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

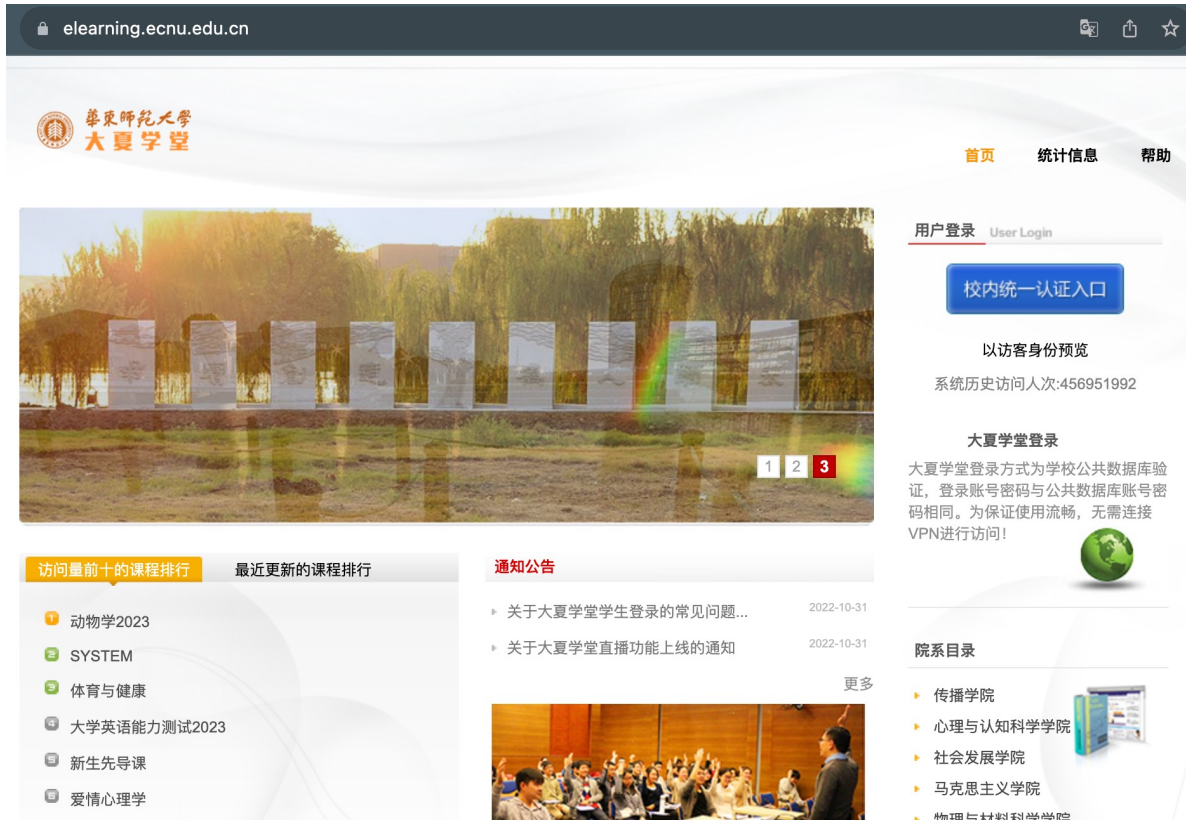
课程结构及考核
course components & grading

- 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/



1. 如何进入大夏学堂

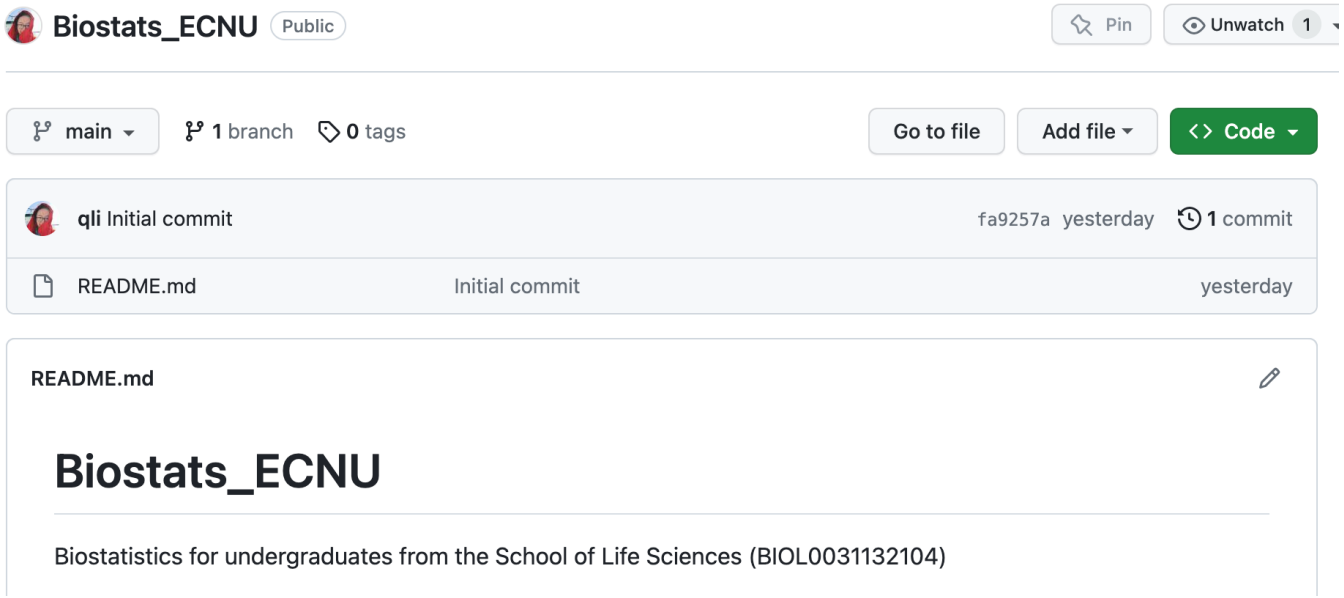
学生均可通过四种方式进入大夏学堂：①学校首页->“教师教育”菜单->“大夏学堂”②学校首页->“快速通道”->“大夏学堂”③教务处首页->快速链接“大夏学堂（数字化教学平台）”④浏览器直接输入网址“https://elearning.ecnu.edu.cn/”。点击“校内统一认证入口”，使用自己的学校公共数据库账号和密码可登陆本人的课程空间。

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

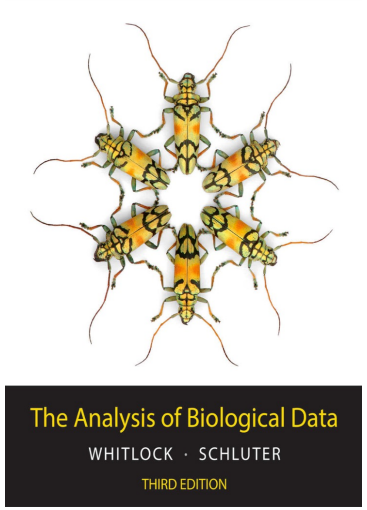


课程大纲 Syllabus (tentative)

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

- 理解统计学的基本概念和方法，以及它们在生物学中的应用;
- 建立假设检验、设计可靠研究、收集和组织数据以及进行正确数据分析的基本原则;
- 侧重于数据，而不是统计学的数学基础;
- 使用计算工具R培养分析技能!
  - 学习曲线较陡峭（但多练习/通过实践学习）

授课教师 About the instructor

- 学术背景：生态学、进化学和生物地理学;
- 研究主题：侧重于生物多样性（生态位/分布范围/性状）;
- 我不是统计学专家，也不是R专家；可能无法回答所有统计学问题;
- 但我通过经验学到了一些方法的应用，知道如何找到解决方案;
- 使用R已有10多年，主要用于统计分析和制作图表;
- 答疑时间：周三下午1-3点，地址：资环楼 #329。

学生 Students

- Who are you?
  - 专业: 生物系;
  - 已学习高数B;
- 对生物统计学课程的期待?

