Biostatistics 生物统计学

(BIOL0031132104)

李 勤

qli@des.ecnu.edu.cn

https://qli.github.io/

华东师范大学·生态与环境科学学院

大纲 outline

- 关于这门课程
- 课程目标
- 关于授课老师和学生
- 统计学基本概念
- 课堂总结和讨论
- 为什么我们使用R

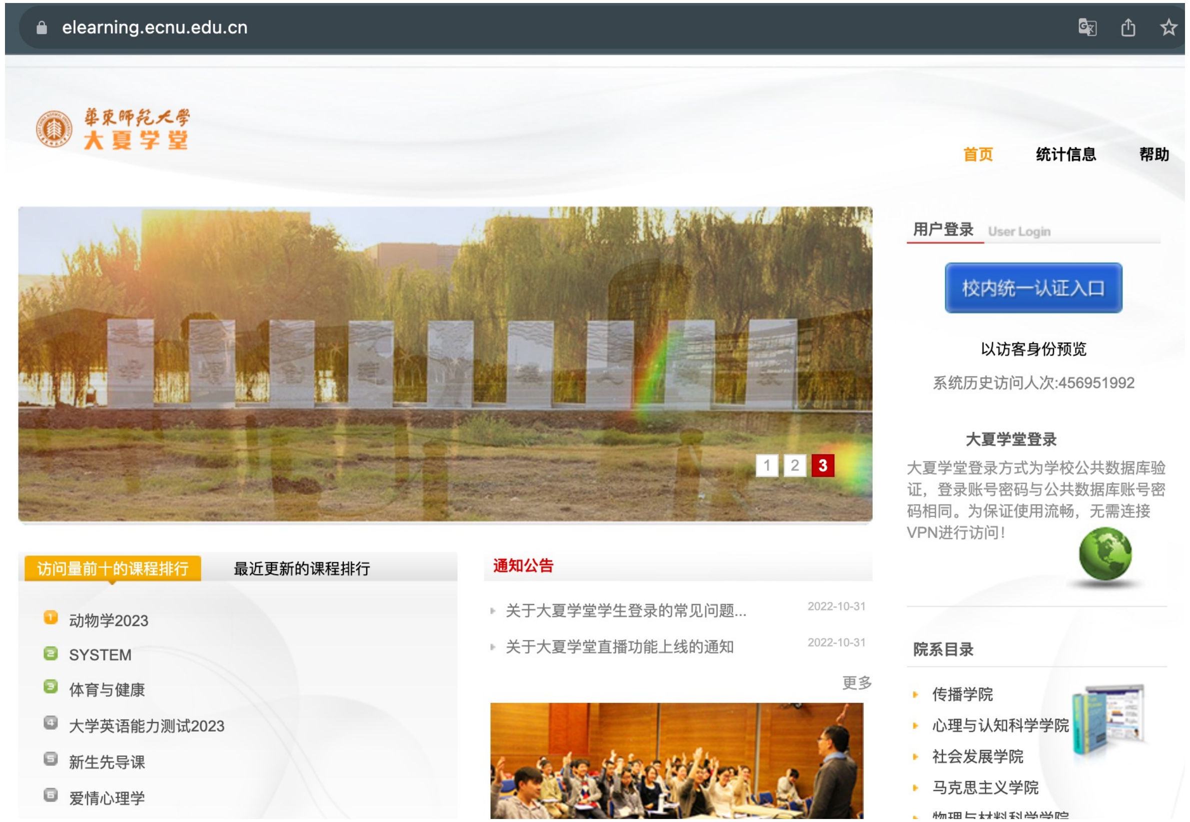
课程结构及考核
course components & grading

- 授课 (课堂参与/讨论/提问 10%)
  - 周一 (1:00 – 3:35 pm 6-8 闵二教113)
  - 周三 (9:50 am – 12:15 pm 3-5 闵二教213)
  - Lecture + R Lab
- 作业 (4次, 35%)
- 期中考试 (15%)
- 期末考试 (40%)

课程网站 website

Blackboard(大夏学堂)

