Lecture 3 – 数据展示 Data Display

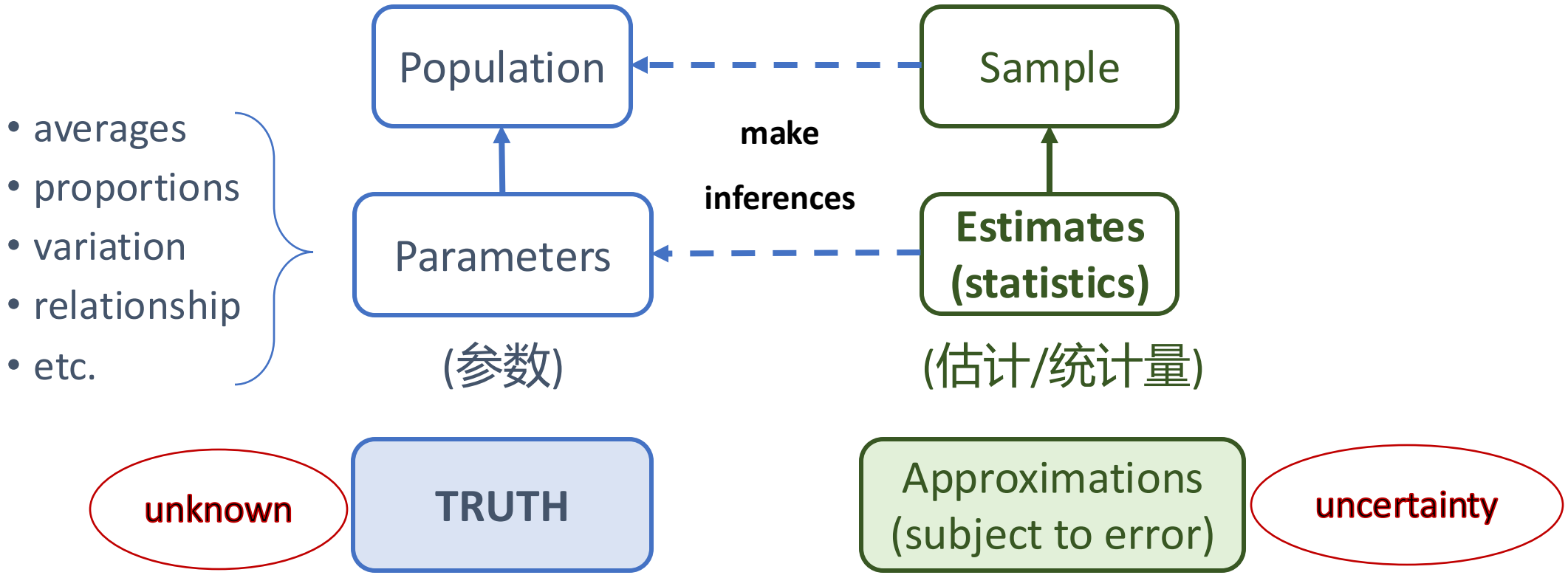
- 内容大纲
  - 回顾Lecture 1 & 2
  - 数据展示– 制图
    - 有效展示的原则 Principles of effective display
    - 展示数据模式的图形类型 Types of graphs to show patterns in data
  - 表格 Table
  - 总结 Summary
  - 课堂讨论 Discussion
  - R lab

第一次课后作业

- 作业内容：
  - Lecture 1-3
  - 统计学基础、基本统计量、数据展示（图形）及R操作
- 作业发布时间：2024年9月28日
- 作业提交截止时间：2024年10月9日

1. 回顾 – 统计学入门

- 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



1. 回顾 – 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
  - （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
  - （2）离散变量（discrete）：以不可分割的单位出现

1. 回顾 – 描述性统计量

- 如果数值变量符合正态分布
  - 算术平均值是最常见的描述频数分布位置的度量
  - 标准差是最常用的分布散布程度的度量
- 当数据形状偏离正态分布时,
  - 中值/中位数和四分位距能提供更为准确的信息
  - 四分位距能更好地表示分布主体的变异程度
  - 而均值和标准差更容易受到极端值的影响

- 比例是类型数据最重要的描述性统计量;

