use a scripted program for analysis*

    a. For example: R (Other programs include MATLAB and SAS).

    b. Menu-based programs leave no record of the analyses you carried out. You will forget. Menus change.

    c. Script files (commands) become your written records of your analyses.

    d. Surround commands in script files with detailed comments on your choices and actions.

    e. R Markdown makes human-readable documents with R script, analysis descriptions, and results (e.g., figures, tables).

    f. Use version control, which keeps a record of changes over time so that you can undo or retrieve earlier versions (e.g., keep script files on Github).

Some data recommendations

2. *Store data in a nonproprietary software format*

    a. For example: use comma- or tab-delimited text files.

    b. Text files can always be read, whereas proprietary formats can become unavailable in future.

    c. You can still use spreadsheet programs to develop the text files (Google Sheets, Excel, FreeOffice, LibreOffice).

3. *Store data in nonproprietary hardware formats*

    a. Keep data on the internet, which probably won't die any time soon.

    b. Backup copies at another location.

Some data recommendations

4. *Leave your data file uncorrected, with all its bumps and warts*

    a. Otherwise you might change something that you later discover was correct.

    b. Make corrections instead using R script so you have a record of the change, and can undo later if necessary.

    c. Corrections directly to the data file go unrecorded – you have no record of the change you made.

    d. Keep comments in your script (command) file that explains reasons for corrections, so you reference and even reevaluate later.

Some data recommendations

5. *Use descriptive names for your data files*

   a. Use names that are short but indicative of file contents.

   b. `sVancIsland_VegBiodiv_2007.csv` **not** `Veg.csv`.

Some data recommendations

6. *Include a "header" first line in data file with descriptive variable names*

   a. Variable names should be descriptive without blanks or commas.

   b. The `read.csv()` command in R assumes by default that the first line of a .csv file is a header line.

lizard	sprintSpeed1984
1	1.43
2	1.56
3	1.64
4	2.13
5	1.96
...	

← this is the header line

Huey, R. B. and A. E. Dunham. 1987. Repeatability of locomotor performance in natural populations of the lizard *Sceloporus merriami*. Evolution 42: 1116-1120.

Some data recommendations

7. *Use plain ASCII text for names and data values*

   a. Plain ASCII includes all letters of English alphabet (uppercase and lowercase), numbers, and many common punctuation marks (_ - * . are ok)

   b. Avoid commas because they separate fields in .csv format

   c. Avoid symbols (e.g., α □ ©  □)

8. *When you add data to a database, add rows not columns*

   a. Set up data files to maximize consistency of column content

   b. Use “long” format rather than “wide”

Data on sprint speed in *wide* format:

lizard	sprintSpeed1984	sprintSpeed1985
1	1.43	1.37
2	1.56	1.30
3	1.64	1.36
4	2.13	1.54
5	1.96	1.82
...		

Data on sprint speed in *long* format:

lizard	sprintSpeed	year
1	1.43	1984
2	1.56	1984
3	1.64	1984
4	2.13	1984
5	1.96	1984
1	1.37	1985
2	1.30	1985
3	1.36	1985
4	1.54	1985
5	1.82	1985
...		

Some data recommendations

9. *A column of data should contain only one data type*
   (i.e., either numerical or character, not both)

R will interpret a column of numbers having even a single character as character data (non-numeric)

don't:

lizard	speed	year
1	1.43	1984
2	1.56	1984
3	1.64?	1984
4	2.13	1984
...		

do:

lizard	speed	year	comment
1	1.43	1984	ok
2	1.56	1984	ok
3	1.64	1984	dubious
4	2.13	1984	ok
...			

Some data recommendations

10. *Record full dates using standardized ISO format*

    a. For dates use YYYY-MM-DD (promoted by the International Organization for Standardization).
    Other formats can be ambiguous.

    b. For datetime use YYYY-MM-DDT hh:mm:ss (T is the time delimiter)

11. *Create a relational database*

    a. Put separate information collected at different scales into different files.

    b. For example: One file for SITE data (temperature, elevation). Another file for measurements of SPECIES collected at those sites. Both files contain the SITE variable, allowing data to be matched as needed (using `match()` command in R).

Some data recommendations

12. *Maintain effective metadata (data about the data)*

    a. Ten years from now you won't remember what the site looked like, which sample you left out of the analysis, or how you assigned a single value for "depth" of a pond.

    b. Record why you collected the data.

    c. Write down details of methods.

    d. Include names of all files associated with the study, definitions for data and treatment codes, missing value codes, definitions, unit of measurement for each variable.

    e. Consider using a metadata standard such as Ecological Metadata Language (EML).