- https://elearning.ecnu.edu.cn/



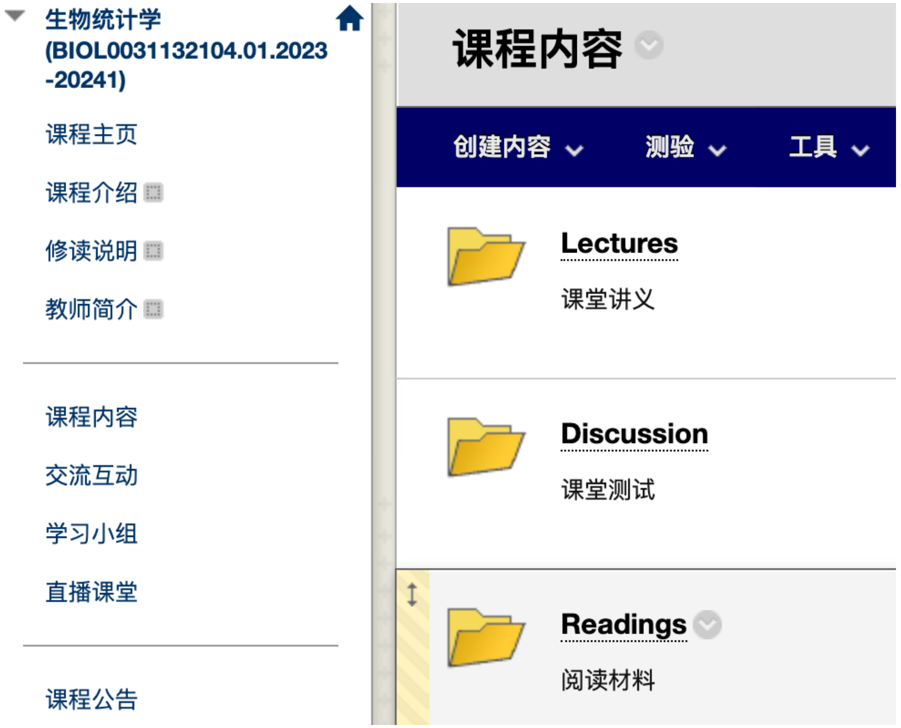
系统历史访问人次:456951992

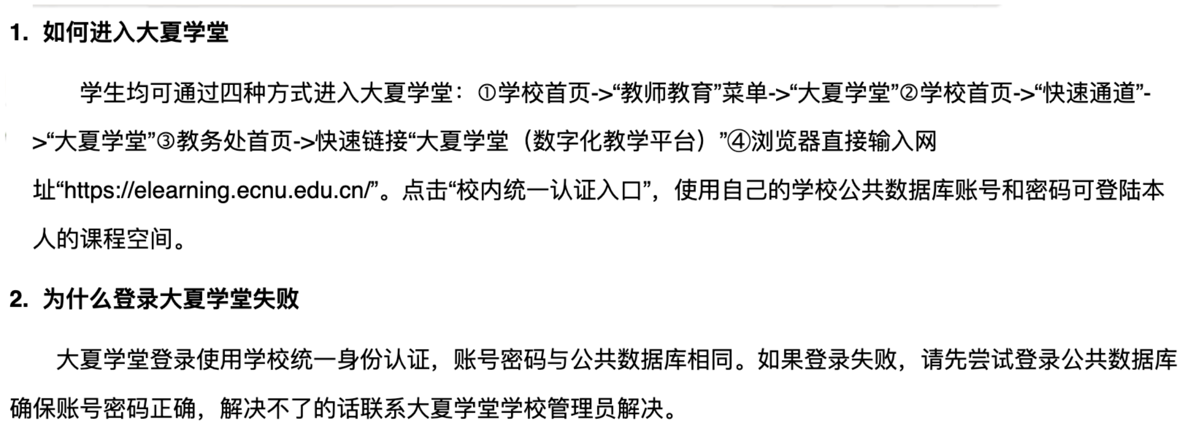
大夏学堂登录

大夏学堂登录方式为学校公共数据库验证，登录账号密码与公共数据库账号密码相同。为保证使用流畅，无需连接VPN进行访问！

关于大夏学堂学生登录的常见问题... 2022-10-31

关于大夏学堂直播功能上线的通知 2022-10-31





1. **如何进入大夏学堂**

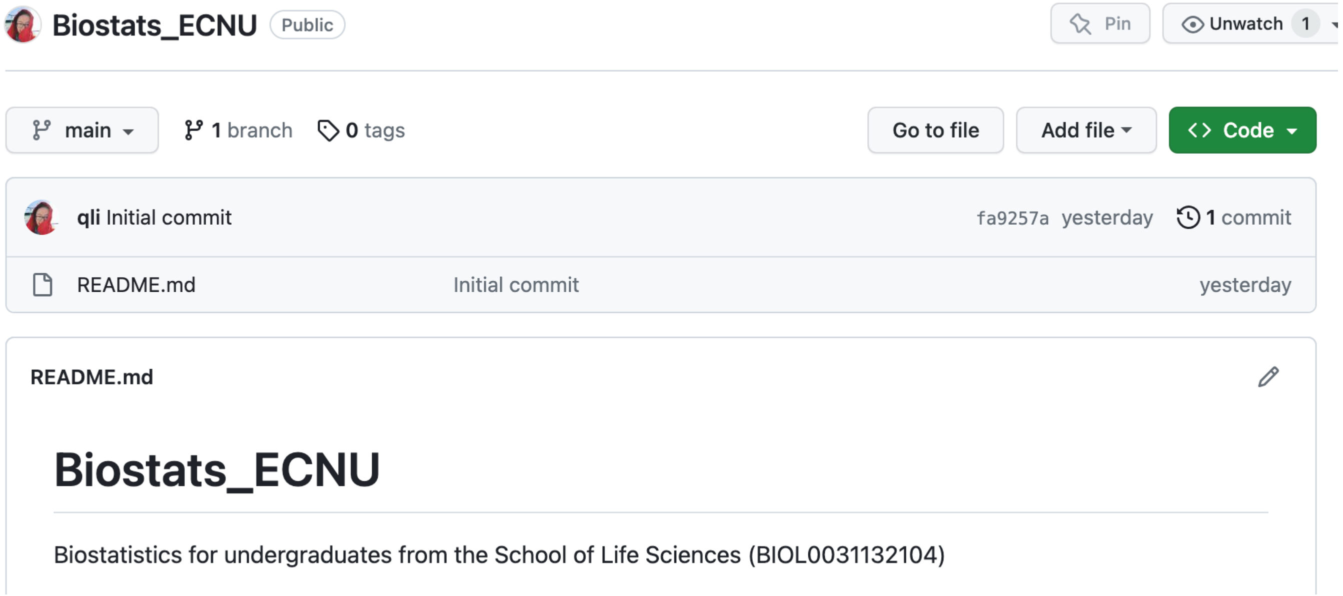
学生均可通过四种方式进入大夏学堂：①学校首页->“教师教育”菜单->“大夏学堂”②学校首页->“快速通道”->“大夏学堂”③教务处首页->快速链接“大夏学堂（数字化教学平台）”④浏览器直接输入网址“https://elearning.ecnu.edu.cn/”。点击“校内统一认证入口”，使用自己的学校公共数据库账号和密码可登陆本人的课程空间。

2. **为什么登录大夏学堂失败**

大夏学堂登录使用学校统一身份认证，账号密码与公共数据库相同。如果登录失败，请先尝试登录公共数据库确保账号密码正确，解决不了的话联系大夏学堂学校管理员解决。

课程网站 website

- https://elearning.ecnu.edu.cn/
- https://github.com/qli/Biostats_ECNU
- 课件PDF定期更新 (每周课前
- 作业也会在网站上发布
- 其它推荐的阅读资料
- 作业相关的R-tips



qli Initial commit fa9257a yesterday 1 commit

README.md Initial commit yesterday

README.md

Biostats_ECNU

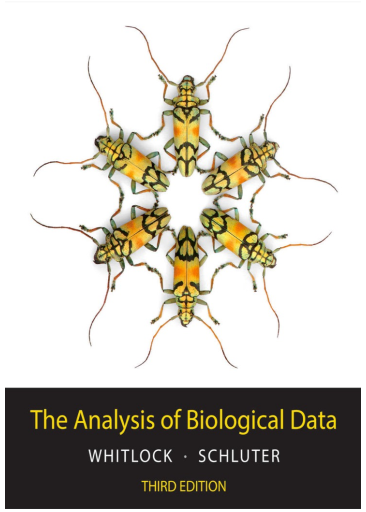
Biostatistics for undergraduates from the School of Life Sciences (BIOL0031132104)

课程大纲 Syllabus (tentative)

1. 绪论——统计学简介
2. 数据描述和展示
3. 概率分布1
4. 概率分布2
5. 比例和频率数据
6. 列联表分析
7. 假设检验
8. 实验设计
9. 两个均值的比较
10. 多个均值的比较
11. 相关性和因果性
12. 线性回归模型
13. 混合线性模型
14. 广义线性模型
15. 非线性回归
16. 多元统计分析

课程简介 About this course

- 这门课程基于
  - 华师大生环学院 - 邢丁亮老师 - 生物统计学
  - UBC - Dr. Dolph Schluter - Quantitative methods in ecology and evolution
- 没有特定教材 (主要以授课内容及推荐阅读为主)
  - Whitlock, M.C. & Schluter, D., The Analysis of Biological Data (3rd edn), W.H. Freeman Publishers, 2020
  - 李春喜, 姜丽娜, 邵云, 张黛静, 生物统计学 (第五版), 科学出版社, 2013
  - An Introduction to R, 2023 (v4.3.1) https://cran.r-project.org/doc/manuals/R-intro.pdf
  - Kabacoff, R., 王韬 (译), R语言实战 (R in action) (第五版), 人民邮电出版社, 2023



课程目标 Course objectives

- Understand basic concepts and methods in Statistics and their applications in Biology.
- Fundamental principles for building hypothesis tests, designing solid studies, collecting and organizing data, and conducting proper data analysis.
- Focus on data, instead of mathematical foundations of statistics.
- Develop analysis skills with the computational tools: R!
    - Steep learning curve (but, practice more/learn by doing).

授课教师 About the instructor

- I was trained in the fields of ecology, evolution and biogeography.
- My research focuses on diversification (niche/range/trait).
- I am not a statistics expert, but have learned from experience.
- I have used R 10+yrs, mainly for statistical analyses, and figures.
- I am not a R expert, but know how to find solutions.
- I won’t be able to answer all stats questions, but I’d like to work with you.
- My office hour: Wed 1-3 pm, Zihuan #329

学生 Students

- Who are you?
    - Major: Biology
    - You have learned Advanced Mathematics B.
- What do you want to take out of this course?
- What are your expectations with theories or skills learned from this course?

Lecture 1 – Introduction to Biostatistics

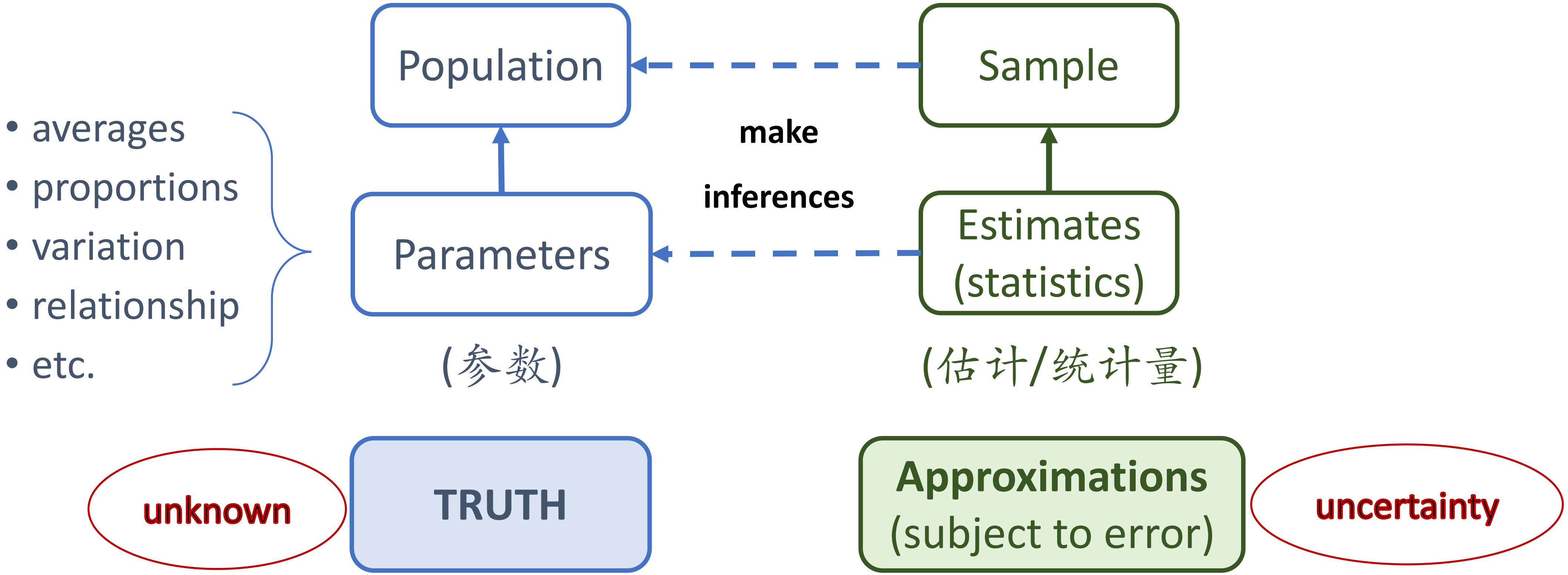
- Outline
  - What is statistics?
  - Sampling: basic concepts
  - Types of data and variables
  - Types of studies
  - Summary
  - Discussion

1. What is statistics?

- **Statistics** (统计学) is the study of methods to describe and measure aspects of nature from samples (样本).
- Statistics is about estimation (估计), the process of inferring an unknown quantity of a target population (总体) using sample data.
- Statistics also gives us tools to quantify the uncertainty (不确定性) of these measures, which mean their departure from the truth.