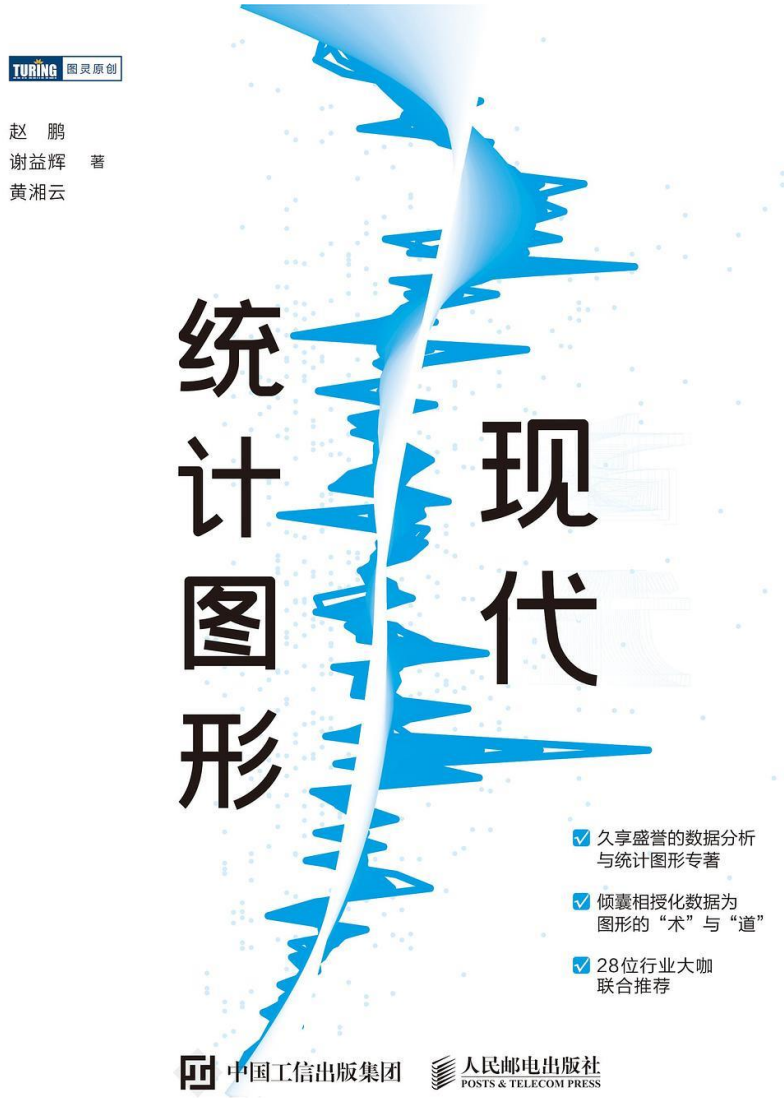
2. 数据展示 Display data

- 学习重点
  - 为什么制作图?
  - 有效展示的原则
  - 实现这些原则的图像类型

2. 数据展示

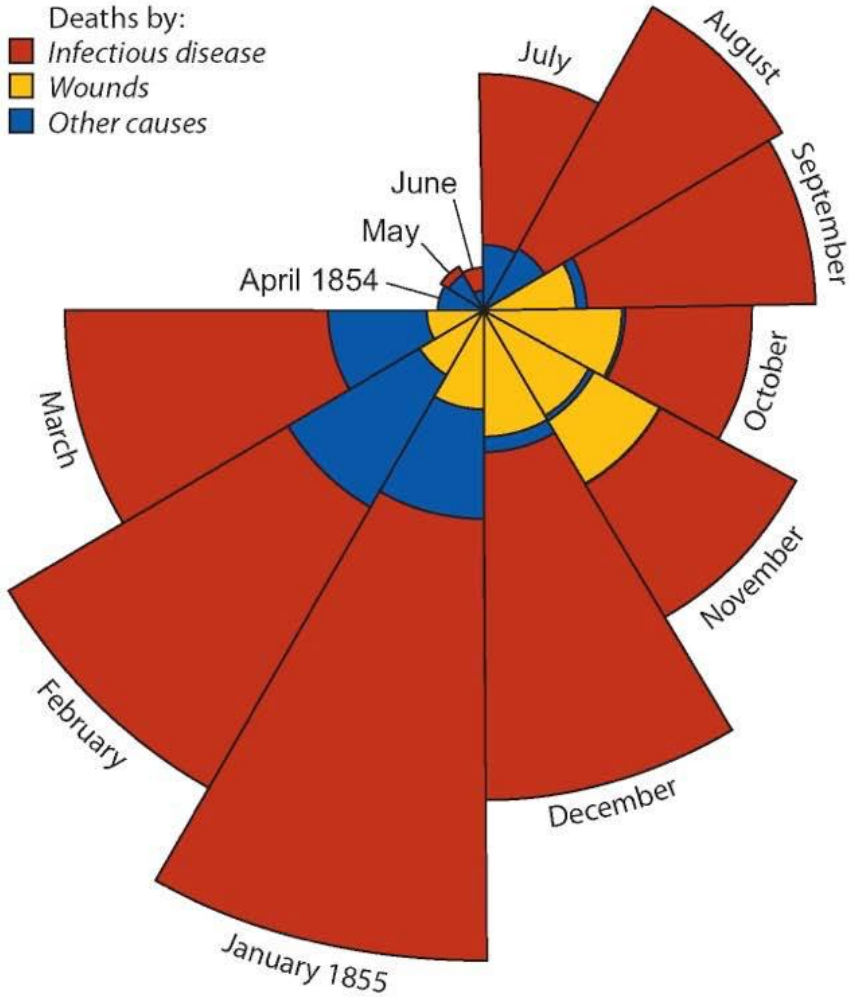
- 制图学 Graphics
  - 为什么做图? 一图胜千言
    - a picture is worth a thousand words