Lecture 1 – Introduction to Biostatistics

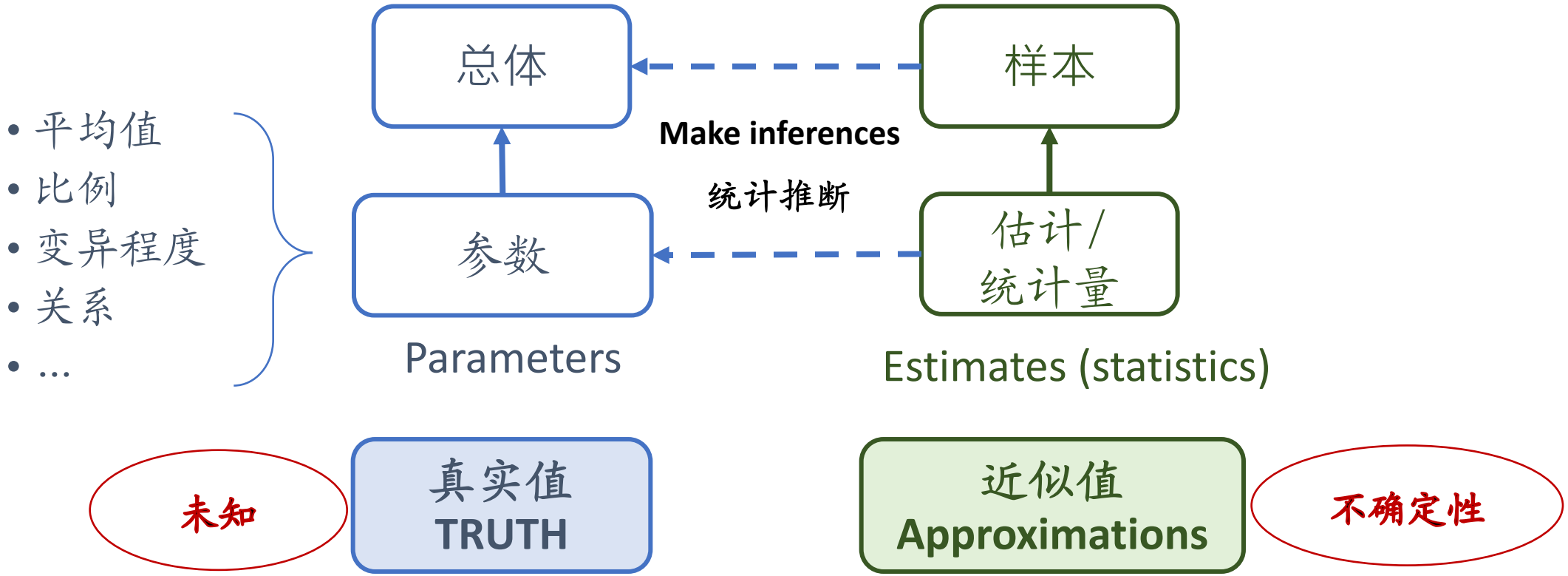
- 内容大纲 Outline
  - 什么是统计学?
  - 抽样: 基本概念
  - 数据和变量的类型
  - 研究的类型
  - 总结
  - 讨论

1. 统计学的基本概念

- 统计学 (**Statistics**) 是研究从样本 (samples) 中描述和测量自然现象的方法;
- 统计学涉及估计 (estimation) 的过程——即使用样本数据 (sample data) 推断目标总体 (a target population) 的未知量 (an unknown quantity);
- 统计学还能量化这些测量估计的不确定性——即它们与真实值的偏差;

