研究分类及 R 的描述性画图入门

Different Types of Research + Introduction to R

杨点溢

Slide content courtesy of Dr Melissa Sands
Department of Government
London School of Economics and Political Science

February 23, 2024

本节课内容

1. 研究分类介绍
2. 描述性推断 Descriptive Inference
3. 预测性推断 Predictive Inference
4. 因果性推断 Causal Inference
5. R 的入门：描述性命令

不同的研究哲学

- 实证主义 Positivism
  - 知识和事实是客观的，不同的人可以得到/复现相同的结果。There are objective facts that are observable and verible in the same way by different individuals.
  - 主要进行实证表述（positive statement）
- 建构主义 Constructivism/解释主义 interpretivism
  - 现实/知识是主观的。我们不可能客观的观察这个世界。Facts are socially embedded and constructed. Reality is subjective, and we cannot claim to objectively observe reality.
  - 常进行规范性表述/价值判断（normative statement/value judgement）

不同的研究哲学

- 实证主义 Positivism
  - 知识和事实是客观的，不同的人可以得到/复现相同的结果。There are objective facts that are observable and verible in the same way by different individuals.
  - 主要进行实证表述（positive statement）
  - 量化研究大多属于实证主义
- 建构主义 Constructivism/解释主义 interpretivism
  - 现实/知识是主观的。我们不可能客观的观察这个世界。Facts are socially embedded and constructed. Reality is subjective, and we cannot claim to objectively observe reality.
  - 常进行规范性表述/价值判断（normative statement/value judgement）
  - 常使用定性的研究方式（采访、文本分析等）

三种量化研究分类

量化研究可以分为以下三个大类

- 描述性推断 Descriptive Inference
- 预测性推断 Predictive Inference
- 因果性推断 Causal Inference

三种量化研究分类

量化研究可以分为以下三个大类

- 描述性推断 Descriptive Inference
- 预测性推断 Predictive Inference
- 因果性推断 Causal Inference
  - 本课程聚焦因果推断

本节课内容

1. 研究分类介绍
2. **描述性推断 Descriptive Inference**
3. 预测性推断 Predictive Inference
4. 因果性推断 Causal Inference
5. R 的入门：描述性命令

描述性推断 Descriptive Inference

总结和探索数据 summarizing and exploring data

- 趋势 Trend
- 分部 Distribution
- 高低多少 How much
- 社会网络 social network
- 量化文本分析 Text Analysis

随着定性定量结合研究和新型数据科学工具的兴起，描述性研究逐渐吃香，潜力很大。

案例：绘制意识形态市场的地图 Marketplace (Bonica, 2014)

开发了一种衡量美国政治候选人和选举金主的意识形态的方法（主要使用选举献金数据）

- 数据：超过 1 亿条美国联邦级和州级竞选献金记录

案例：绘制意识形态市场的地图 Marketplace (Bonica, 2014)

开发了一种衡量美国政治候选人和选举金主的意识形态的方法（主要使用选举献金数据）

- 数据：超过 1 亿条美国联邦级和州级竞选献金记录
- 使用同时捐助不同级别和机构选举的金主生成 “意识形态坐标”
  - 同空间选举献金分数 common-space campaign finance scores (CFscores)

案例：绘制意识形态市场的地图 Marketplace (Bonica, 2014)

开发了一种衡量美国政治候选人和选举金主的意识形态的方法（主要使用选举献金数据）

- 数据：超过 1 亿条美国联邦级和州级竞选献金记录
- 使用同时捐助不同级别和机构选举的金主生成 “意识形态坐标”
  - 同空间选举献金分数 common-space campaign finance scores (CFscores)
- 假设：贡献者一般在意识形态上更喜欢相近的候选人
  - 根据候选人理想点与贡献者理想点之间的距离分配贡献金额