- 参考书《现代统计图形》
  - 在线阅读 https://bookdown.org/xiangyun/msg/



2. 数据展示

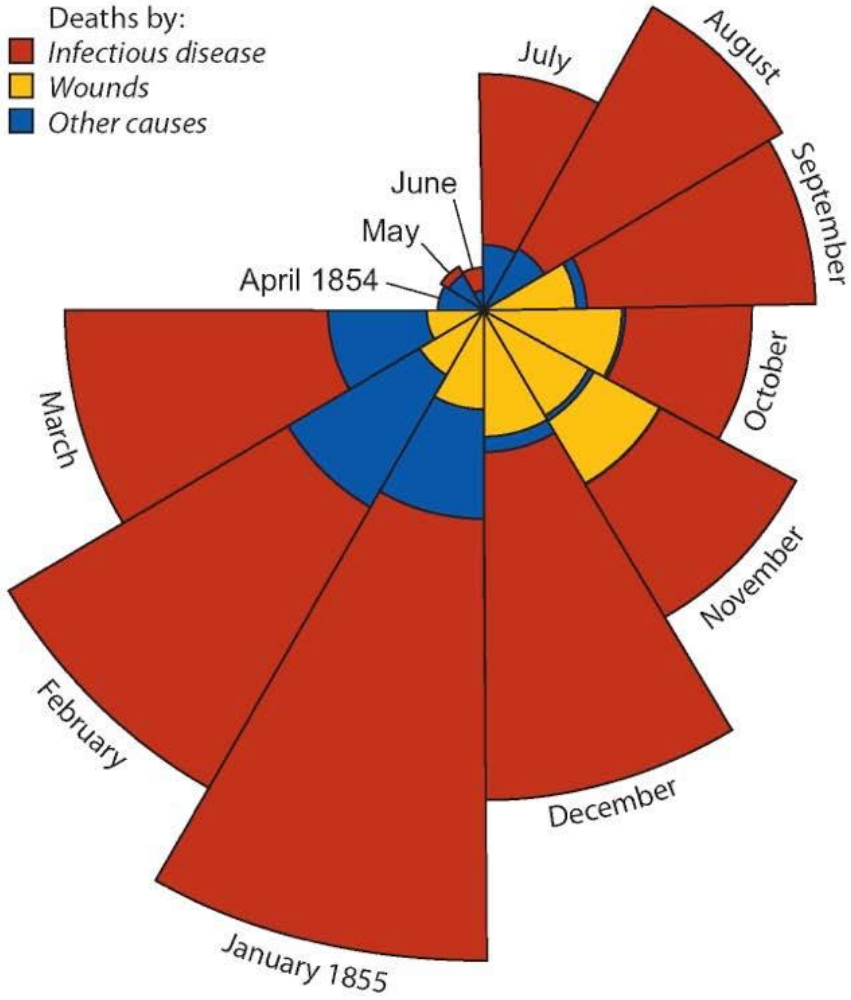
- 制图学 Graphics
  - 为什么做图?
    - 人眼是一个模式检测器（a pattern detector）
    - 图形能够直观地比较组间的测量值，并揭示变量之间的关系
    - 它们是向更广泛的受众传达结果的最佳方式
  - 例子： 提灯女士的玫瑰图（南丁格尔）
    - 克里米亚战争中英国军队的死亡人数
    - 真正影响战争伤亡的是缺乏有效的医疗护理!



(Whitlock & Schluter 2020)

2. 数据展示

- 制图学 Graphics
  - 为什么做图?
    - 人眼是一个模式检测器（a pattern detector）
    - 图形能够直观地比较组间的测量值，并揭示变量之间的关系
    - 它们是向更广泛的受众传达结果的最佳方式
  - 但我们应该尽量避免用饼图
    - 因为人眼对角度大小的敏感性不如长度
    - 替换：条形图或克利夫兰点图



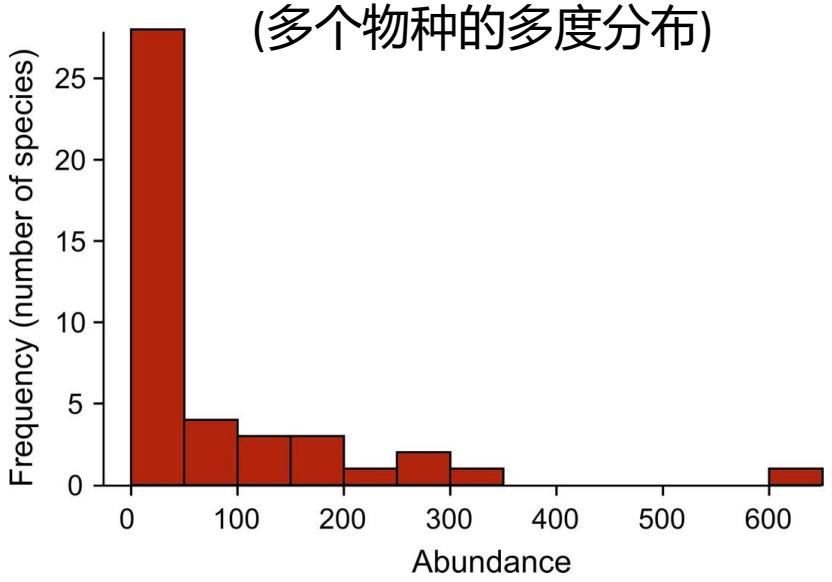
(Whitlock & Schluter 2020)

2. 数据展示

- 为什么要使用图形?
  - 揭示数据中的模式（To reveal pattern in data）
  - 任何数据分析或统计程序的第一步都是绘制数据图表并观察数据，因为分析和展示在很大程度上是一致的。
  - 最佳选择取决于数据的类型，是数值变量还是类型变量，以及目的是展示一个变量的测量结果还是两个变量之间的关联。