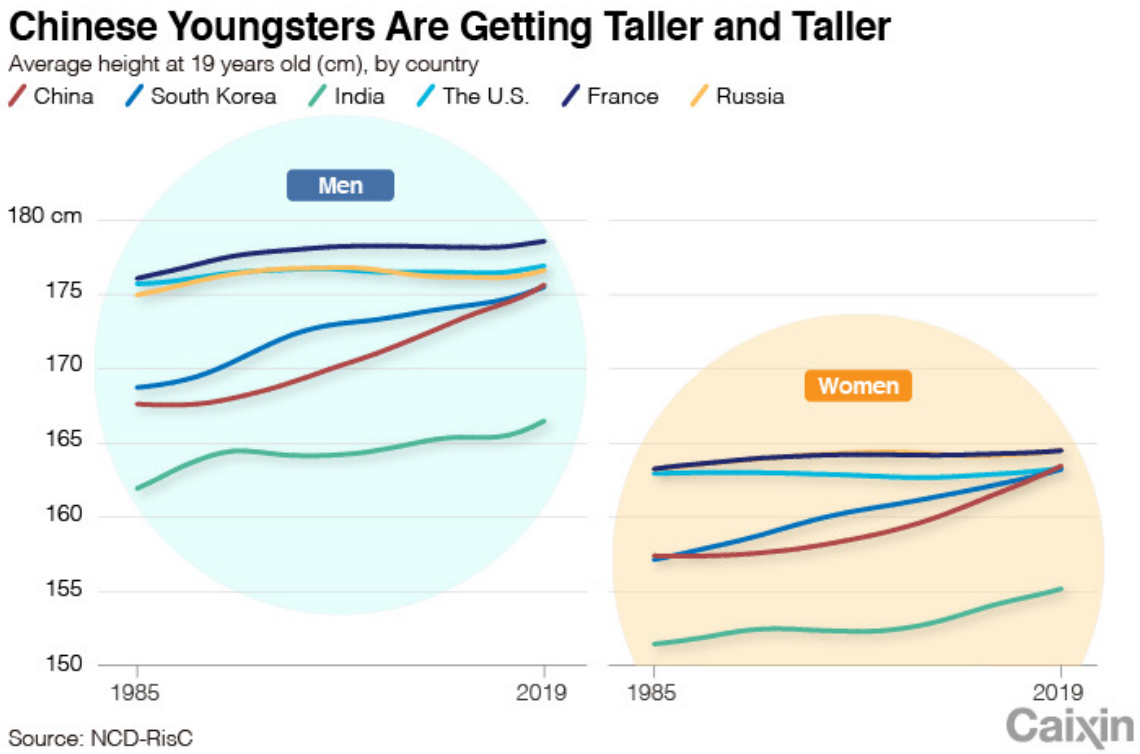
1. 统计学的基本概念

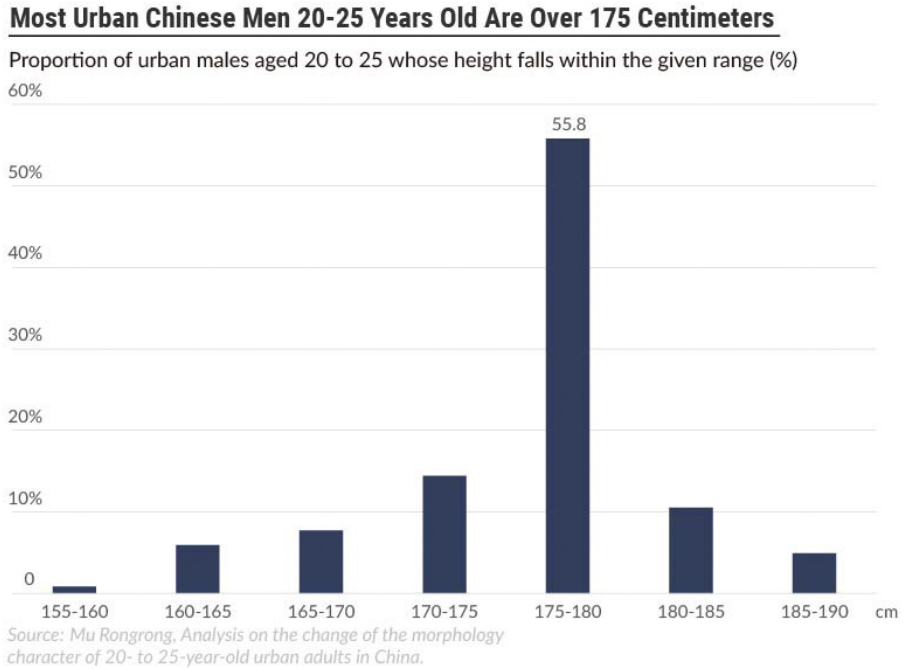
- About estimation of population (总体), with sample data (样本).

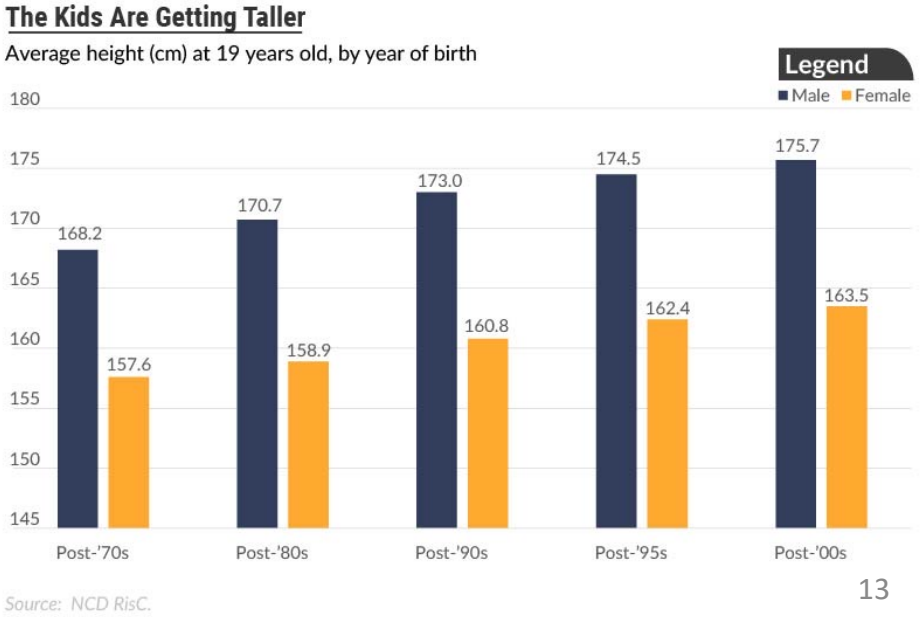


1. 统计学的基本概念

- 例子: 身高
  - 均值 vs 分布







1. 统计学的基本概念

- 例子：统计学涉及使用样本数据进行估计。
  - 身高 Body heights

- 你能提出一些现实中的生物学问题吗？
  - 如何解决？

1. 统计学的基本概念

- 统计学还涉及假设检验 (hypothesis testing);
- 统计假设是关于总体参数 (a population parameter) 的具体声明;
  - 例子：身高
  - 假设：男性的平均身高高于女性;
  - 如何进行检验？

1. 统计学的基本概念

- 为什么我们需要统计学？
  - 这是一种分析数据、得出结论并做出明智的决策的基本工具;
  - 它提供了一种有结构性的(structured)和客观的(objective)方法来处理数据中的不确定性和变异性;
  - 可重复性 (Reproducibility)
    - 重复研究的含义

1. 统计学的基本概念

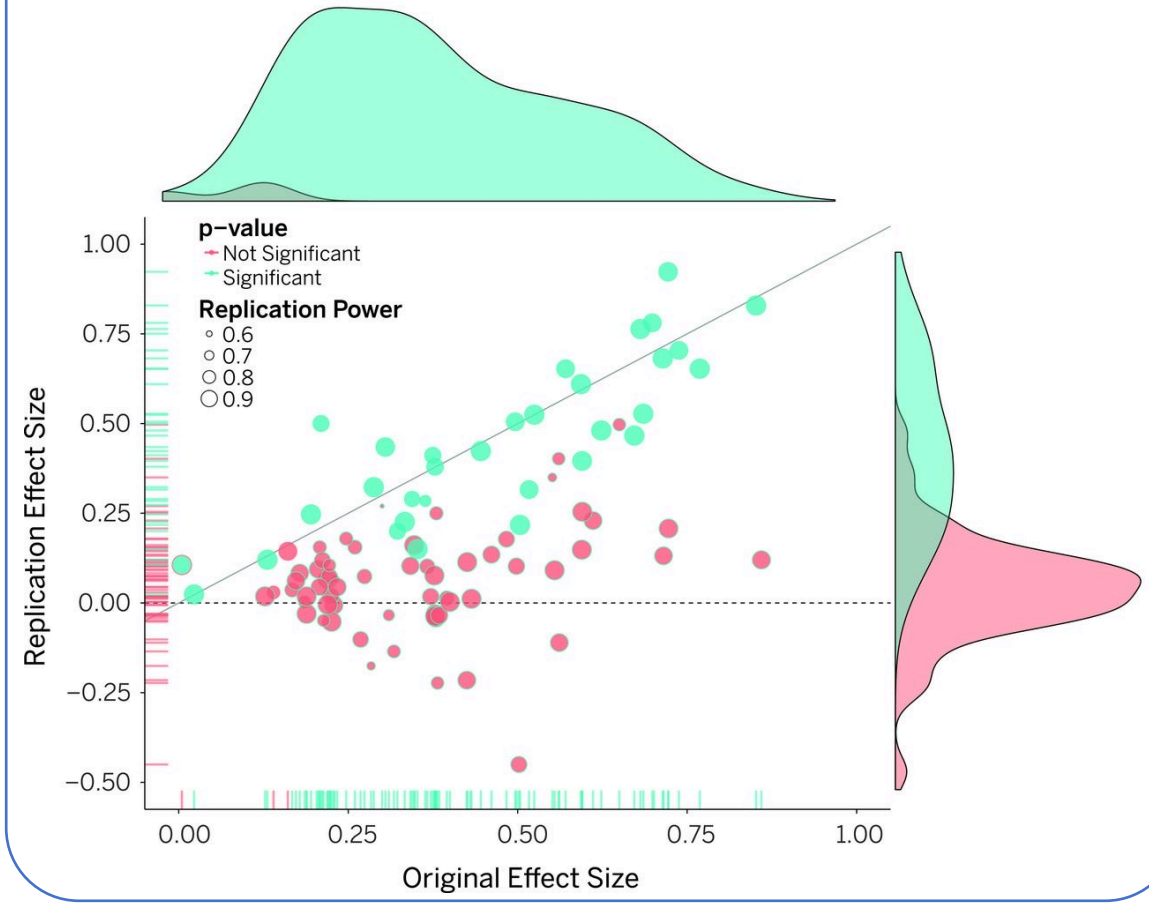
- 可重复性 Reproducibility
  - *“Replication can increase **certainty** when findings are reproduced, and promote **innovation** when they are not. This project … suggests that there is still more work to do to verify whether we know what we think we know.”*
- 可重复性的危机?
  - 失效的生物材料
  - **缺乏分析数据的知识**
  - 不正确的实验室操作
  - 低估负面的结果
  - …