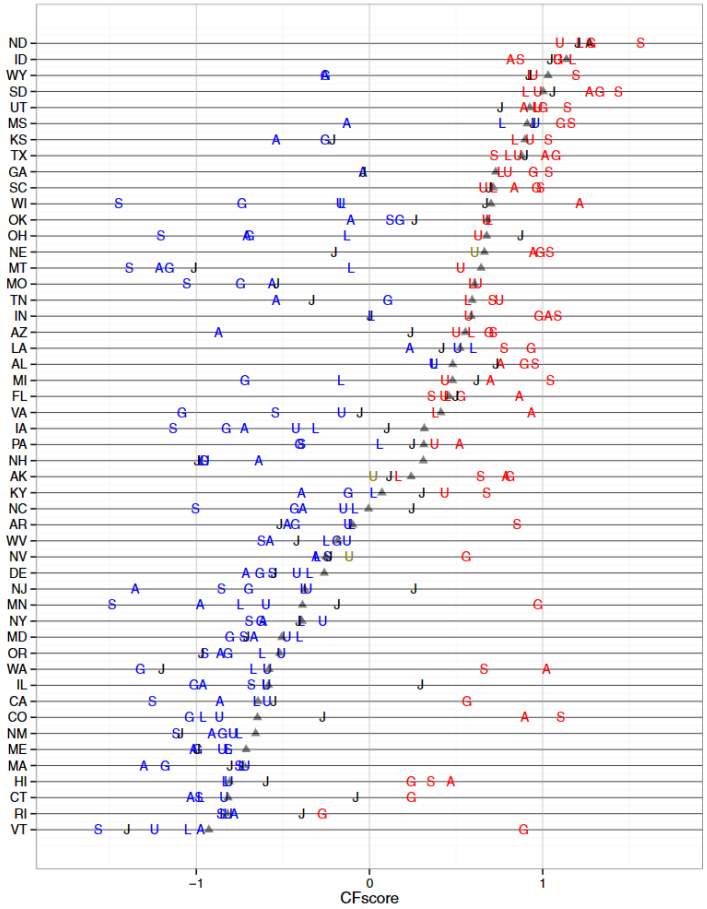
案例：绘制意识形态市场的地图 Marketplace (Bonica, 2014)

开发了一种衡量美国政治候选人和选举金主的意识形态的方法（主要使用选举献金数据）

- 数据：超过 1 亿条美国联邦级和州级竞选献金记录
- 使用同时捐助不同级别和机构选举的金主生成 “意识形态坐标”
  - 同空间选举献金分数 common-space campaign finance scores (CFscores)
- 假设：贡献者一般在意识形态上更喜欢相近的候选人
  - 根据候选人理想点与贡献者理想点之间的距离分配贡献金额
- 依靠向各个职位的候选人提供捐助的捐助者来搭建跨机构和政治层面的对比性

案例：绘制意识形态市场的地图 Marketplace (Bonica, 2014) cont'd

FIGURE 3 Ideological Summary of State Politics (2010)



Note: The symbols are interpreted as follows: G = Governor, A = Attorney General, S = Secretary of State, J = State Supreme Court (median), L = Lower Legislative Chamber (median), U = Upper Legislative Chamber (median), black triangle = mean ideal point of candidates elected in the state. The symbols are color coded by party (Dem = Blue; Rep = Red).

注：美国各州官员的意识形态立场

- 蓝色= 民主党
- 红色= 共和党
- G= 州长
- A= 检察长
- S= 州务卿
- J= 州法院中位水平
- L= 下院中位水平
- U= 上院中位水平
- 黑色三角 = 以上平均

本节课内容

1. 研究分类介绍
2. 描述性推断 Descriptive Inference
3. **预测性推断 Predictive Inference**
4. 因果性推断 Causal Inference
5. R 的入门：描述性命令

预测性推断 Predictive Inference

预测样本外的数据点 forecasting out-of-sample data points

- 根据已有的数据建立模型
- 与描述性推断非常相似
- 我们今后学到的量化模型也可以用于预测
  - 预测性研究可以大力出奇迹（因果推断则较为复杂）
  - 不用考虑内生性
  - 数据越多越好
- 未来方向：机器学习

案例：预测投票行为 (Nickerson and Rogers, 2014)

- 竞选使用数据构建预测模型来精准拉票
- 这些模型为选民数据库中的每个公民生成三种类型的“预测分数”：
  - 行为评分：使用过去的行为和人口统计信息来计算公民投票、献金等的概率。
  - 支持分数：调查公民样本关于他们的候选人/问题的支持。用于衡量总体偏好。
  - 响应度得分：使用随机现场实验 (Random Field Experiments) 的结果来预测公民将如何响应竞选活动。对异质处理影响 (heterogeneous treatment effects) 进行建模并用于预测目标人群的处理反应。
    - 这也可以被视为因果推断，见下文

案例：预测投票行为 (Nickerson and Rogers, 2014) cont'd

"大数据"