2. 数据展示

- 为什么要使用图形?
  - 揭示数据中的模式

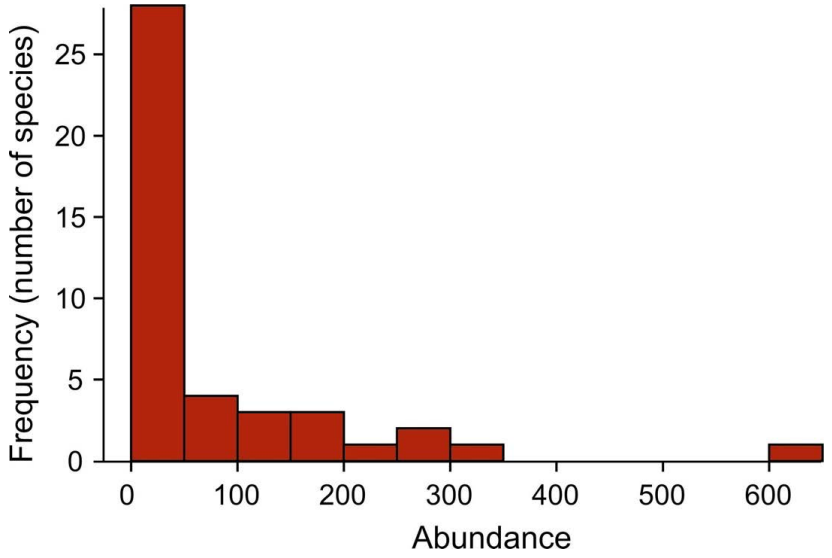


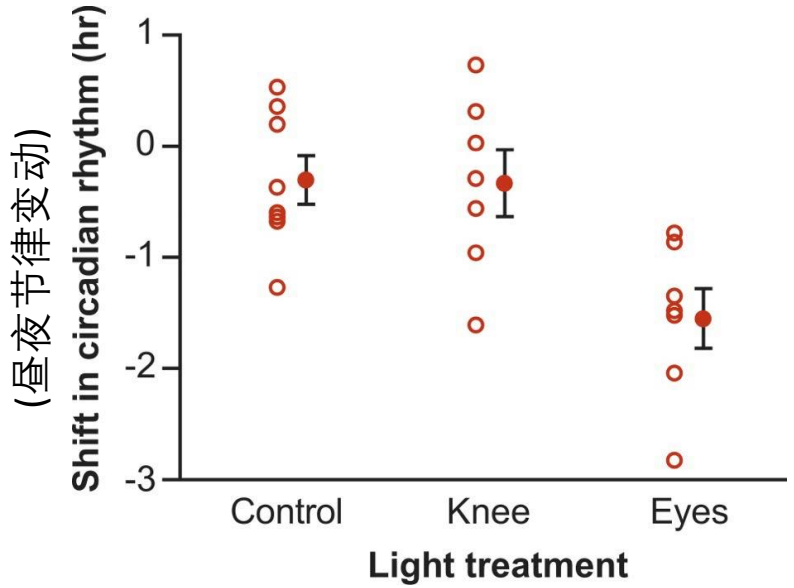
  - 频数分布 Frequency distributions
    - 频数：具有特定测量值的观测值的数量
    - 频数分布显示变量的每个值在样本中出现的概率。
    - 模式：分布的位置、范围和形状

(Whitlock & Schluter 2020)

2. 数据展示

- 为什么要使用图形?
  - 揭示数据中的模式
  - 频数分布 Frequency distributions
    - 分布的位置、范围和形状
  - 变量之间的关联
    - 两个或多个变量之间的关系
    - 组间差异