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. 统计学的基本概念

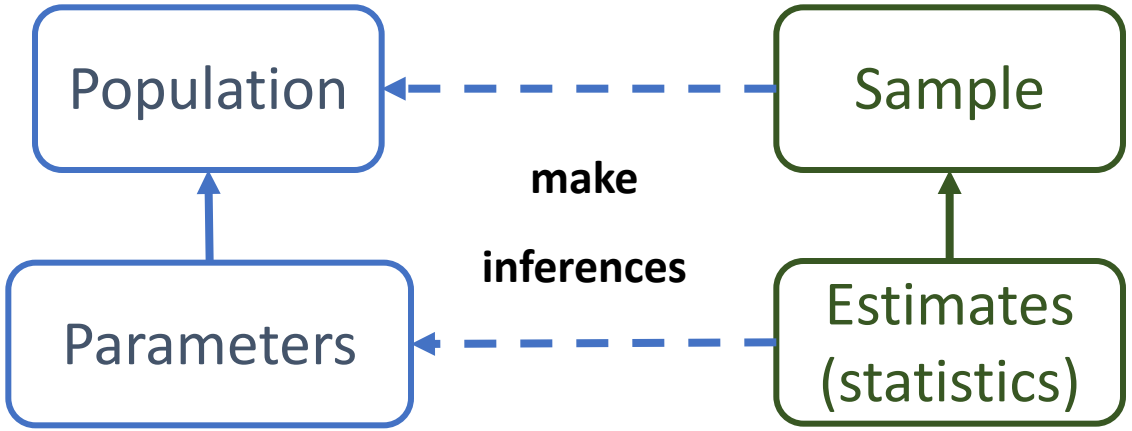
- 在生物学中为什么需要统计学？
  - 通过有效地统计学应用，我们可以提高对自然及其规律的理解。

- 还有其他原因吗？

2. 抽样的基本概念

- 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)

vs： *versus*



2. 抽样的基本概念

- 总体 Population
  - 研究中感兴趣的所有个体的整体集合 (the entire set);
  - 通常是大量的个体
  - 例如，全球人类、所有居住在中国的人、上海的所有人、华东师范大学的所有人、华东师范大学的所有学生（空间尺度的差异）;
- 样本 Sample
  - 从总体中选择/观察/测量的个体的子集 (a subset);
  - 个体数量要小得多;
  - 例如，感兴趣的总体中10-30%的个体;

2. 抽样的基本概念

- 总体 Population
  - 研究中感兴趣的所有个体的整体集合 (the entire set);
  - 参数 (parameter)
    - 描述总体特征的一些量 (例如，平均值、比例、变异度、相关等)
- 样本 Sample
  - 从总体中选择/观察/测量的个体的子集 (a subset);
  - 统计量 (statistic/estimate)
    - 从样本计算得出的与参数相关的一些量;
  - 随机抽样 (simple random sampling)
    - 每个个体被抽中的概率相等 (equal probability/chance);

2. 抽样的基本概念

feline high-rise syndrome
猫高层综合症

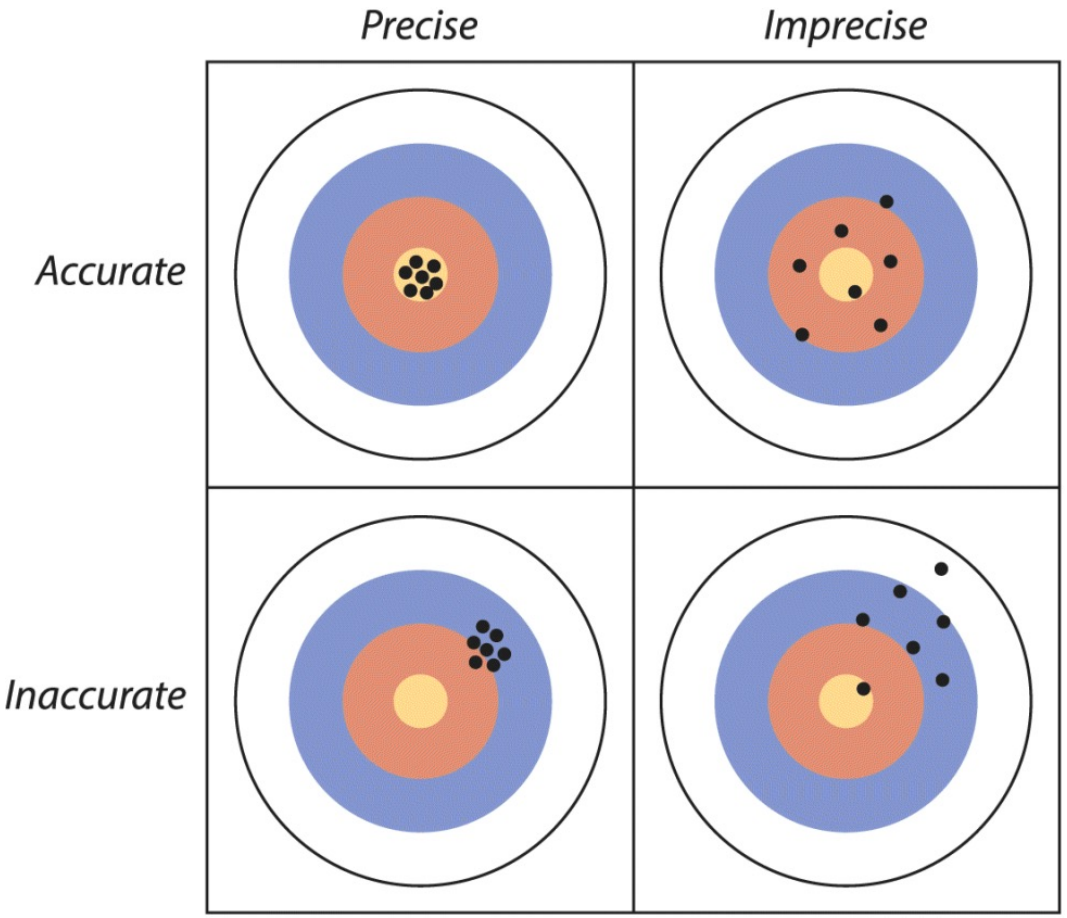
Courtesy of Richard Watherwax/watherwax.com