1. What is statistics?

- About estimation of population (总体), with sample data (样本).



- averages
- proportions
- variation
- relationship
- etc.

Population

Sample

make inferences

Parameters

Estimates (statistics)

(参数)

(估计/统计量)

unknown

TRUTH

Approximations (subject to error)

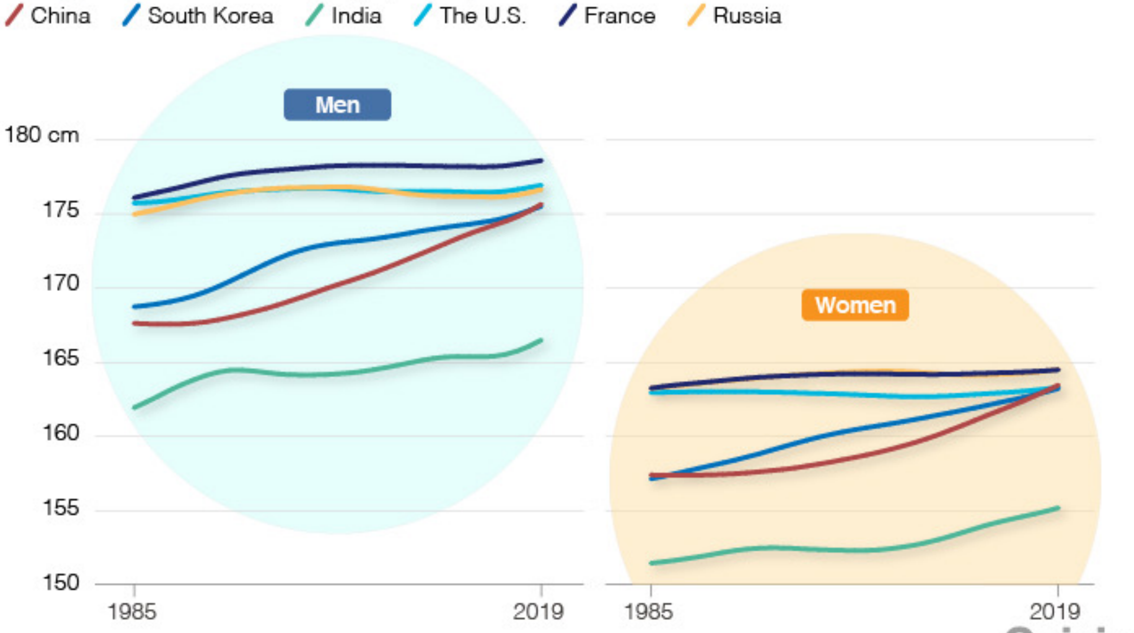
uncertainty

1. What is statistics?

- Examples: body heights
  - averages vs distributions

Chinese Youngsters Are Getting Taller and Taller

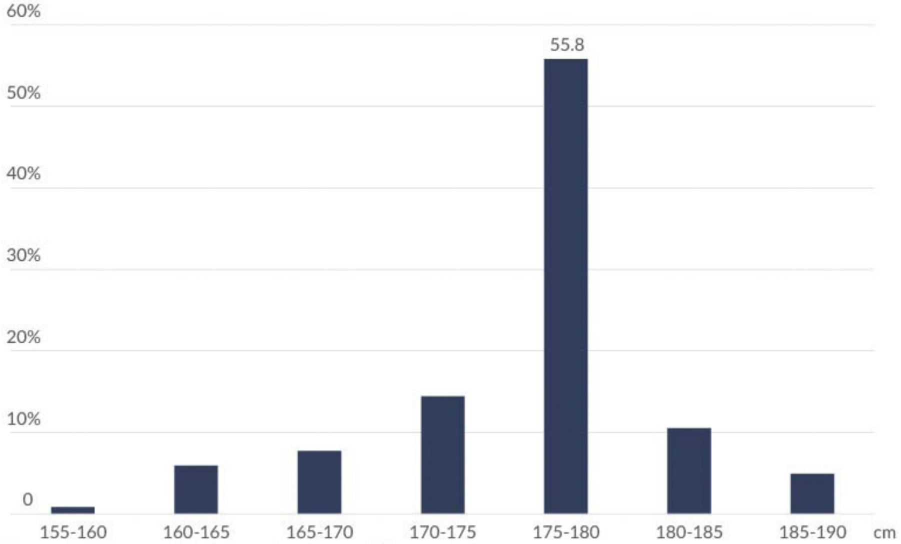
Average height at 19 years old (cm), by country



Source: NCD-RisC

Most Urban Chinese Men 20-25 Years Old Are Over 175 Centimeters

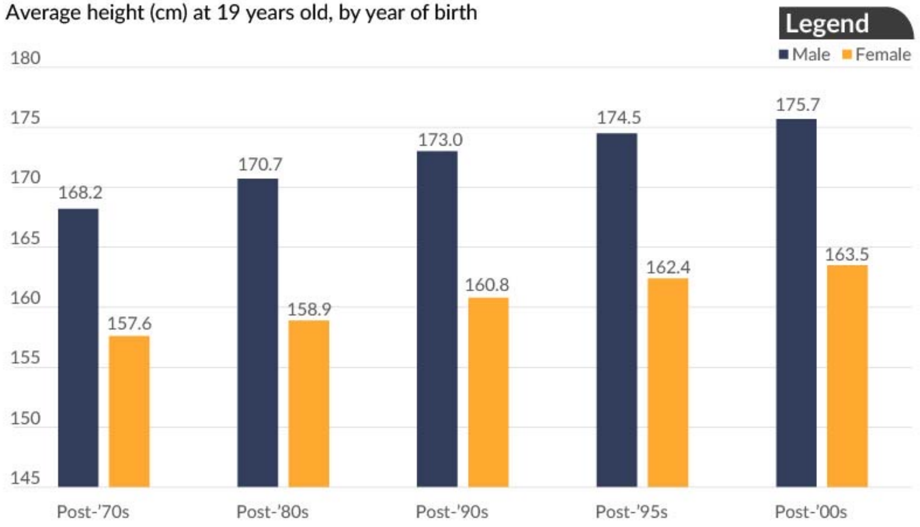
Proportion of urban males aged 20 to 25 whose height falls within the given range (%)



Source: Mu Rongrong, Analysis on the change of the morphology character of 20- to 25-year-old urban adults in China.

The Kids Are Getting Taller

Average height (cm) at 19 years old, by year of birth



Source: NCD RisC.

1. What is statistics?

- Examples: statistics is about estimation, with sample data.
    - Body heights

- Can you raise some real *biology questions* and think about how would you approach them?

1. What is statistics?

- Statistics is also about hypothesis testing (假设检验).
- A statistical hypothesis is a specific claim regarding a population parameter (总体的参数).
  - Example: body height
  - Hypothesis: Men have a higher average height than women.
  - How to test it?