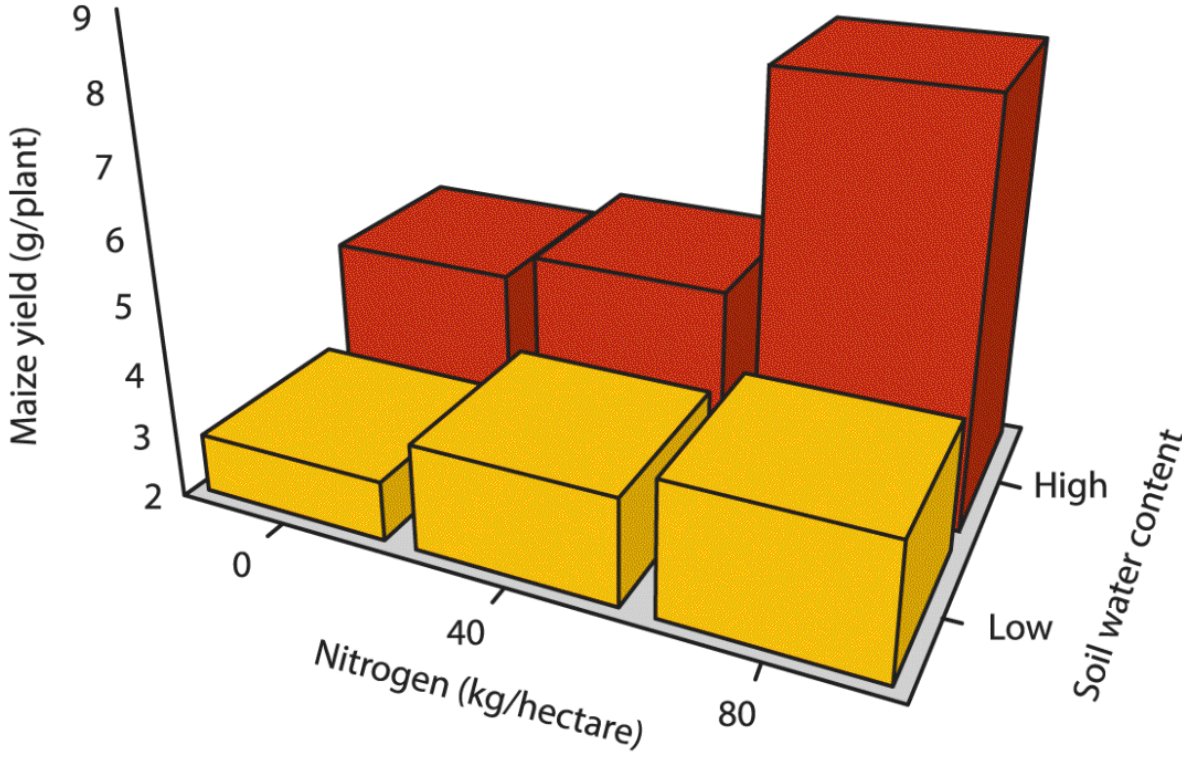




(Whitlock & Schluter 2020)

2. 数据展示

- 糟糕的图形
  - 盆栽玉米的实验结果
  - 组合条件：三个氮素浓度及两个土壤含水量
  - 柱形高度代表玉米的平均产量。
- 请问，你觉得有哪些错误/不足？

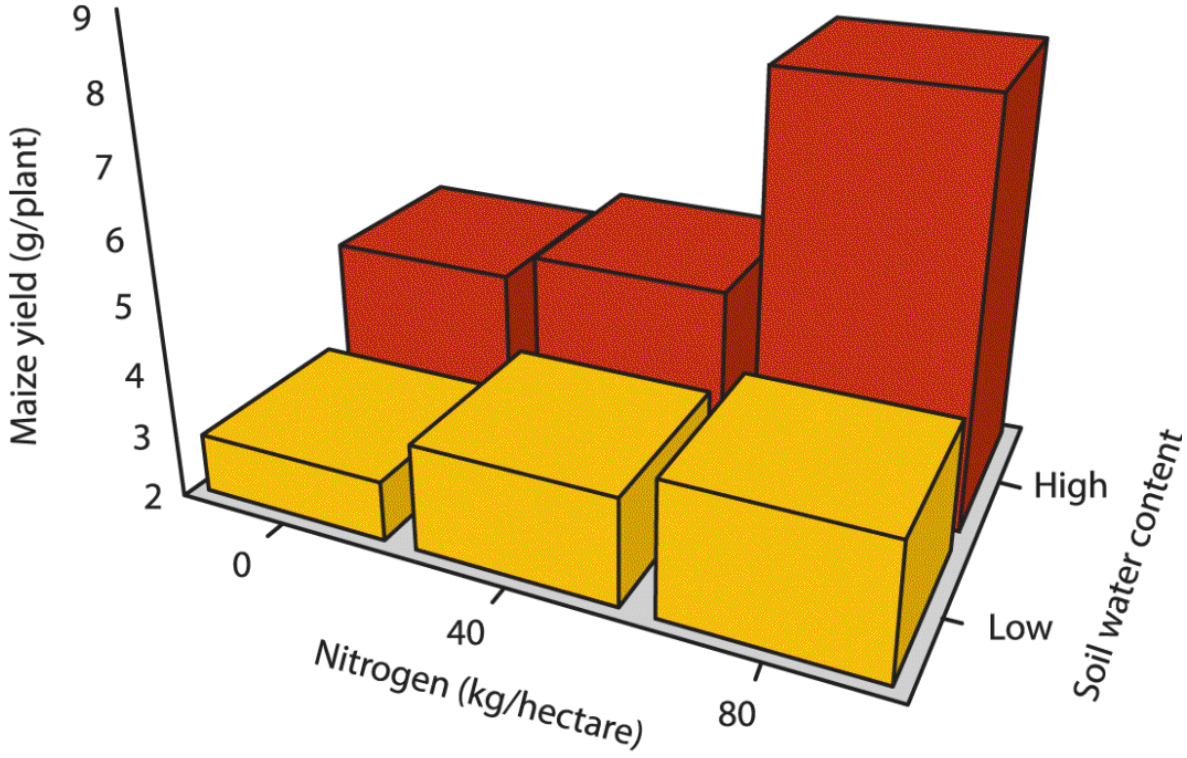


Whitlock & Schluter. *The Analysis of Biological Data*. 3e © 2020 W. H. Freeman and Company