- 选民档案 (Voter Files)
- 消费者和财产数据
- 人口普查数据（社区特征）
- 以前的参与（例如捐赠、志愿服务）

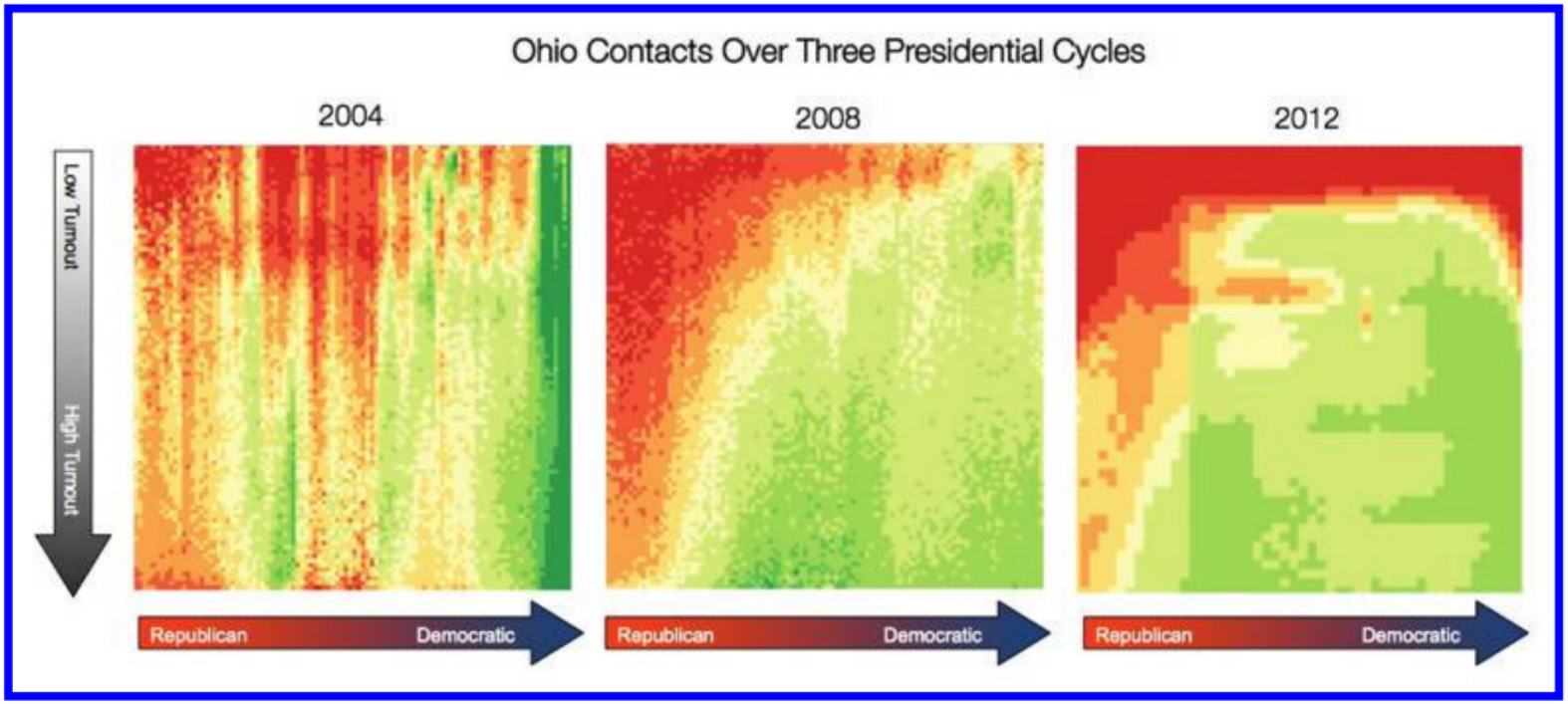
分析使用

- 简单回归（例如 OLS）
- 监督学习算法 (Supervised learning algorithms)

案例：预测投票行为 (Nickerson and Rogers, 2014) cont'd

Figure 1

Heatmap of Ohio Contacts over Three Presidential Cycles



Source: Derived from Catalist, LLC.

Notes: The x-axis is likelihood of supporting a Democratic candidate over a Republican candidate, ranging from 0 (left) to 100 (right). The y-axis is likelihood of voting ranging, from 100 (low) to 0 (high). Colors (or in grayscale, shade) of each cell indicate how many direct contacts were made by a particular campaign. In the grayscale version of the heatmap, darker means more contacts. In the color version, dark red represents the least contacts and dark green the most contacts. Readers can see the color heatmap in the online version of this paper.

本节课内容

1. 研究分类介绍
2. 描述性推断 Descriptive Inference
3. 预测性推断 Predictive Inference
4. **因果性推断 Causal Inference**
5. R 的入门：描述性命令

因果性推断 Causal Inference

因果推断 = 对反事实 (Counterfactual)的推断
因果影响 = 两个其他起始条件相同的平行世界中，由于 x 因素的不同而导致 y 因素的不同
举例:

- 假如房价没那么高，年轻人的生育意愿会 (would) 增加吗？
- 假如打击 LGBT 宣传，LGBT 会少吗？
- 假如我努力，会买得起房吗？

因果推断的根本问题 Fundamental Problem of Causal Inference

可惜没如果！反事实是不可被观察的 (The counterfactual is unobservable)