🚹 > 🚺

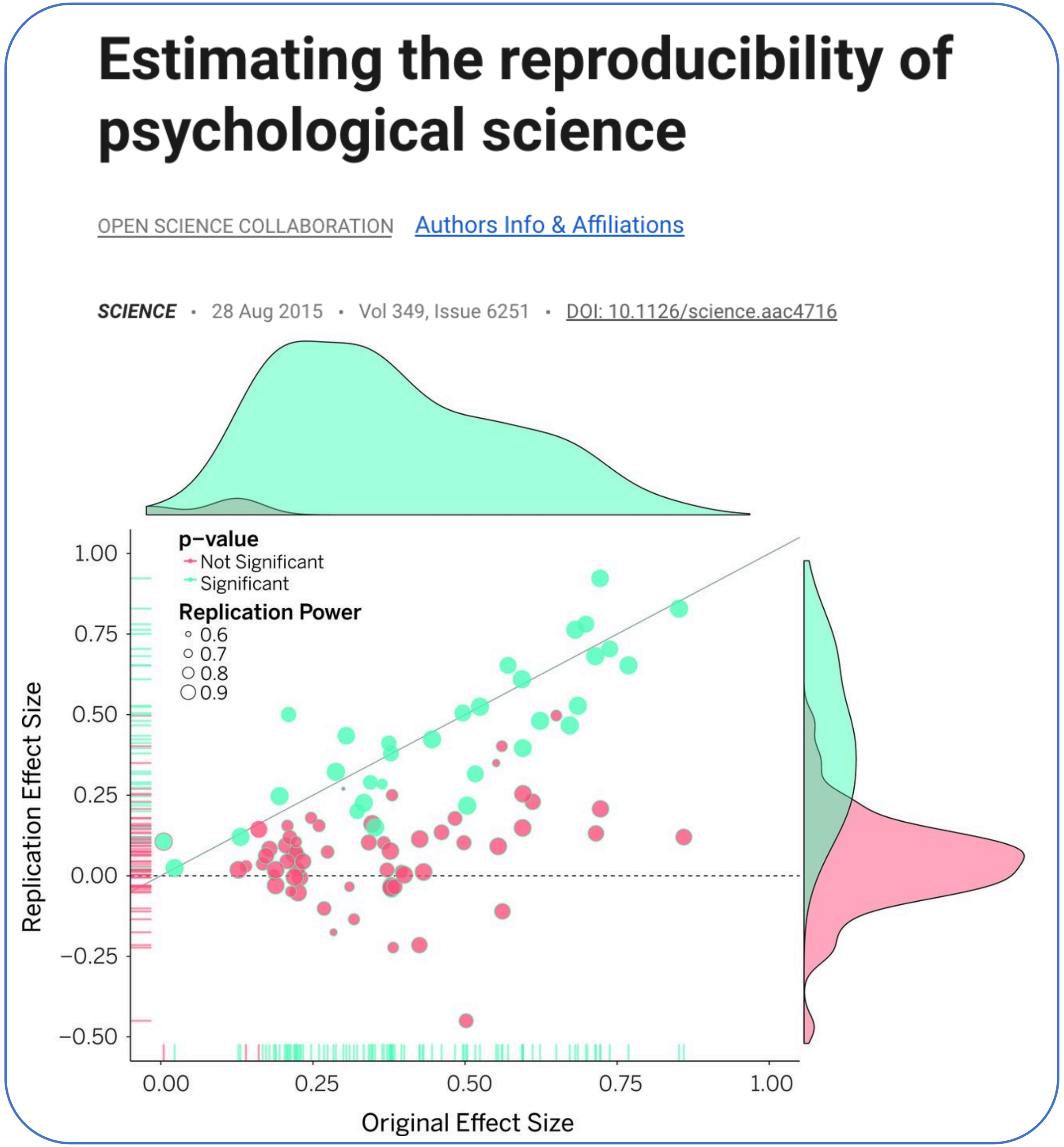
1. What is statistics?

- Why we need statistics?
  - A fundamental tool that allows researchers to analyze data, draw conclusions, and make informed decisions.
  - It provides a **structured** (结构性的) and **objective** (客观的) approach to dealing with uncertainty and variability in data.
  - Reproducibility (可重复性)
    - The meaning of replicated studies

1. What is statistics?

- Reproducibility
  - *"Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project … suggests that there is still more work to do to verify whether we know what we think we know."*
- Why crisis?
  - Invalidated Biological Materials
  - **Lack of Knowledge to Analyze Data**
  - Incorrect Laboratory Practices
  - Undervaluing Negative Results
  - …

Better Stats training is needed!



Estimating the reproducibility of psychological science

OPEN SCIENCE COLLABORATION Authors Info & Affiliations

SCIENCE • 28 Aug 2015 • Vol 349, Issue 6251 • DOI: 10.1126/science.aac4716

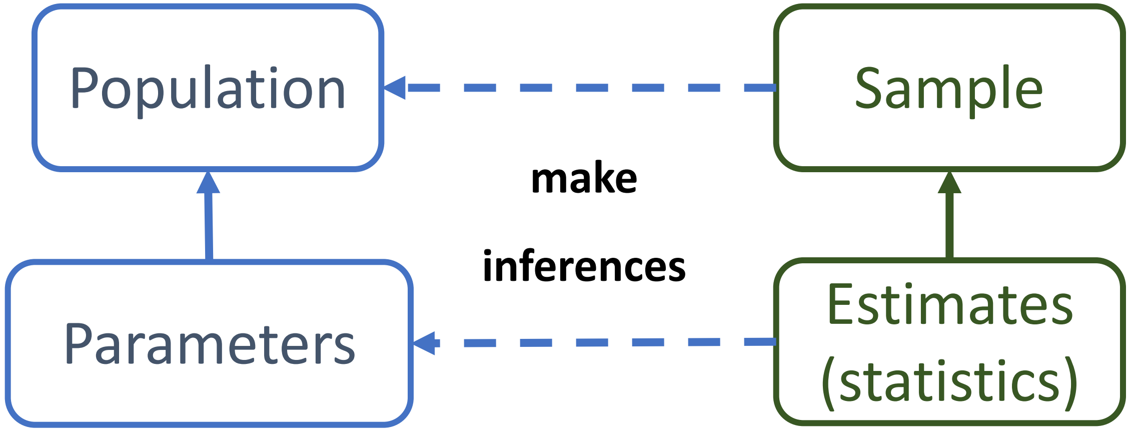
1. What is statistics?

- Why we need statistics in Biology?
  - By using statistics effectively, we can improve our understanding of nature.
  - *Any other reasons?*

Break 5 min

2. Sampling: basic concepts

- Population (总体) *vs* sample (样本)
- Estimation (估计) *vs* hypothesis testing (假设检验)
- Parameter (参数) *vs* estimate/statistic (统计量)
- Probability (概率)
- Sampling distribution (抽样分布)
- Standard error (标准误)
- Confidence interval (置信区间)
- Effect size (效应大小)
- *P*-value (P值)



2. Sampling: basic concepts

- Population (总体)
  - The entire collection of individual units that a researcher is interested in
  - Generally, a large number of individuals
  - E.g., all human beings (globally); all people living in China (regionally); all people in Shanghai; all people at ECNU; all students at ECNU (locally);
- Sample (样本)
  - A subset of individuals selected/observed/measured from the population
  - A much smaller number of individuals
  - E.g., 10-30% of the interested populations

2. Sampling: basic concepts

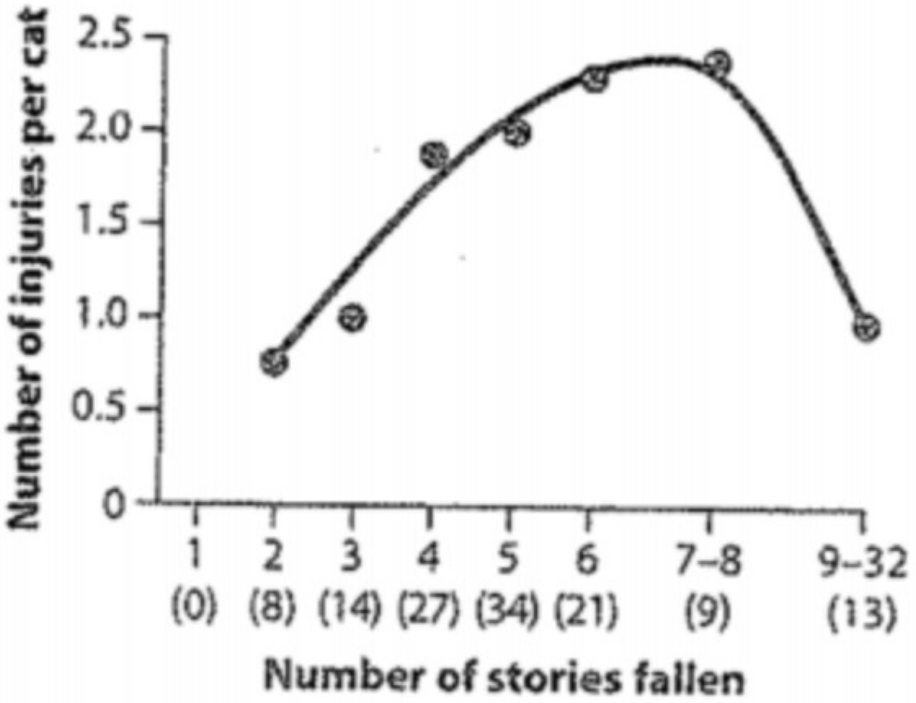
- Population (总体)
  - The entire collection of individual units
  - Parameter (参数)
    - some quantities describing populations (e.g., averages, proportions, measures of variation, measures of relationship…)
- Sample (样本)
  - A subset of individuals
  - Statistic/estimate (统计量)
    - some quantities related to parameters but calculated from a sample
  - Simple random sampling (随机抽样)
    - every individual has the same chance to be sampled (equal probability).

2. Sampling: basic concepts

feline high-rise syndrome
猫高层综合症

- Population *vs* Sample
  - Cat fell from buildings in NYC (vet clinics)
  - The injury rate increases with floors
  - But, it decreases with higher floors!
- Explanations
  - The autors: terminal velocity (6/7th floors) → cat relaxes → this change to its muscles cushions the impact when the cat meet the pavement
  - W & S: samples are biased!
    - fewer samplings at both lower & higher floors →
    - cats from vet clinics ≠ all fallen cats!