- 总体 *vs* 样本
  - 纽约市楼房上掉落的猫（兽医诊所的数据）
  - 伤害程度 (injury rate) 随楼层增加而增加
  - 但是，更高楼层的伤害程度反而降低了！
- 可能的解释
  - 论文作者: 达到终端速度 (terminal velocity at 6/7th floors) → 猫会放松 → 肌肉的这种变化缓冲了到达地面的冲击！
  - W & S（参考教材作者）:
    - 样本存在偏差 (samples are biased)!
    - 较低楼层和较高楼层的样本较少 →
    - 来自兽医诊所的猫 ≠ 所有掉落的猫！

(Whitlock & Schluter 2020; Data from Diamond 1988)

2. 抽样的基本概念

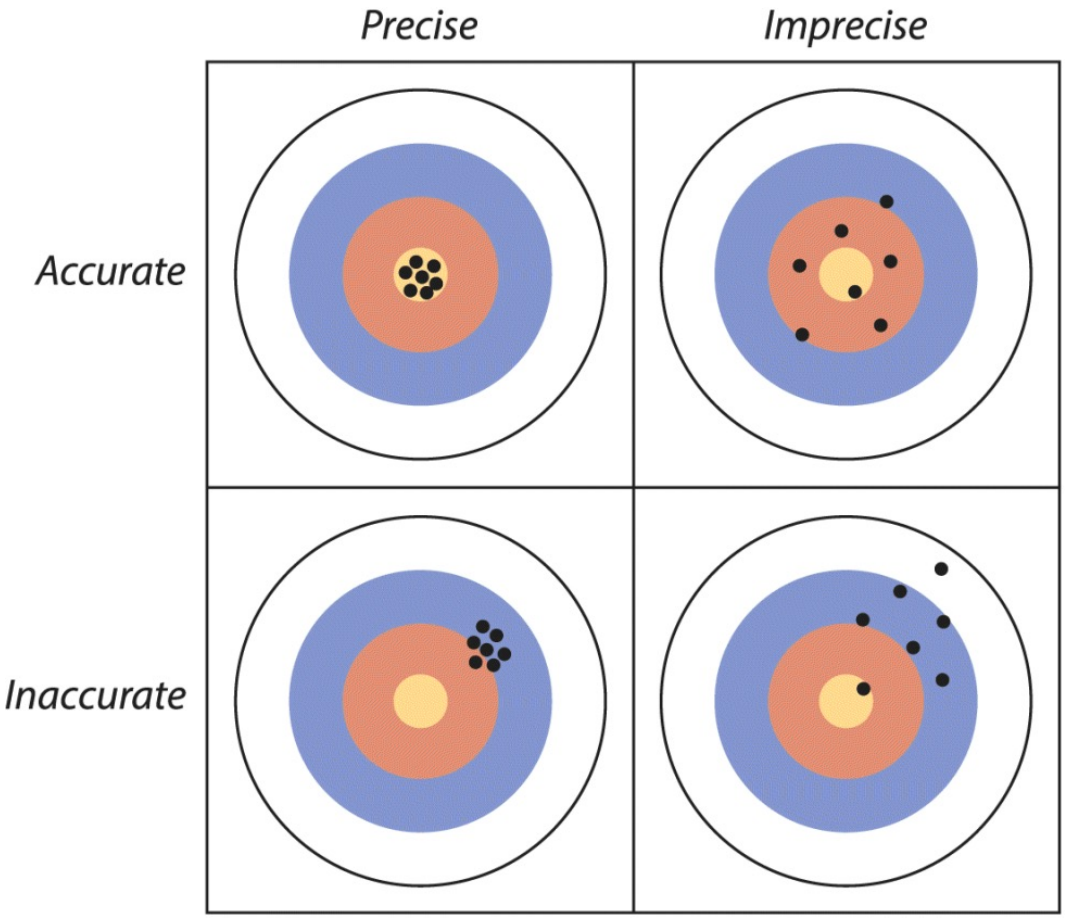
- 好样本的特征 (good samples)
  - 以打靶为例
    - 每个点：总体参数估计的一个估计值
    - 多个点：来自重复的样本
      - Repeated samples
  - 抽样误差 (sampling error)
    - 由抽样引起的估计值与真实值（正在估计的总体参数）之间的偶然差异 (chance difference)



(Whitlock & Schluter 2020)

2. 抽样的基本概念

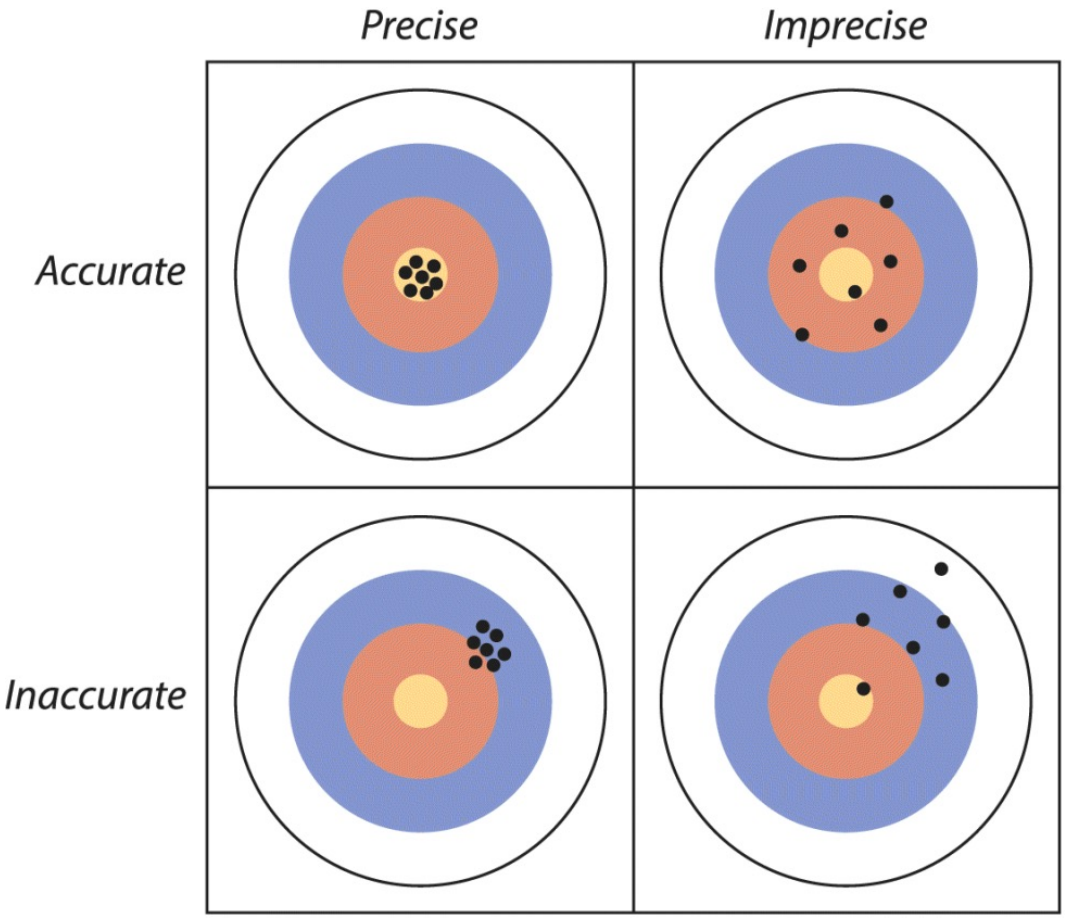
- 好样本的特征 (good samples)
  - 抽样误差 (sampling error)
  - 准确度 (Accuracy)
    - 估计是准确的或无偏的(accurate/ unbiased)，意味着我们可能获得的所有估计值的均值都集中在真实的总体参数上（靶心）。
    - 偏差 (bias) 是我们获得的估计值与真实的总体参数之间的系统性差异（低估或高估）。