- 一个房价低的假想中国并不存在
- 一个打击 LGBT 宣传的假想欧美不存在/没打击 LGBT 宣传的假想中国不存在
- 一个假想中努力过得“我”不存在

所以没有简单的办法进行对比！

观察性研究 vs 实验

两种办法解决因果推断的根本问题：

- 实验 (experiments)
  - 往往涉及随机取样（random Sampling）
  - 一定涉及干预 (interference/manipulation)
    - 分为实验组 (treatment group) 和控制组 (control group)
    - 控制组配发安慰剂 (placebo)
- 观察性研究 (observational studies)
  - 基于研究者的观察 based on what the resercher observes
  - 不涉及研究者对实验对象的干预 does not involve interference/maniupulatoin

实验（在可行范围内）永远优于观察性研究

- 准实验 (quasi-experiment) 或者 自然实验 (natural experiment)
  - 最后几节课学的 DID SC SDID IV RDD 都是准实验
- “如同”随机取样和随机干预

因果推断范例：Gerber et al. (2008)

研究背景：

- 基于理性自利行为的选民投票率理论通常无法预测显着的投票率，除非它们考虑到公民从履行公民义务中获得的效用 (搭便车问题 free-rider problem)。
- 这种效用有两个方面：按照规范行事所带来的内在满足感 (intrinsic satisfaction) 和遵守规范的外在激励 (extrinsic incentives)。
- Gerber、Green 和 Larimer (2008) 在大规模实地实验中测试了内在动机 (intrinsic motives)，方法是对选民施加不同程度的外部压力。
  - 在 2006 年 8 月密歇根州初选之前向 180,002 个家庭发送一系列邮件。
  - Y_i (因变量)：是否在初选（primary）投票（是/不是）
  - D_i (处理/treatment)：不同的邮件

因果推断范例：Gerber et al. (2008) cont'd

- 公民责任 (Civic Duty)
  - 鼓励去投票
- 霍桑效应 (Hawthorne's effect)
  - 鼓励去投票
  - 收信者被告知研究人员将检查他们是否投票
- 家庭自身压力
  - 鼓励去投票
  - 收信者被告知政府对是否投票有记录
  - 告知收信者自己的家庭成员是否在最近两次选举中投票
- 邻居压力
  - 和家庭自身压力相似，不过信里同时协商街区的邻居是否在最近两次选举中投票

社会压力的实验处理 (Gerber et al., 2008)

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____
9999 BRIAN JOSEPH JACKSON		Voted	_____
9991 JENNIFER KAY THOMPSON		Voted	_____

实验结果 (Gerber et al., 2008)

	Control (Not Mailed)	Civic Duty (Encouraged to vote)	Hawthorne (Encouraged & Monitored)	Self (Encouraged, Monitored, Shown Own Past Voting)	Neighbors (Encouraged, Monitored, Shown Own & Others' Past Voting)
Percent Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

协变量平衡 (covariate balance) (Gerber et al., 2008)

当 $n \simeq 180,000$ 时，协变量几乎完全平衡：

TABLE 1. Relationship between Treatment Group Assignment and Covariates (Household-Level Data)

	Control	Civic Duty	Hawthorne	Self	Neighbors
	Mean	Mean	Mean	Mean	Mean
Household size	1.91	1.91	1.91	1.91	1.91
Nov 2002	.83	.84	.84	.84	.84
Nov 2000	.87	.87	.87	.86	.87
Aug 2004	.42	.42	.42	.42	.42
Aug 2002	.41	.41	.41	.41	.41
Aug 2000	.26	.27	.26	.26	.26
Female	.50	.50	.50	.50	.50
Age (in years)	51.98	51.85	51.87	51.91	52.01
N =	99,999	20,001	20,002	20,000	20,000

Note: Only registered voters who voted in November 2004 were selected for our sample. Although not included in the table, there were no significant differences between treatment group assignment and covariates measuring race and ethnicity.