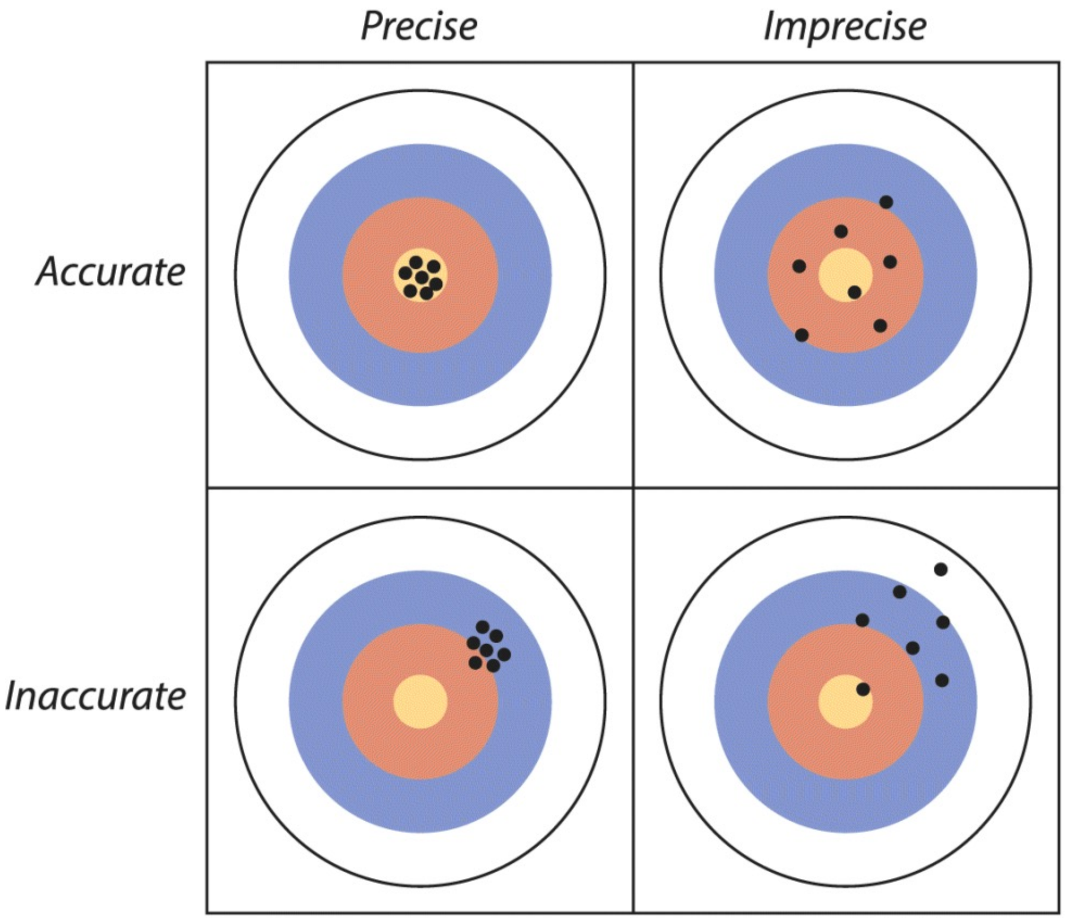
Courtesy of Richard Watherwax/watherwax.com



(Whitlock & Schluter 2020; Data from Diamond 1988)

2. Sampling: basic concepts

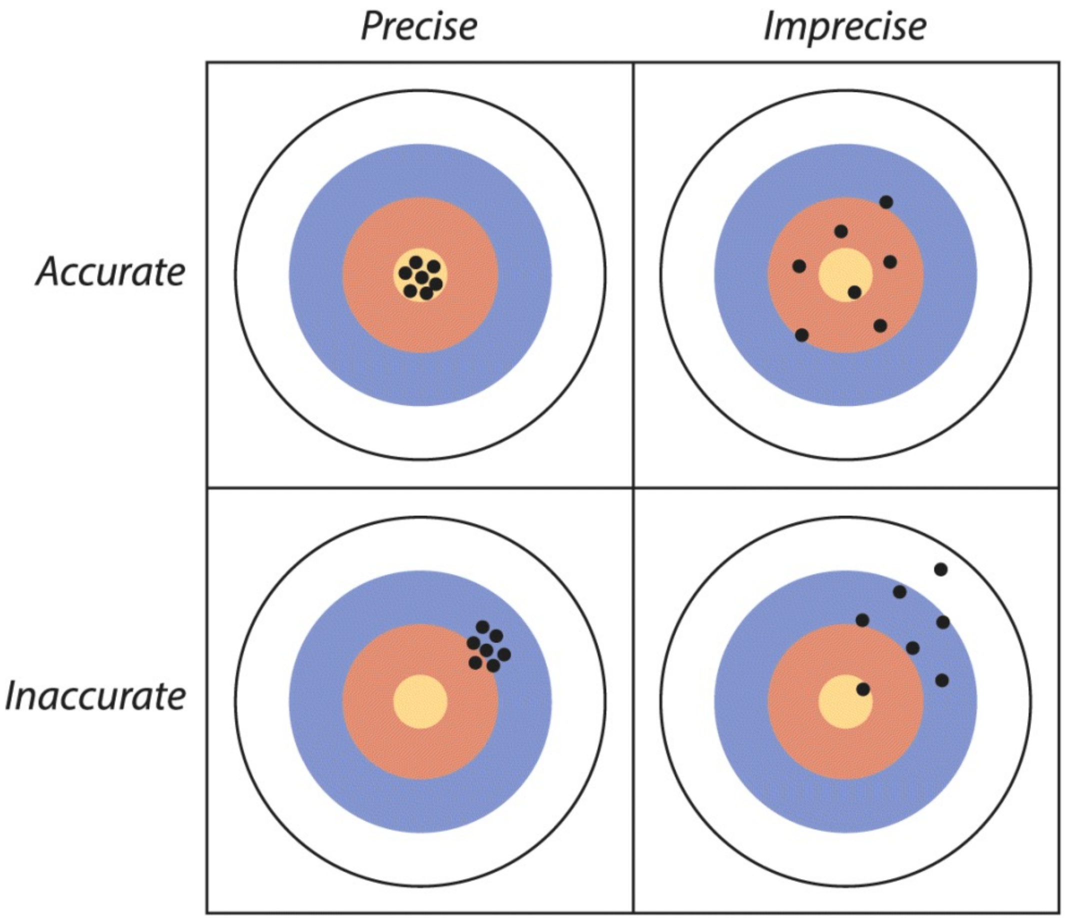
- Properties of good samples
  - An analogy (以打靶为例)
    - Each point: an estimate of the population
    - Multiple points: from repeated samples
  - Sampling error (抽样误差)
    - This **chance difference** (偶然差异) between an estimate and the truth (the population parameter being estimated) **caused by sampling**



(Whitlock & Schluter 2020)

2. Sampling: basic concepts

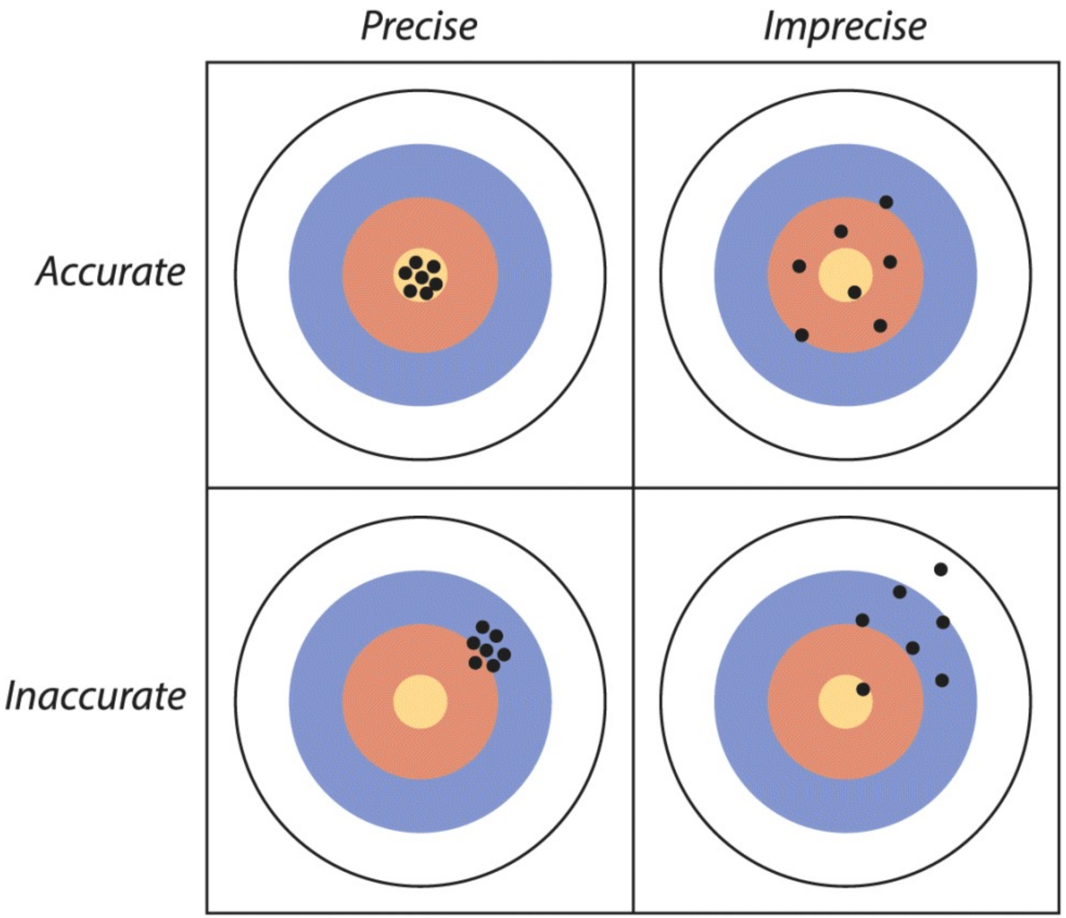
- Properties of good samples
    - Sampling error (抽样误差)
    - Accuracy (准确度)
        - An estimate is accurate (or unbiased), meaning that the average of all estimates that we might obtain is centered on the true population value.
        - Bias (偏差/系统误差) is a systematic discrepancy between the estimates we would obtain and the true population parameter (underestimate/overestimate).