(Whitlock & Schluter 2020)

2. 抽样的基本概念

- 好样本的特征 (good samples)
  - 抽样误差 (sampling error)
  - 准确度 (Accuracy)
  - 精确度 (Precision)
    - 由于抽样误差引起的估计值的分散程度;
    - 大样本受偶然差异的影响较小，因此其它条件相等的情况下，较大的样本 (larger samples) 将具有较低的抽样误差和较高的精确度。



(Whitlock & Schluter 2020)

2. 抽样的基本概念

- 总体 *vs* 样本
  - W & S: 样本存在偏差 (samples are biased)!
    - 来自兽医诊所的猫 ≠ 所有掉落的猫!
  - 偏差 (biases)
    - 如果未受伤和死亡的猫无法到达兽医诊所,
    - 那么从两三层楼掉落的猫的受伤程度可能被高估 (**overestimated**);
    - 而对于从高层楼摔下的猫的受伤程度可能被低估 (**underestimated**);

feline high-rise syndrome
猫高层综合症

Courtesy of Richard Watherwax/watherwax.com

(Whitlock & Schluter 2020; Data from Diamond 1988)

2. 抽样的基本概念

- 除了抽样之外的其它偏差来源
  - 测量过程
    - 例如，使用拉伸的胶带测量树的直径;
  - 估计值的不足
    - 数学公式的缺陷，例如物种多样性如丰富度和许多多样性指数的情况;

2. 抽样的基本概念

- 随机样本 Random sample
  - 大多数统计方法的常规要求/前提假设
  - 两个标准 (two criteria)
    - 总体中的每个个体都必须有相等的概率被抽样 (equal chance);
      - 难点：一些个体可能难以被抽中;
    - 被选中的个体间必须是相互独立的 (independent) ;
      - 对于非独立抽样，样本大小实际上比我们认为的要小；反过来，这将导致精确度的估计不准;

2. 抽样的基本概念

- 随机样本 Random sample
  - 总体中的每个个体都有相等且独立的机会被抽样;

- 随机抽样使得偏差最小化
  - 并让我们可以估计抽样误差的大小;
  - 例子?

2. 抽样的基本概念

- 如何进行随机抽样 (random sampling)?
  - 创建总体的个体列表（N），为每个个体分配一个数字，介于1和总体总数之间;
  - 确定要抽样的个体数（称为n）——样本大小;
  - 使用随机数生成器，生成介于1和总体总数之间的n个随机整数;
  - 抽取编号与随机数生成器生成的编号相匹配的个体;

2. 抽样的基本概念

- 如何进行随机抽样 (random sampling)?
  - 从这门课中抽取n名学生:
  - 总共有59名学生，选择6名学生（请注意，这两个数字其实都很小）;
  - 如何在R中实现?
    - sample(x=6, n=59)

2. 抽样的基本概念

- 如何进行随机抽样 (random sampling)?
- 抽样误差 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;

2. 抽样的基本概念

- 如何进行随机抽样 (random sampling)?
- 抽样偏差 sampling biases
  - 便利样本（the sample of convenience ）： 基于研究人员容易获得的个体的样本（例如，受伤的猫）;
  - 志愿者偏差（volunteer bias）： 由志愿者和他们所属的总体之间的系统差异引起;
    - 更注重健康和更主动;
    - 更低收入（如果志愿者是有偿的）;
    - 生病更严重，因为面临死亡的个体可能会尝试任何事情;
    - 更可能有空闲时间;

3. 数据和变量的类型

- 收集数据（collecting data）
  - 针对随机抽样得来的样本，我们可以开始测量变量（variables）；
  - 变量是不同个体的特征或测量；
    - 例如，身高、生长速率、生物量、繁殖率等；
  - 数据是对样本个体进行的一个或多个变量的测量；

3. 数据和变量的类型

- 数据类型
  - 分类/类型变量（Categorical variables）
    - 定性特征：描述属于某一类别或组的特征;
    - （1）定性变量（Nominal）：变量没有固有的顺序;
      - 例如，性别（男性或女性）、存活状态（活着或死亡）、传粉媒介（昆虫、鸟类、风），语言（普通话、英语、广东话、法语等）;
    - （2）定序变量（Ordinal）：变量可以被排序（但未知绝对大小）;
      - 例如，尺寸等级（小、中、大）、生命周期阶段（卵、幼虫、幼体、亚成体、成体）;
  - 数值变量（Numerical variables）

3. 数据和变量的类型

- 数据类型
  - 分类/类型变量（Categorical variables）
    - （1）定性变量（Nominal）：变量没有固有的顺序;
    - （2）定序变量（Ordinal）：变量可以被排序（但未知绝对大小）;
  - 数值变量（Numerical variables）
    - 测量值具有数值范围（a numerical scale）的变量;
    - （1）连续变量（continuous）：某个范围内取任何实数值（real-number）;
      - 例如，身高、体积、生物量、体温等;
    - （2）离散变量（discrete）：以不可分割的单位出现（indivisible units）;
      - 例如，车祸数量、物种多度、物种丰富度等;

3. 数据和变量的类型

- 数据类型
  - 分类/类型变量（Categorical variables）
  - 数值变量（Numerical variables）
- 区分和转换（distinguishment & transformation）
  - 被编号的变量并不意味着它是数值变量;
    - 例如，家庭1、家庭2、或个体1、个体2等;
  - 数值数据可以通过分组转换为分类数据;
    - 转换后包含较少的信息（丢失绝对数值的信息）;
    - 例如，“高于平均值”和“低于平均值”;