本节课内容

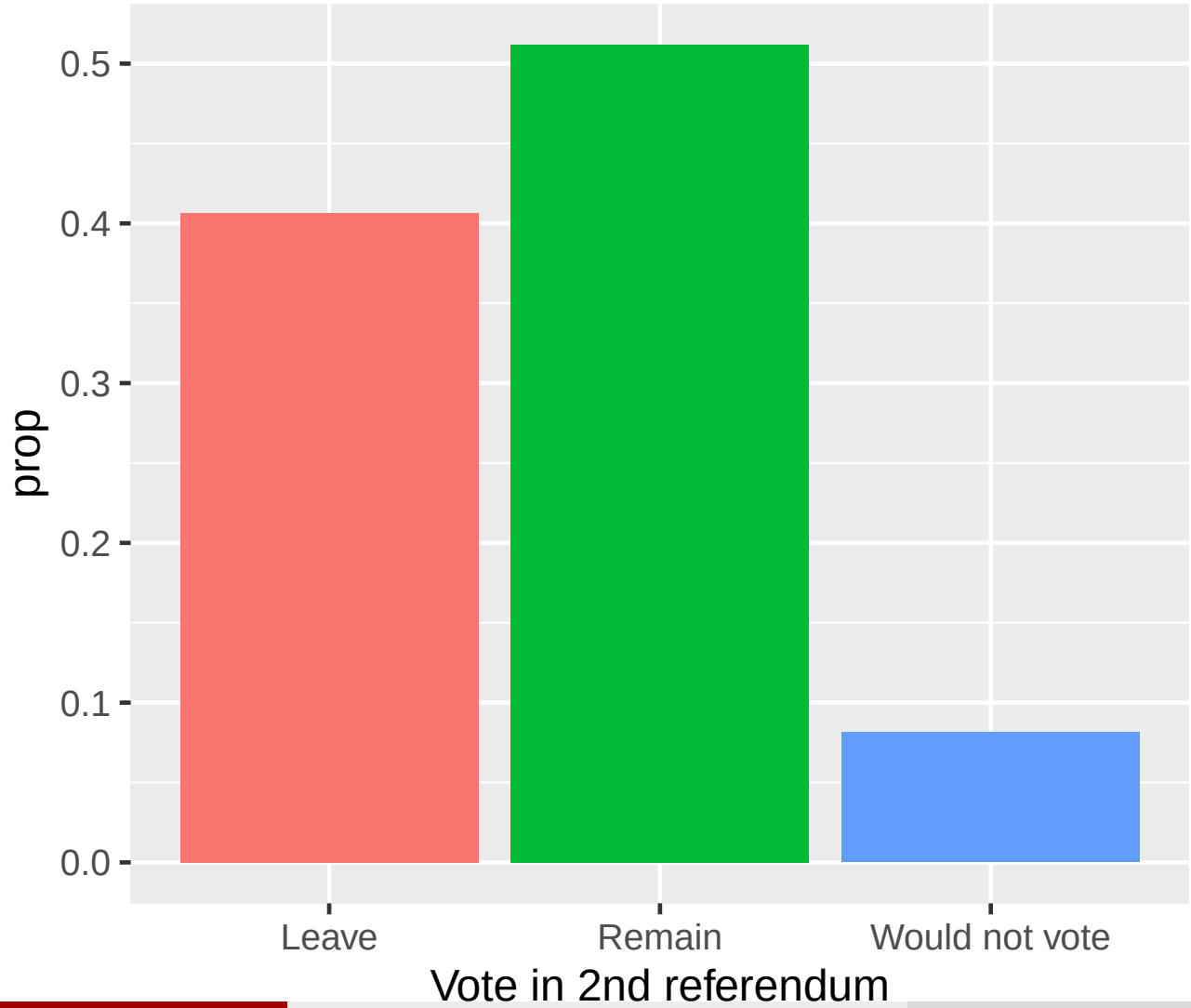
1. 研究分类介绍
2. 描述性推断 Descriptive Inference
3. 预测性推断 Predictive Inference
4. 因果性推断 Causal Inference
5. R 的入门：描述性命令