(Whitlock & Schluter 2020)

2. Sampling: basic concepts

- Properties of good samples
  - Sampling error (抽样误差)
  - Accuracy (准确度)
  - Precision (精确度)
    - The spread of estimates resulting from sampling error
    - Larger samples are less affected by chance, and so, all else being equal, larger samples will have lower sampling error and higher precision than smaller samples.



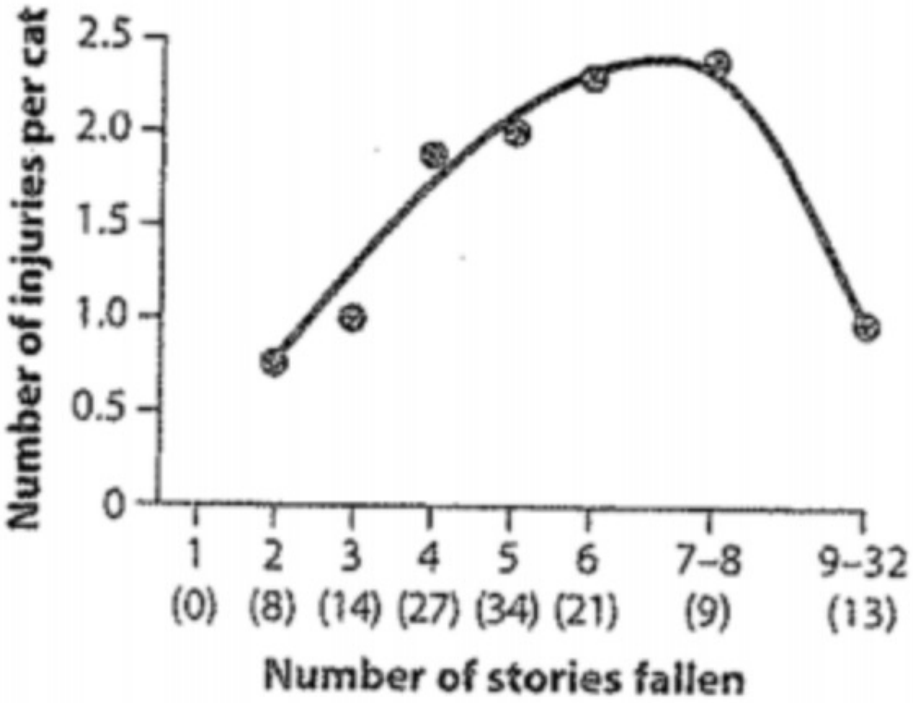
(Whitlock & Schluter 2020)

2. Sampling: basic concepts

feline high-rise syndrome
猫高层综合症

- Population *vs* Sample
  - W & S: samples are biased!
    - cats from vet clinics ≠ all fallen cats!
  - Biases
    - If uninjured and dead cats do not make it to the pet hospital,
    - Injury rates for cats falling only two or three floors are likely to be **overestimated**;
    - Injury rates for cats falling many stories might be **underestimated**.

Courtesy of Richard Watherwax/watherwax.com



(Whitlock & Schluter 2020; Data from Diamond 1988)

2. Sampling: basic concepts

- Other sources of bias in addition to sampling
  - Measurement process
    - e.g., use a stretched tape to measure tree diameter
  - Inadequate estimator
    - mathematical formula, e.g., the case of species diversity like richness and many indices for beta diversity

2. Sampling: basic concepts

- Random sample (随机样本)
  - A common requirement for the most statistical methods
  - Two criteria (This is not as easy as it sounds!)
    - Every unit in the population must have an equal chance of being included in the sample.
      - Some members may be difficult to collect.
    - The selection of units must be independent.
      - With non-independent sampling, our sample size is effectively smaller than we think; and in turn, this will cause miscalculations of estimation precision.

2. Sampling: basic concepts

- Random sample (随机样本)
  - Each member of a population has an equal and independent chance of being selected.
- Random sampling minimizes bias and makes it possible to measure the amount of sampling error.
- Examples?

2. Sampling: basic concepts

- How to take a random sample?
  - Create a list of every unit in the population of interest (*N*), and give each unit a number between one and the total population size.
  - Decide on the number of units to be sampled (call this number *n*).
  - Using a random-number generator, generate *n* random integers between one and the total number of units in the population.
  - Sample the units whose numbers match those produced by the random-number generator.

2. Sampling: basic concepts

- How to take a random sample?
  - Sample *n* student from this course
    - 59 students in total, and select 6 students (to be noted, both numbers are small.)
    - How to do this in R?

2. Sampling: basic concepts

- How to take a random sample?
- Sampling biases
    - The sample of convenience: a sample based on individuals that are easily available to the researcher (e.g., injured cats).
    - Volunteer bias: resulting from a systematic difference between the pool of volunteers and the population to which they belong.
        - more health conscious and more proactive;
        - low-income (if volunteers are paid);
        - more ill, because individuals who are dying anyway might try anything;
        - more likely to have time on their hands;

Break 5 min

3. Types of data and variables

- Data (数据) to collect
  - With a sample in hand, we can begin to measure variables (变量).
  - A variable is any characteristic or measurement that differs from individual to individual.
    - e.g., body height, growth rate, biomass, reproductive rate, etc.
  - Data are the measurements of one or more variables made on a sample of individuals.

3. Types of data and variables

- Data type
    - Categorical variables (分类/类型变量)
        - Qualitative characteristics: describing membership in a category or group
        - Nominal (定性变量): variables have no inherent order;
            - E.g., gender (male vs female), survival (alive or dead), pollinator vectors (insect, bird, wind), language (Mandarin, English, Cantonese, French etc.)
        - Ordinal (定序变量) : variables can be ordered (but with unknown magnitude);
            - E.g., size class (small, medium, or large), life stage (egg, larva, juvenile, subadult, adult);
    - Numerical variables (数值变量)

3. Types of data and variables

- Data type
  - Categorical variables (分类/类型变量)
    - Nominal (定性变量): variables have no inherent order;
    - Ordinal (定序变量) : variables can be ordered (but with unknown magnitude);
  - Numerical variables (数值变量)
    - Quantitative measurements that have magnitude on a numerical scale
    - Continuous (连续变量): can take on any real-number value within some range
      - E.g., body heights, body size, biomass, body temperature, etc.
    - Discrete (离散变量): come in indivisible units
      - E.g., num. of car accident, abundance, species richness , etc.