Some data recommendations

13. *Deposit your data when you publish it (and even if you don't publish)*

    a. Online data archives, e.g., Dryad (https://datadryad.org/search), Ecological Archives, https://esapubs.org/archive/, Genbank

Finding data (besides collecting your own)

- I stole the lizard data from a scatter plot in the original article.
- There is no copyright on published data, which is useful when you need an example or are carrying out a meta-analysis.
- Graphics tool: https://www.datathief.org/
- Online data archives, e.g., Ecological Archives, https://esapubs.org/archive/, Genbank, Dryad (https://datadryad.org/search)
- Permissions/conditions may be required to publish results from archives.

Review some basic principles

1. What is the definition of probability?
2. The chance difference between an estimate from a random sample and the population parameter being estimated is called _____.
3. A systematic discrepancy between estimates of a parameter from a sample and the true value of the population parameter is called _____.
4. The probability distribution of values of a sample estimate that we *might* obtain when we sample a population, and their probability of occurrence, is called the _____.
5. The standard deviation of the probability distribution for a sample estimate is called the _____.
6. A range of values surrounding the sample estimate that is likely to contain the true value of the population parameter is called a _____ .
7. The probability of obtaining a discrepancy from the null hypothesis as extreme as that observed, if the null hypothesis were true, is called the _____.

True or false?

1. If a 95% confidence interval for the mean is calculated from a random sample as $0.9 \leq \mu \leq 3.1$, then there's a 95% chance that the mean lies between 0.9 and 3.1.

2. The probability that an individual is blinking when a photograph is taken is about 0.04. This means that if a photo of 2 people is taken, the probability that at least one is blinking is 0.04 + 0.04 = 0.08.

3. The *P*-value from a test of treatment effects in a clinical trial was calculated as 0.12. The null hypothesis of no treatment effect was <u>not</u> rejected ($P > 0.05$). We can conclude that the treatment is ineffective.

4. The *P*-value from a test of treatment effects in a clinical trial was calculated as 0.12. The null hypothesis of no treatment effect was <u>not</u> rejected ($P > 0.05$). We can conclude that the treatment effect (the difference between treatment means) is small.

5. The *P*-value from a test of treatment effects in a clinical trial was calculated as 0.001. The null hypothesis of no treatment effect was <u>soundly</u> rejected. We can conclude that the treatment effect (the difference between treatment means) is large.

How did you do so far?

If you fared poorly on the answers to these questions, review these core concepts in your stats book (e.g, Chapters 4 and 6 of Whitlock and Schluter (2015).

- Estimation vs hypothesis testing
- Sample estimate ("effect size") vs population parameter
- Probability
- Sampling distribution
- Standard error
- Confidence interval
- Effect size
- *P*-value

True or false?

6. The *P*-value from a test of treatment effects in a published clinical trial was calculated as 0.001. This means that if we repeated the experiment using the same population and sample size, it would likely again reject the null hypothesis of no treatment effect.

7. The effect size from a test of treatment effects in a published clinical trial was estimated to be large. This means that if we repeated the experiment using the same population and sample size, it would likely again estimate a large effect size.

True or false?

1. The *P*-value from a test of treatment effects in a published clinical trial was calculated as 0.001. This means that if we repeated the experiment using the same population and sample size, it would likely again reject the null hypothesis of no treatment effect.

2. The effect size from a test of treatment effects in a published clinical trial was estimated to be large. This means that if we repeated the experiment using the same population and sample size, it would likely again estimate a large effect size.

These last two questions concern the reproducibility of results from research studies.

Results reproducibility: obtain the same results from an independent study with procedures as closely matched to the original study as possible.

Inferential reproducibility: draw the same conclusions from either an independent replication of a study or a reanalysis of the original study.

Goodman et al (2016) *Science Translational Medicine* 8.341: 341ps12.

Results on reproducibility

A large-scale collaborative effort (27 authors) obtained an initial estimate of the reproducibility of psychological science (Open Science Collaboration 2015, *Science*). The study repeated 100 experimental and observational studies previously published in 3 psychology journals. Replications maintained high fidelity to the original designs, including sample sizes, in consultation with original authors.

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

facilitated each step of the process and maintained the protocol and project resources. Replication materials and data were required to be archived publicly in order to maximize transparency, accountability, and reproducibility of the project (https://osf.io/ezcuj).