R 的入门

请参考 1.R

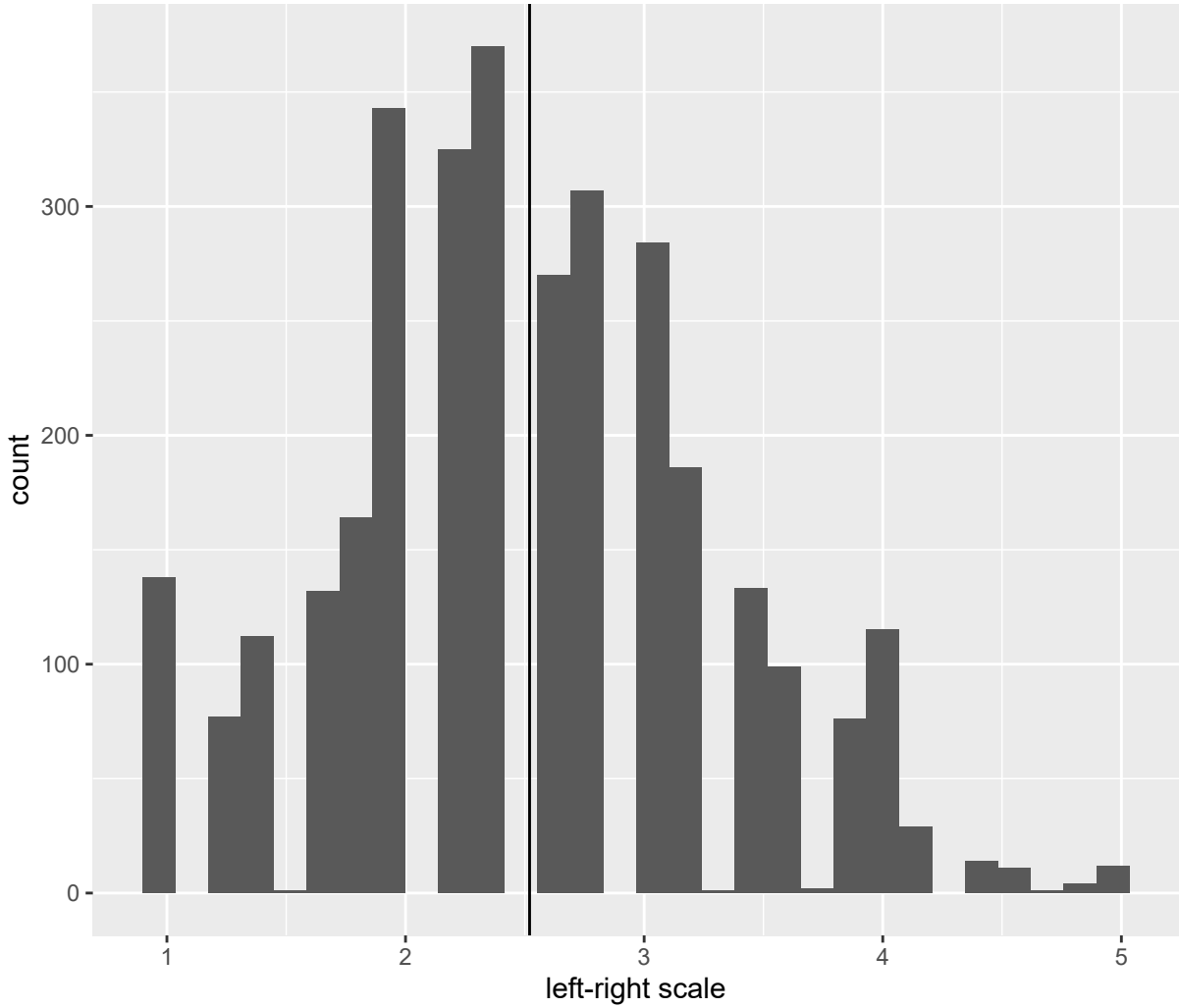
条形图 Bar Graph

- 横轴为分类 categorical 变量
- 纵轴表示分部多少



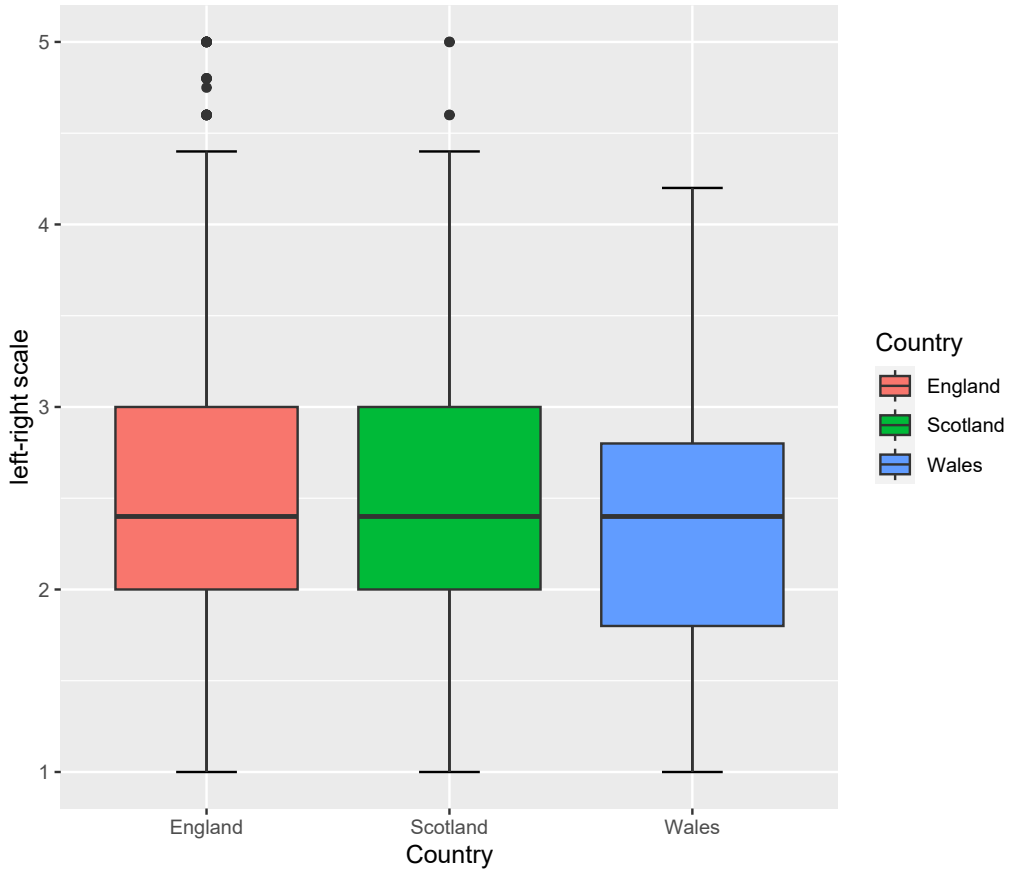
直方图 Histogram

- 横轴为连续 continuous 变量
- 纵轴表示分部多少



箱形图 Boxplot

- 展现数据分散情况
- 箱子下中上线分别为 P25(下四分位/lower quartile) P50（中位数/median） P75 (上四分位/upper quartile)