3. 数据和变量的类型

- 变量之间的关系（the relationship between variables）
  - 统计学的一个主要用途是通过检验变量之间的相关性来推断它们之间的关系;
    - $Y \sim X$
  - 目标：评估一个解释变量对响应变量的预测或影响效果如何;
    - 解释变量/自变量（explanatory/independent variables）：X
    - 响应变量/因变量（response/dependent variables）：Y

3. 数据和变量的类型

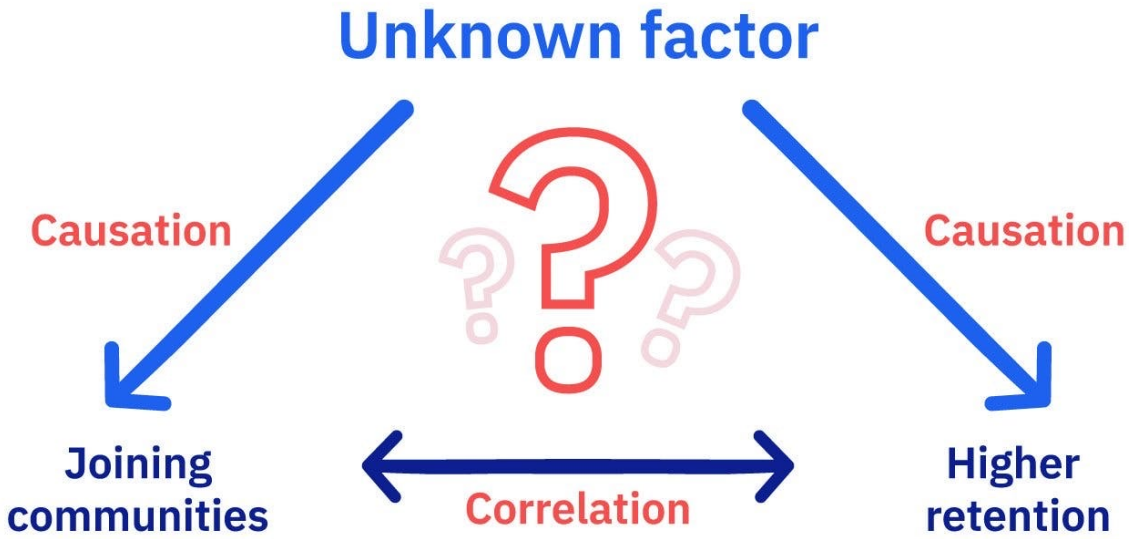
- 变量之间的关系（the relationship between variables）
  - 统计学的一个主要用途是通过检验变量之间的相关性来推断它们之间的关系；
    - $Y \sim X$
- 例如：
  - 毒理实验中毒素的剂量是解释变量，而生物的存活是响应变量；
  - 哪个变量是解释变量，哪个是相应变量？
    - 生命阶段中植物/动物生物量的变化？
    - 随着调查样方的增大，物种丰富度的变化？

4. 研究的类型

- 生物学中的数据通常来自两类研究:
  - 试验研究（ an experimental study）:
    - 研究人员将不同处理随机分配给个体（例如，临床试验，营养添加实验；对照组与试验组）；
  - 观察研究 （an observational study）:
    - 如果处理的分配不是由研究人员进行的，则该研究是观察性的（例如，对丰富度、物候等进行的现场调查）；
- 试验研究的优势（advantage）
  - 随机分组最小化了混淆变量（confounding variables）的影响，从而能探讨变量之间的因果关系（cause-and-effect）；
  - 而观察研究只能指出变量之间的相关性（associations）；

4. 研究的类型

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



4. 研究的类型

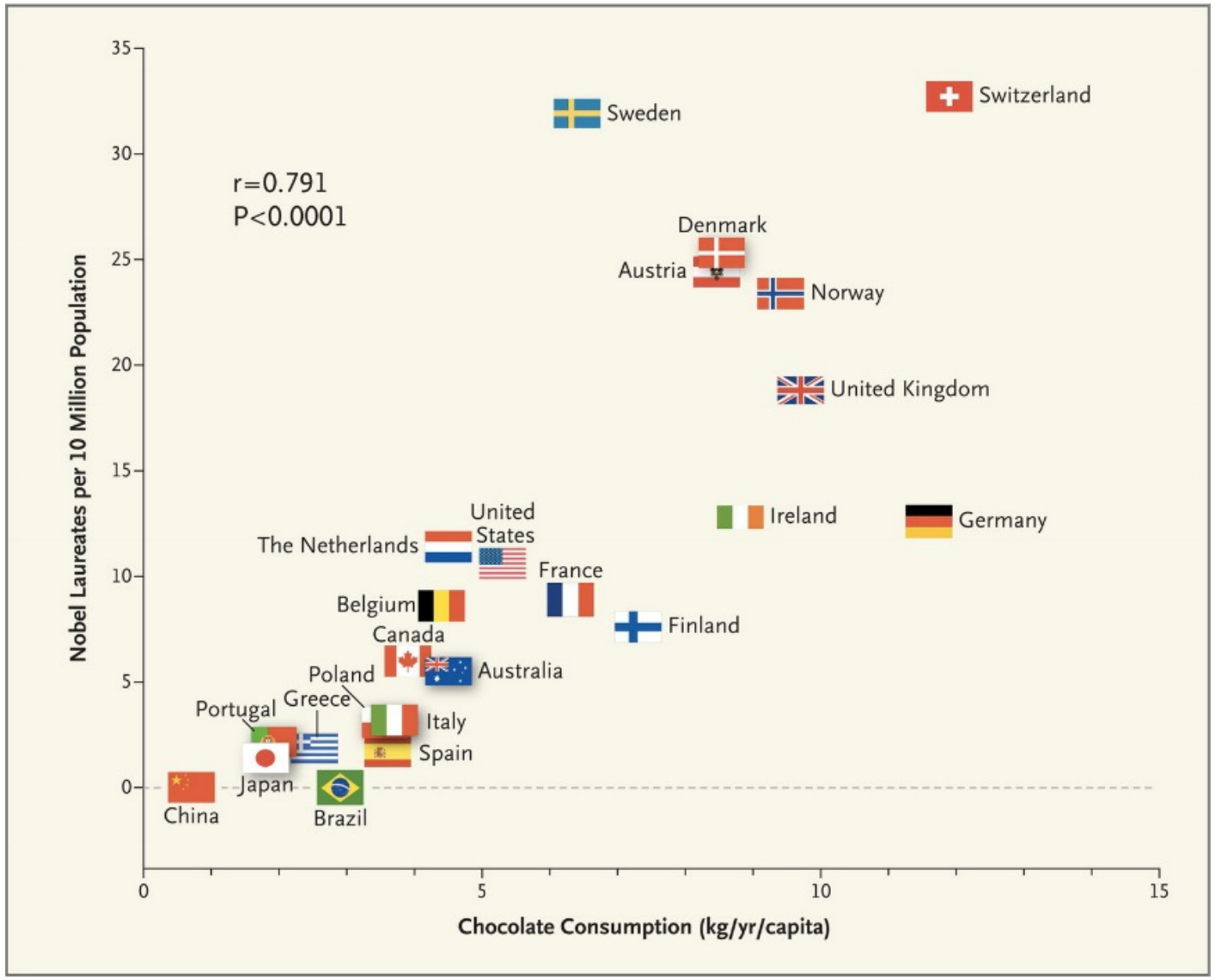
- Correlation *vs* Causation

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. 研究的类型

- 试验研究与观察研究
  - 观察研究可以揭示变量之间可能的因果关系;
  - 例如，关于人类自愿吸烟的健康后果的研究都是观察性研究;
    - 因为在伦理上不可能为人们分配吸烟和不吸烟的处理以评估吸烟的影响;
    - 非人类动物（例如，小鼠）吸烟的健康危害研究有助于证明吸烟对人类健康是危险的。

5. 小结 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).