3. Types of data and variables

- Data type
  - Categorical variables (分类/类型变量)
  - Numerical variables (数值变量)
- Distinguishment & transformation
  - A variable is indexed by a number does not mean it is a numerical variable.
    - E.g., family 1, family 2, …; individual 1, individual 2, etc.
  - Numerical data can be reduced to categorical data by grouping.
    - The result contains less information.
    - E.g., “above average” and “below average”

3. Types of data and variables

- The relationship between variables
  - One major use of statistics is to infer the relationships between two or more variables by examining their associations.
    - $Y \sim X$
  - Goal: to assess how well one of the variables (deemed the explanatory variable) <u>predicts</u> or <u>affects</u> the other variable (called the response variable).
    - Explanatory/independent variables (解释变量/自变量): X
    - Response/dependent variables (响应变量/应变量): Y

3. Types of data and variables

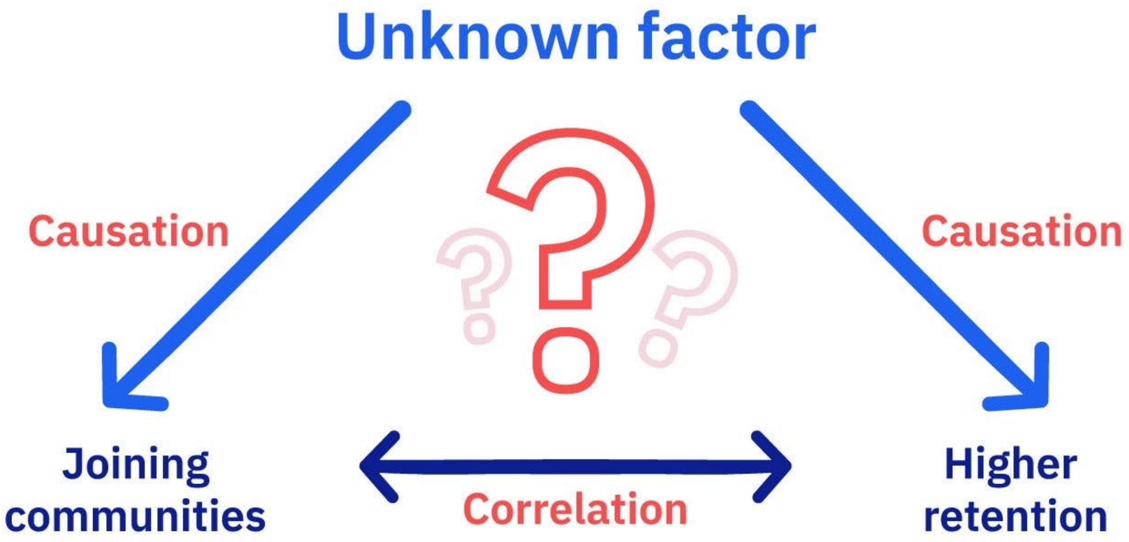
- The relationship between variables
  - One major use of statistics is to infer the relationships between two or more variables by examining their associations.
    - $Y \sim X$
- Examples
  - The administered dose of a toxin in a toxicology experiment would be the explanatory variable, and organism survival would be the response variable.
  - The change of plant/animal biomass with life stages?
  - The species richness with increasing surveyed areas?

4. Types of studies

- Data in biology are obtained from either an experimental or an observational study
  - A study is experimental if the researcher assigns treatments randomly to individuals (e.g., clinical trials, nutrient-addition experiments; control vs. experiment).
  - A study is observational if the assignment of treatments is not made by the researcher (e.g., field surveys for richness, phenology, …).
- Advantage of experiments:
  - randomization minimizes the influence of confounding variables (混淆变量), allowing to determine cause-and-effect relationships between variables.
  - Observational studies can only point to associations.

4. Types of studies

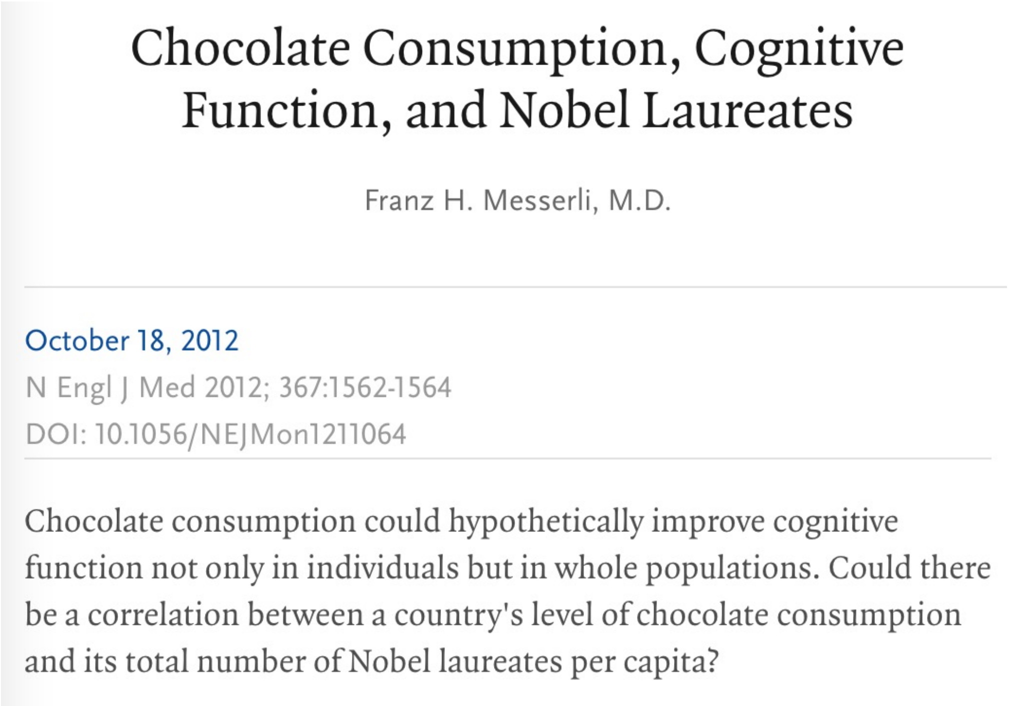
- Correlation (association) ≠ Causation (cause-and-effect)
- 相关性不等于因果关系！



© 2022 Amplitude Inc.

4. Types of studies

- Correlation *vs* Causation

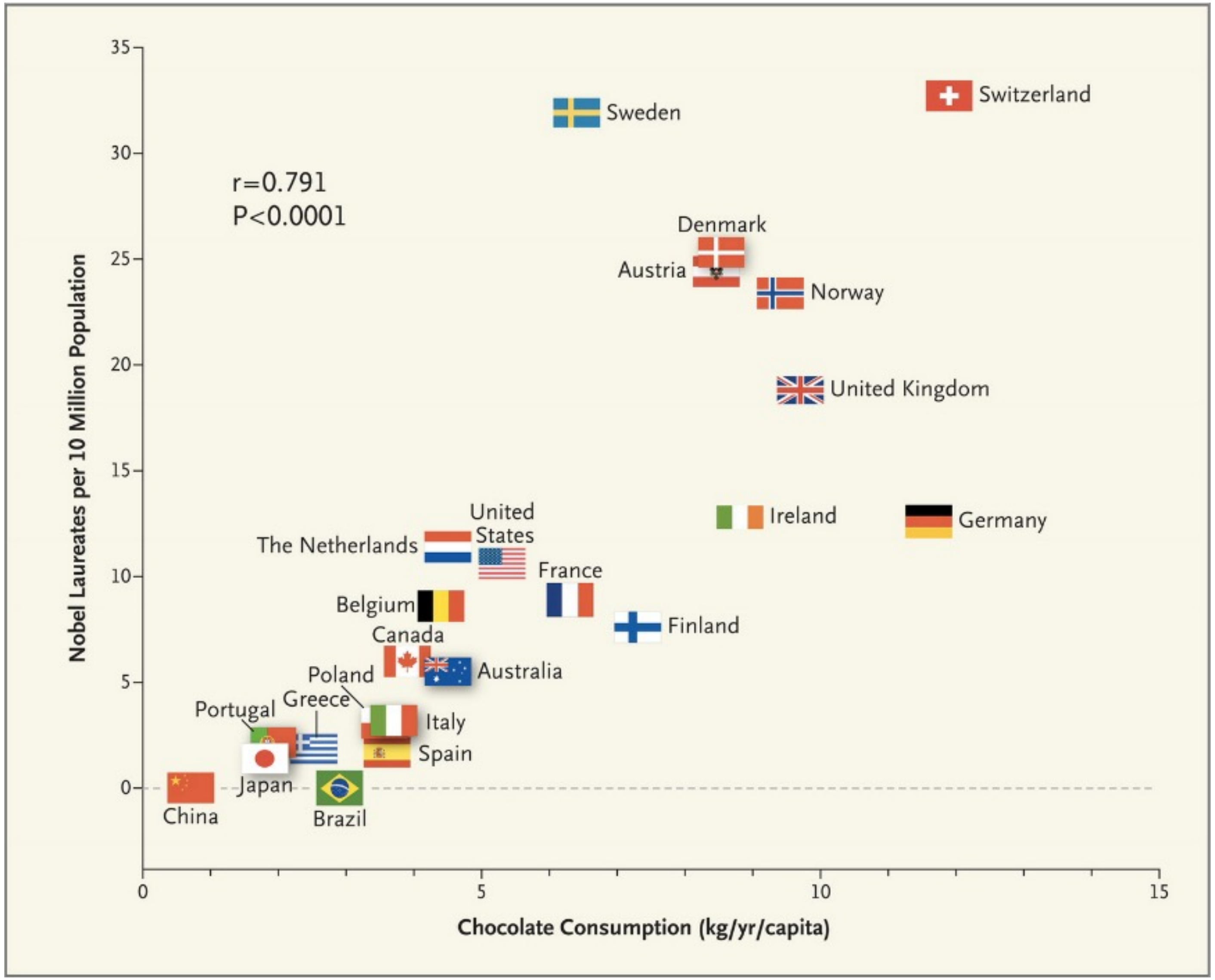


Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

October 18, 2012
N Engl J Med 2012; 367:1562-1564
DOI: 10.1056/NEJMon1211064

Chocolate consumption could hypothetically improve cognitive function not only in individuals but in whole populations. Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?