In total, 100 replications were completed by 270 contributing authors. There were many different research designs and analysis strategies in the original research. Through consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs. Analyses converted results to a common effect size metric [correlation coefficient (r)] with confidence intervals (CIs). The units of analysis for inferences about reproducibility were the original and replication study effect sizes. The resulting open data set provides an initial estimate of the reproducibility of psychology and correlational data to support development of hypotheses about the causes of reproducibility.

Sampling frame and study selection

We constructed a sampling frame and selection

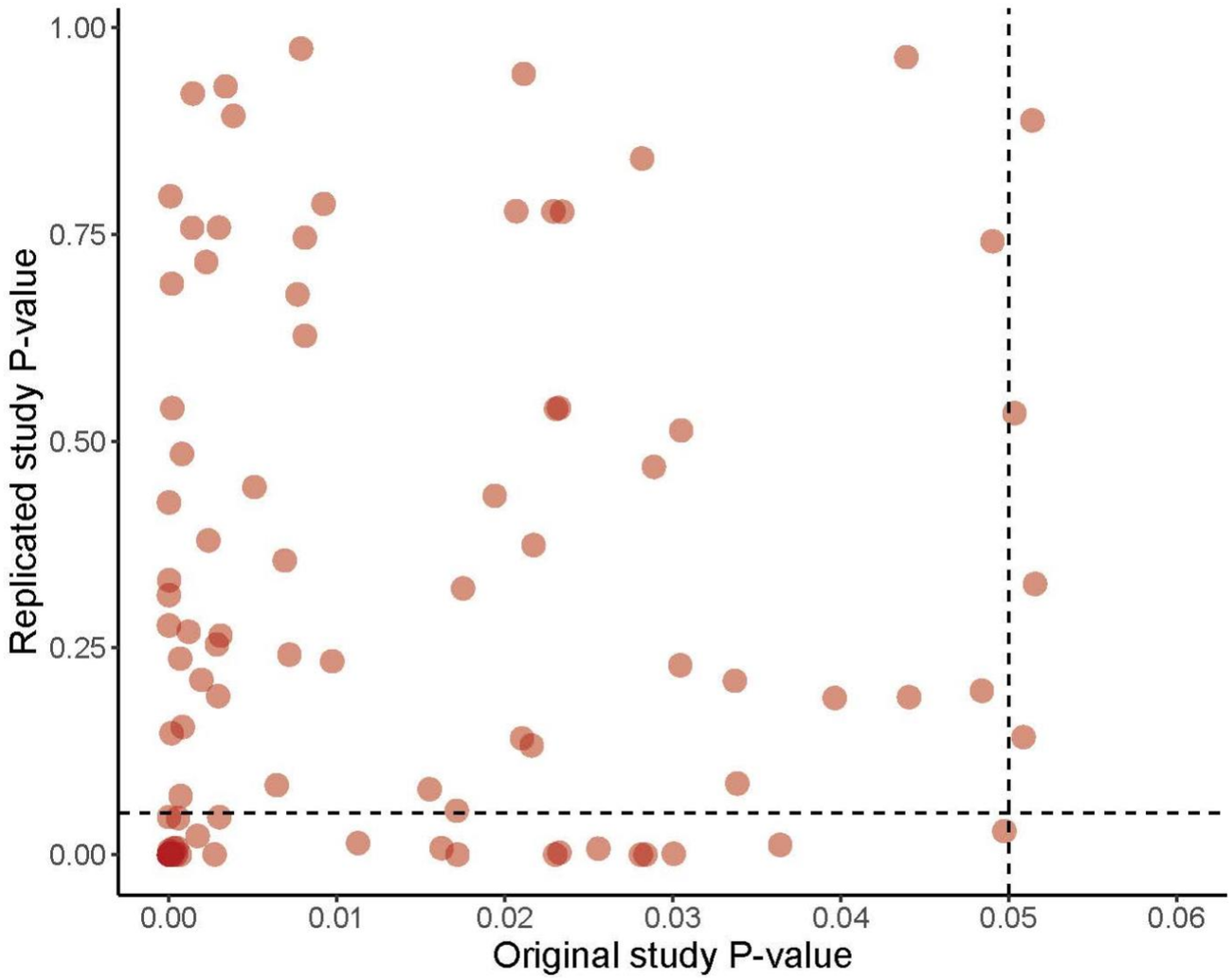
Results on reproducibility

97% of original studies had statistically significant results ($P < .05$), whereas 36% of replications had statistically significant results.

The median *P*-value of the replication effects (0.20) was about 30 times that of the original (0.007).

Correlation between *P*-values: $r_S = 0.30$.

(Note the different scales of the two axes.)