- “胡须” (Whiskers) 上下限为 P25 或者 p75±1.5IQR 内最近的值
  - IQR=P75-P25; 之外都是异常值 (outliers)，用点表示.



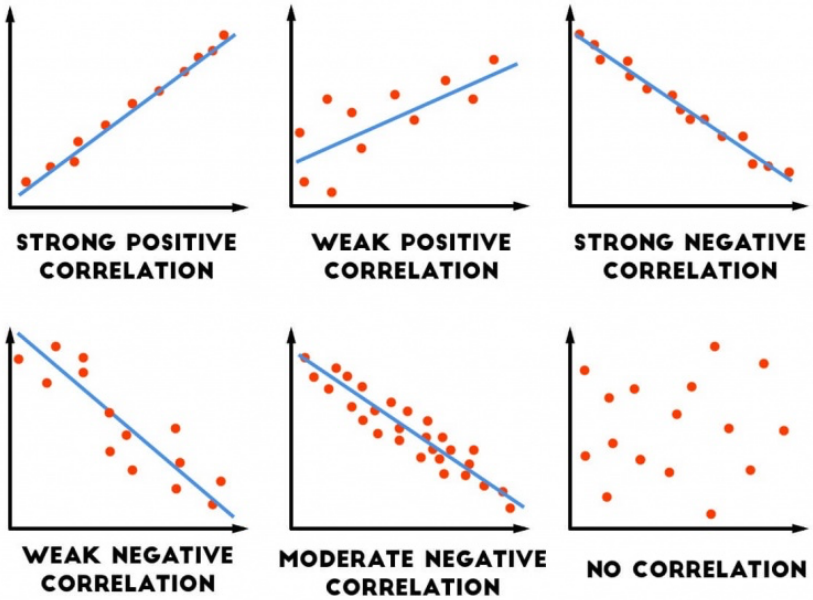
Cov 与 Corr

协方差 Covariance: $Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$

- Cov>0 正相关; Cov<0 负相关 (Cov 的值没有限制)

相关性 Correlation: $Corr(x, y) = R(x, y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$