(with outdated data)

4. Types of studies

- Experimental vs Observational studies
    - Observational studies can suggest the possibility of cause-and-effect relationships between variables.
    - E.g., studies of the health consequences of voluntary cigarette smoking in people are all observational studies, because it is ethically impossible to assign smoking and nonsmoking treatments to people to assess the effects of smoking.
        - The health hazards of smoking in nonhuman animals (e.g., mouse) have helped make the case that cigarette smoking is dangerous to human health.

5. Summary

- Statistics is the study of methods for measuring aspects of populations from samples and for quantifying the uncertainty of the measurements.
- Much of statistics is about estimation, and allows hypothesis testing.
- The goals of sampling are to increase the accuracy and precision of estimates and to ensure that it is possible to quantify precision.
- In a random sample, every individual in a population has the same chance of being selected, and the selection of individuals is independent.
- Variables are either categorical or numerical, measured from experimental or observational studies.
- In studies of association between two variables, the explanatory variable is typically used to predict the response variable (Correlation ≠ Causation).

6. Discussions

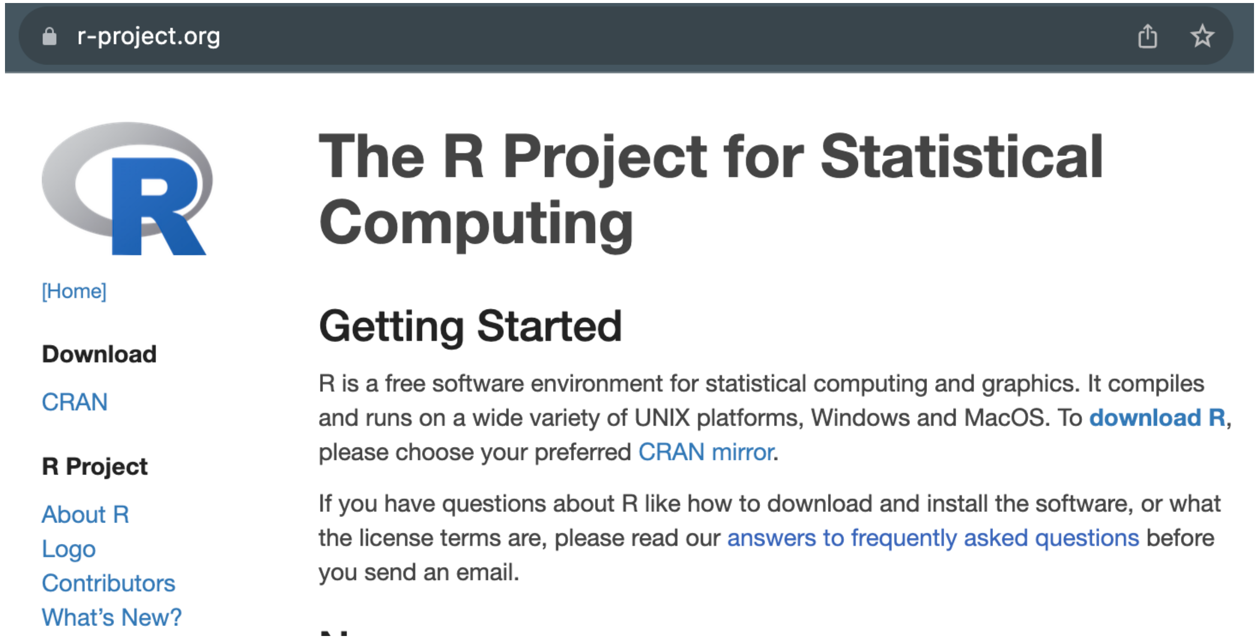
- 1. Which of the following numerical variables are continuous? Which are discrete?
  - a. Number of injuries sustained in a fall
  - b. Fraction of birds in a large sample infected with avian flu virus
  - c. Number of crimes committed by a randomly sampled individual
  - d. Logarithm of body mass

6. Discussions

- 2. The average age of piñon/pinyon pine trees in the coast ranges of California was investigated by placing 500 10-hectare plots randomly on a distribution map of the species using a computer. Researchers then found the location of each random plot in the field, and they measured the age of every piñon pine tree within each of the 10-hectare plots. The average age within the plot was used as the unit measurement. These unit measurements were then used to estimate the average age of California piñon pines.
  - What is the population of interest in this study?
  - Why did the researchers take an average of the ages of trees within each plot as their unit measurement, rather than combine into a single sample the ages of all the trees from all the plots?

About R

- https://www.r-project.org/
  - An Introduction to R, 2023 (v4.3.1) https://cran.r-project.org/doc/manuals/R-intro.pdf



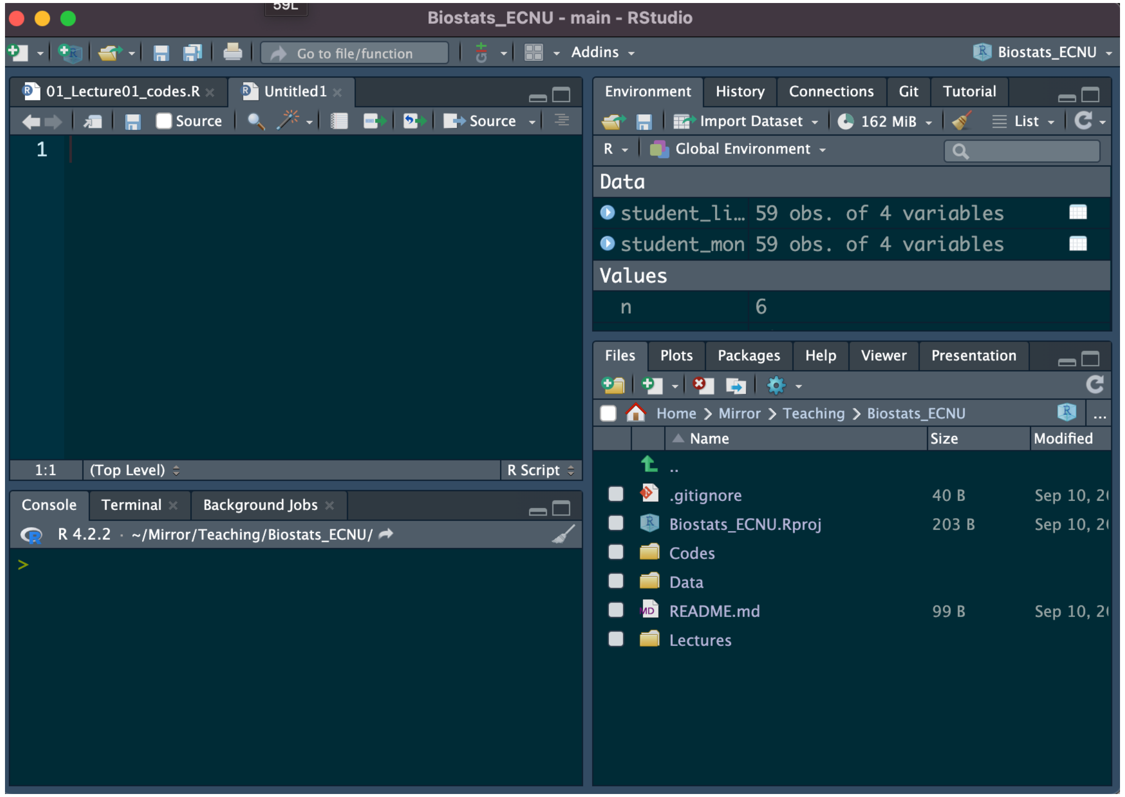
The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Rstudio



Data

```
student_li… 59 obs. of 4 variables
student_mon 59 obs. of 4 variables
```

Values

```
n           6
```

Name	Size	Modified
..		
.gitignore	40 B	Sep 10, 2(
Biostats_ECNU.Rproj	203 B	Sep 10, 2(
Codes		
Data		
README.md	99 B	Sep 10, 2(
Lectures		

About R

- Advantages
  - Powerful, flexible, and free!
  - Runs on all computer platforms (I used Mac, but similar with Windows).
  - New stuff always coming online – yet all in a common language
    - Various packages for data cleaning, stats models, plot figures.
  - Superb data management and manipulation capabilities.
- Disadvantages
  - R uses scripts to execute commands rather than menus and a mouse.
  - It can sometimes be difficult to do otherwise simple things.
  - There are several kinds of data objects to remember.
  - Some variation in command syntax, e.g., plot() vs ggplot().

About R – bad things

- R uses scripts to execute commands rather than menus and a mouse.
- It can sometimes be difficult to do otherwise simple things.
- It is not a great spreadsheet.
    - Use a dedicated spreadsheet program for text files (e.g., .csv).
- 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).
- Some variation in command syntax, e.g., plot() vs ggplot().
- 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.