5. 小结 Summary

- 统计学是研究从样本中测量总体特征和量化测量不确定性的方法;
- 统计学的很大一部分涉及参数估计，并进行假设检验;
- 抽样的目标是提高估计的准确度和精确度，并确保能够量化精确度;
- 在随机样本中，总体中的每个个体被选中的概率相同，且个体之间是独立的;
- 变量可分为类型或数值变量，可以通过试验或观察研究得到;
- 研究两个变量间的关联时，通常使用解释变量来预测响应变量;
  - 相关 ≠ 因果关系

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

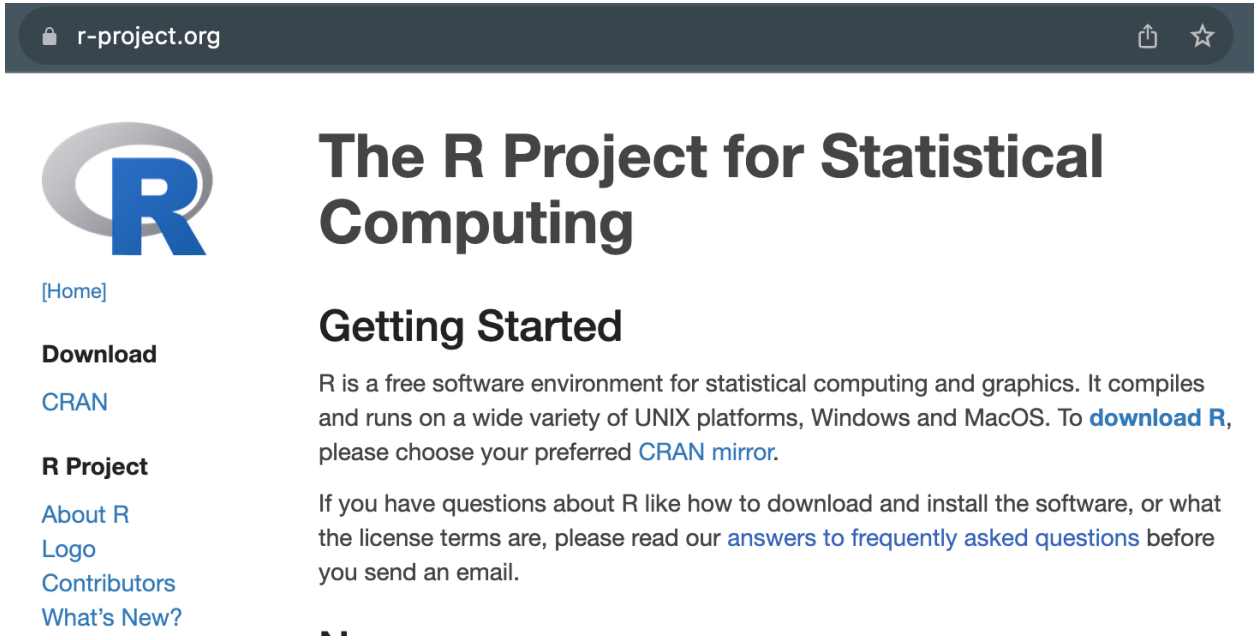
6. Discussions

- 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?

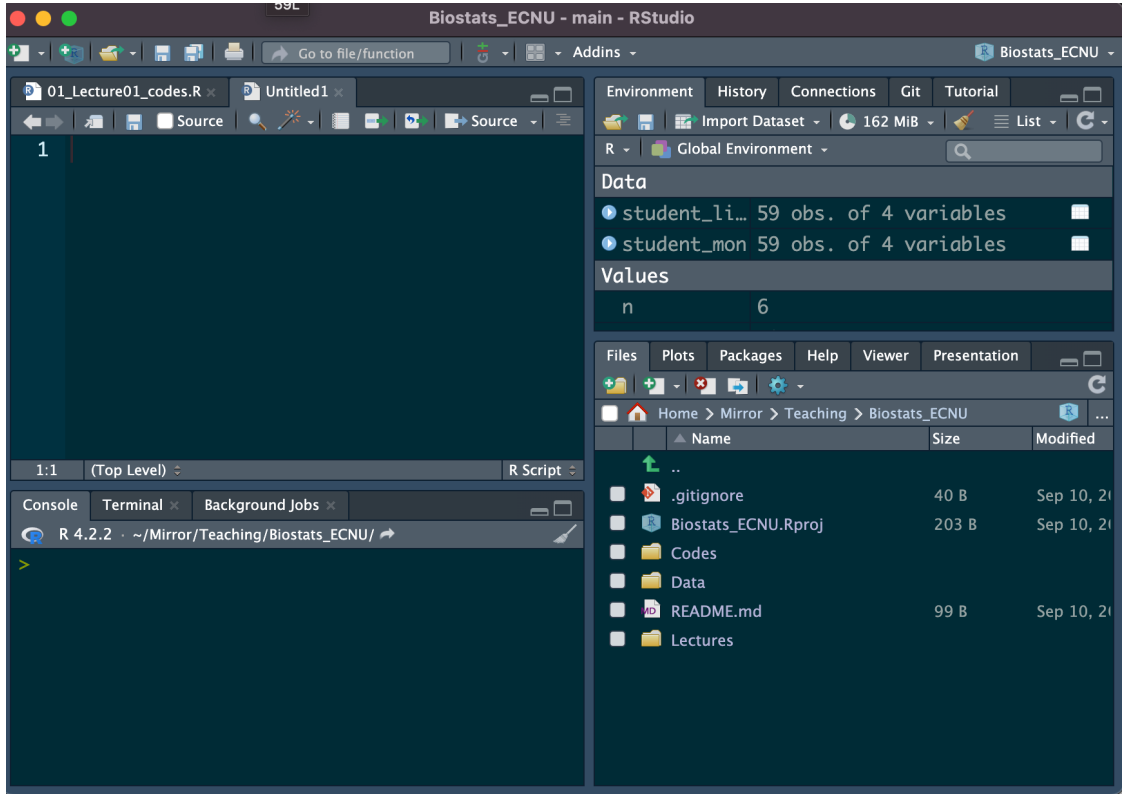
About R

- https://www.r-project.org/
  - An Introduction to R, 2023 (v4.3.1)

https://cran.r-project.org/doc/manuals/R-intro.pdf



Rstudio



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.