2. 数据展示

- 糟糕的图形
- 错误/不足?
  - 图形隐藏了具体数据
  - 难以看出数据中的模式
  - 大小失真
  - 图形元素不清晰

  - 尽量避免 3D 图形，除非它是可以交互操作的

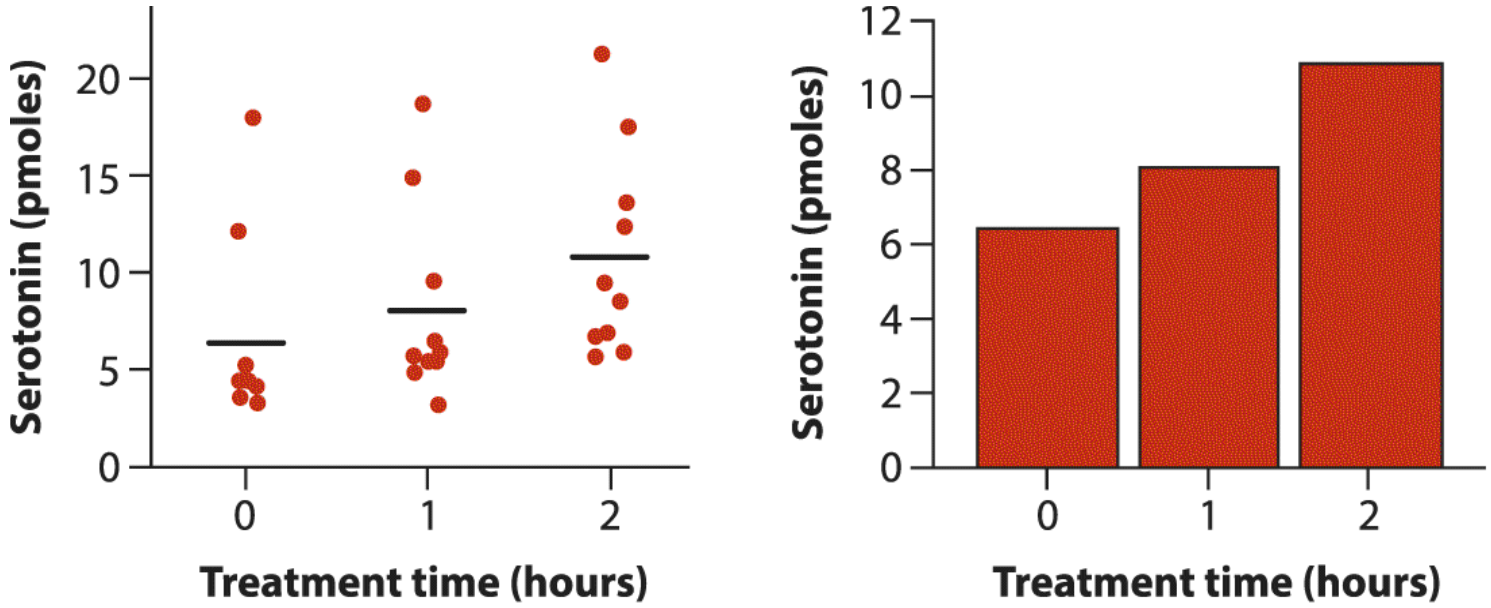


Whitlock & Schluter. *The Analysis of Biological Data.* 3e © 2020 W. H. Freeman and Company

2. 数据展示

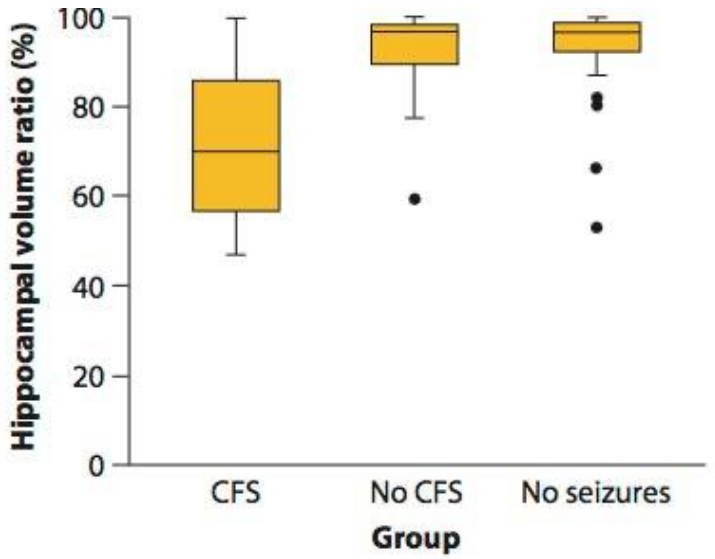
- 提高图表有效性（effectiveness）的四个有用原则（principle）
  - 显示数据 show the data
  - 使数据中的模式一目了然 make patterns in the data easy to see
  - 如实表示大小 represent magnitudes honestly
  - 清晰地绘制图形元素，尽量减少杂乱无章 draw graphical elements clearly

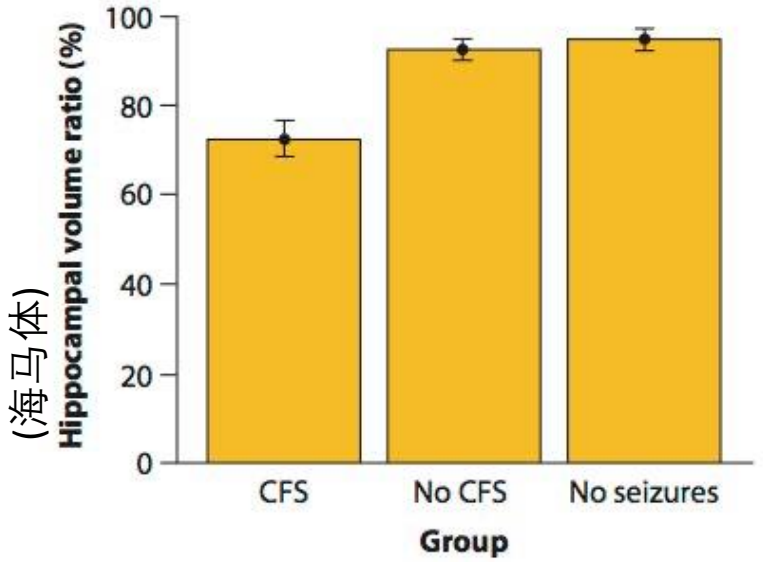
- 原则 1：显示数据
  - 显示单个数据点
  - 如：散点图（scatter plot）揭示了隐藏在条形图（bar graph）中的模式
    - 每个数据点都是 30 只蝗虫中的某一只的血清素水平
    - 实验将其高密度笼养 0、1 或 2 小时（0 为对照组）
    - 不同处理之间有明显的变化，但平均值与极值?