- $-1 \leq Corr \leq 1$
- 和-1/1 越近说明相关性越强

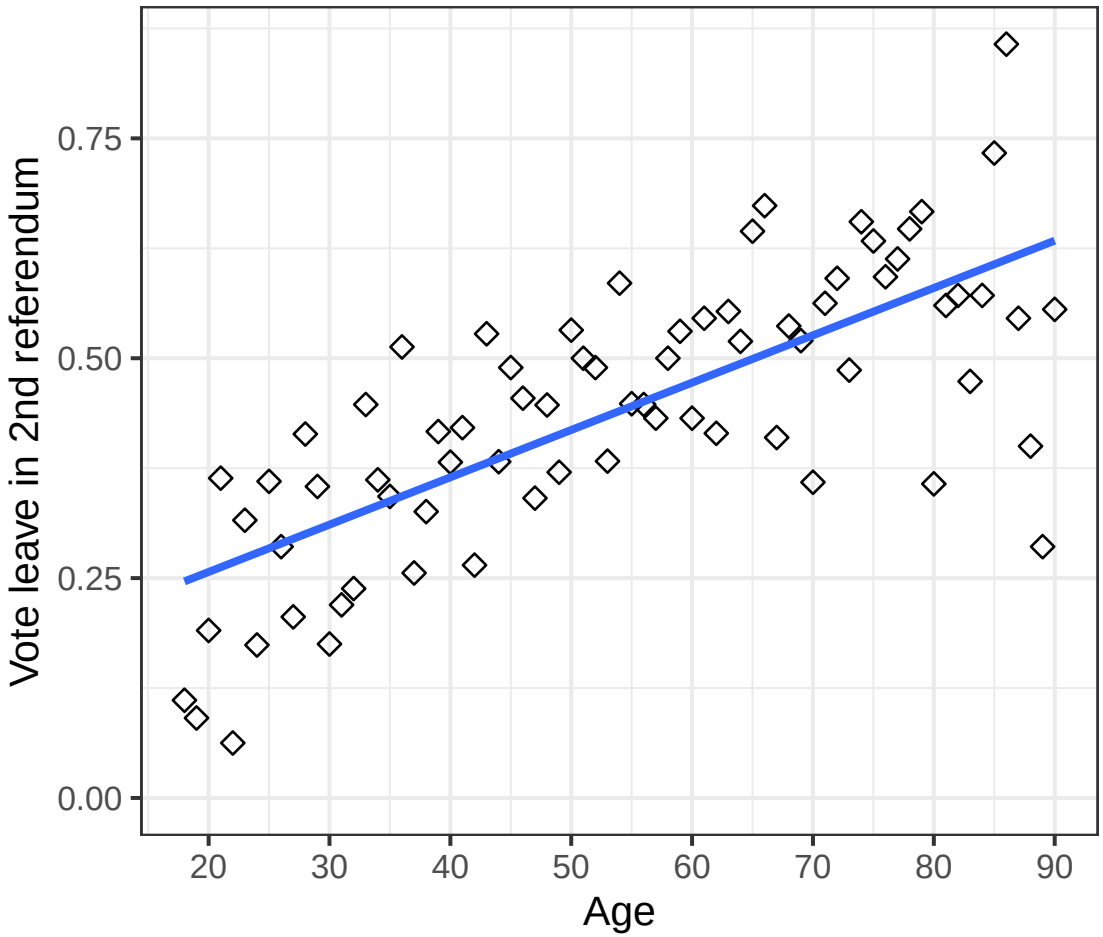


R^2

- 当只有两个变量时，$R^2 = R * R = Corr(x, y)^2$。
- $0 \leq R^2 \leq 1$
- 因变量的变化中有多少可以被自变量解释
  It is the proportion of the variation in y that can be predicted by x(s).

散点图与回归线 Scatter plot (with regression line)

- 找到“最佳拟合线” Find "the line of best fit"
- 怎么计算呢？最小二乘法 (Least Squares)
  - 最小化点到线的垂直距离平方的和
  - 从而最小化点到线的平均垂直距离



“表 1: 描述性统计” Table 1

- 因为通常是文章的第一个表，所以描述性统计也通常叫做“表 1” (Table 1).
- 对于连续性 continous 的数据总结，见下表
  - 必须有的内容：平均数 (Mean) 和标准差 (SD)
  - 可以有的内容：最小值 (Min) 最大值 (Max) 中位数 (Median) P25(Q1) P75(Q3)

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
bill_length_mm	165	1	43.9	5.5	32.1	44.5	59.6
bill_depth_mm	81	1	17.2	2.0	13.1	17.3	21.5
flipper_length_mm	56	1	200.9	14.1	172.0	197.0	231.0
body_mass_g	95	1	4201.8	802.0	2700.0	4050.0	6300.0
year	3	0	2008.0	0.8	2007.0	2008.0	2009.0

“表 1: 描述性统计” Table 1 cont'd

- 对于分类 categorical 数据总结，见下表
  - 必须有的内容：数量 (N) 和百分比 (% or pct)

		N	%
species	Adelie	152	44.2
	Chinstrap	68	19.8
	Gentoo	124	36.0
island	Biscoe	168	48.8
	Dream	124	36.0
	Torgersen	52	15.1
sex	female	165	48.0
	male	168	48.8
	NA	11	3.2

也可以分类（比如说实验组和对照组）对比

		female (N=165)		male (N=168)			
		Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
bill_length_mm		42.1	4.9	45.9	5.4	3.8	0.6
bill_depth_mm		16.4	1.8	17.9	1.9	1.5	0.2
flipper_length_mm		197.4	12.5	204.5	14.5	7.1	1.5
body_mass_g		3862.3	666.2	4545.7	787.6	683.4	79.9
year		2008.0	0.8	2008.0	0.8	0.0	0.1
		N	Pct.	N	Pct.		
species	Adelie	73	44.2	73	43.5		
	Chinstrap	34	20.6	34	20.2		
	Gentoo	58	35.2	61	36.3		
island	Biscoe	80	48.5	83	49.4		
	Dream	61	37.0	62	36.9		
	Torgersen	24	14.5	23	13.7		

References I

Bonica, A. (2014). Mapping the Ideological Marketplace. *American Journal of Political Science*, *58*(2), 367–386. https://doi.org/10.1111/ajps.12062

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, *102*(1), 33–48. https://doi.org/10.1017/S000305540808009X

Nickerson, D. W., & Rogers, T. (2014). Political campaigns and big data. *Journal of Economic Perspectives*, *28*(2), 51–74. https://doi.org/10.1257/jep.28.2.51