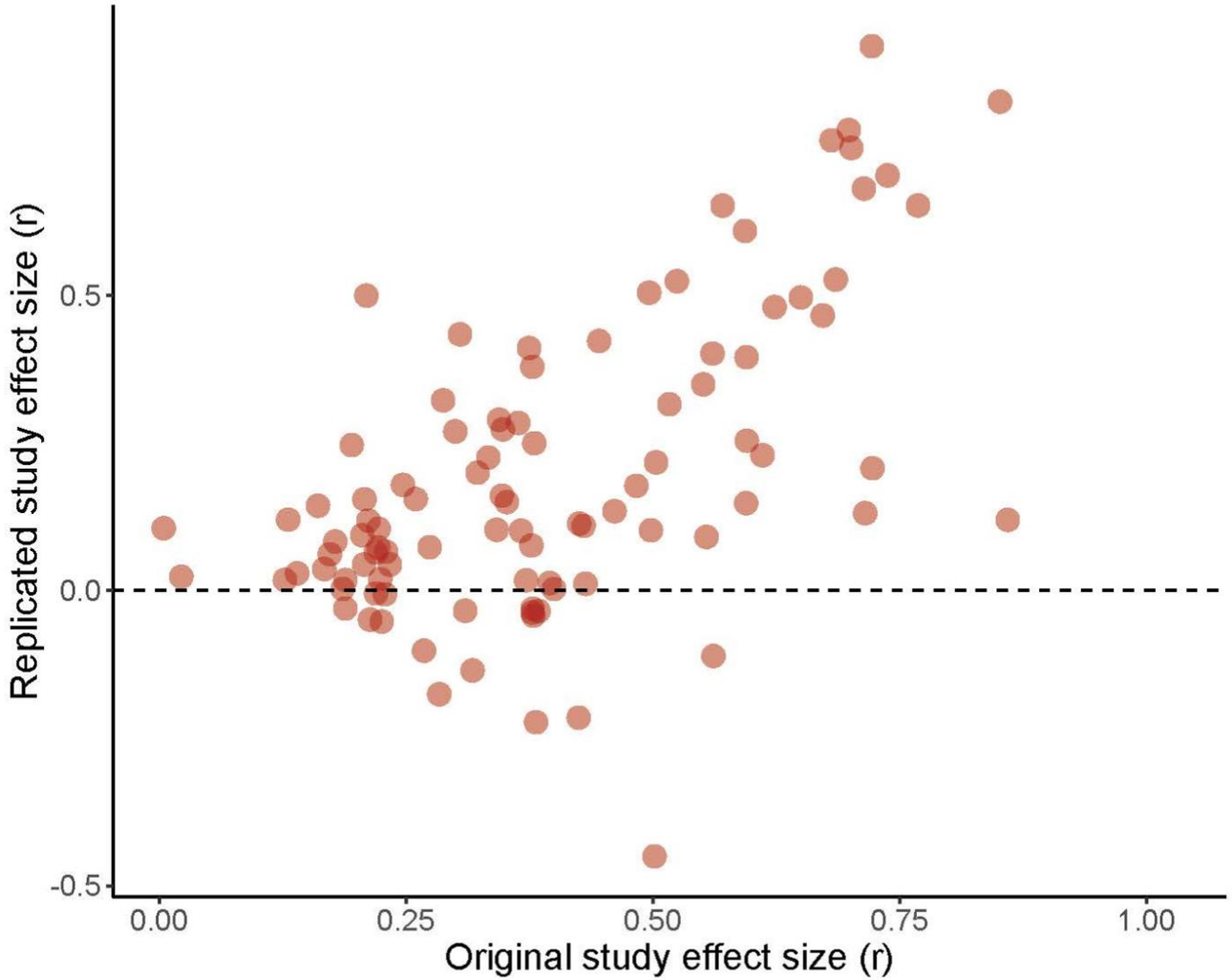


Results on reproducibility

The median effect size of the replication effects (0.12) was about a third that of original effects (0.40).

Correlation between effect sizes: r_S = 0.51.

Note the different scales of the two axes.



Results on reproducibility

How can these results be explained?

Published results are biased, the result of low-power research designs and publication bias. Effect sizes are more reproducible than *P*-values.

"Progress occurs when existing expectations are violated and a surprising result spurs a new investigation. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know."

An objective of this course is to improve understanding of concepts to aid in this process. There is much to do!

Discussion paper:

Wainer (1984) How to display data badly.

Find at the course web site.

Need two presenters for next week: 20-25 minute presentation

Need two discussion moderators