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 原则 1：显示数据
  - 当数据点较多时，箱形图（Box plot）可以代替散点图。
  - 哪种图形更有效？为什么？





一项调查 107 名耐药性癫痫患者海马体体积损失的研究（Cook et al.，1993 年）：

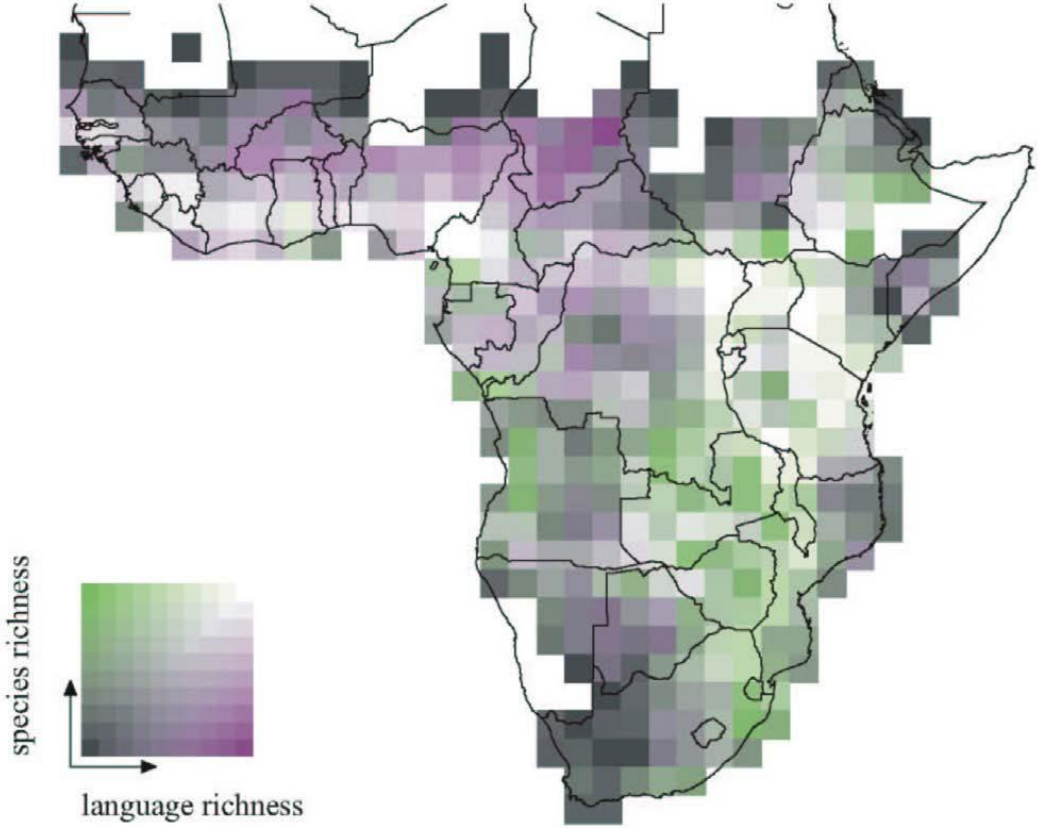
- 描述了海马体积损失（用核磁共振成像测量，即海马小半部分的体积除以大半部分的体积，以百分比表示）与患者病史之间的关联。
- 根据患者在童年时期是否有发热性癫痫发作（CFS）、非发热性癫痫发作（无CFS）和无癫痫发作的记录进行分组。

- 原则2：使数据中的模式一目了然

> *“Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency”* – Tufte (1983)

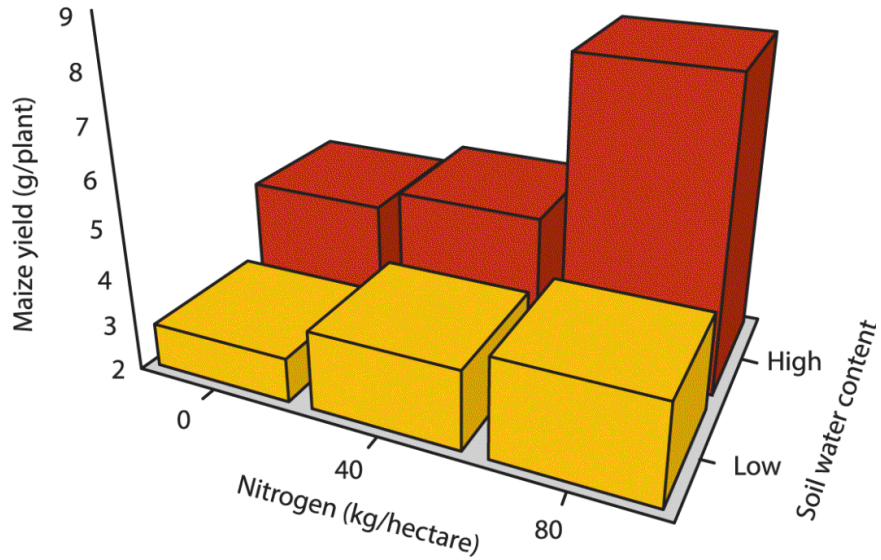
显示非洲大陆地图上每个方格内鸟类物种数量和不同人类语言数量（Reproduced from Moore et al., 2002）

- 这些数据的模式是什么?
- 你花了多长时间才 “看” 到?
- 是否很容易看出变量之间的相关性强度?



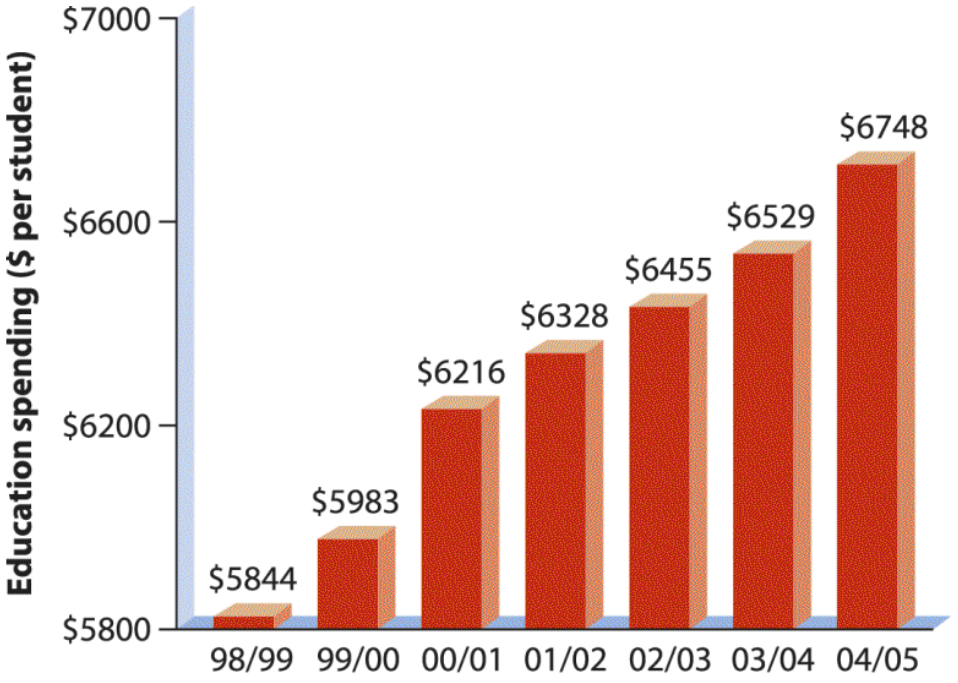
- 原则2：使数据中的模式一目了然
  - 显示相同数据的方法不止一种，可以尝试用不同的方法显示数据。
  - 远离三维效果和杂乱无章的图形，因为它们容易掩盖数据中的模式。
  - 数据中的主要模式是否较为容易辨识?
  - 避免在一张图形中包含过多信息。

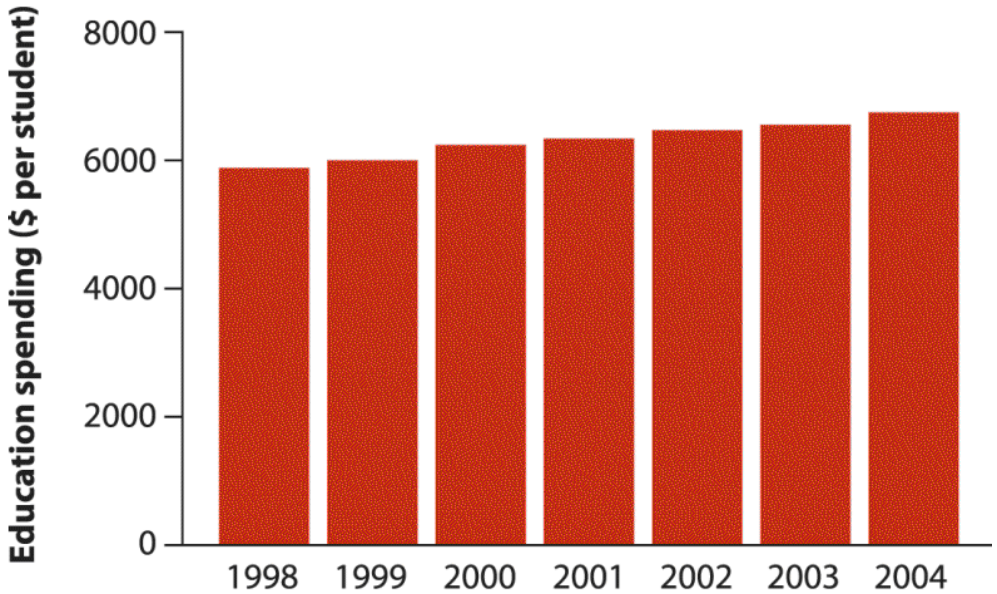
- 你将如何改进这幅图?



Whitlock & Schluter. *The Analysis of Biological Data.* 3e © 2020 W. H. Freeman and Company

- 原则3：如实表示大小
  - 重要的刻度涉及图形纵坐标轴上的最小值——基线（baseline）
  - 条形图的基线必须始终为零，因为人眼会本能地认为条形图的高度和面积与幅度成正比。

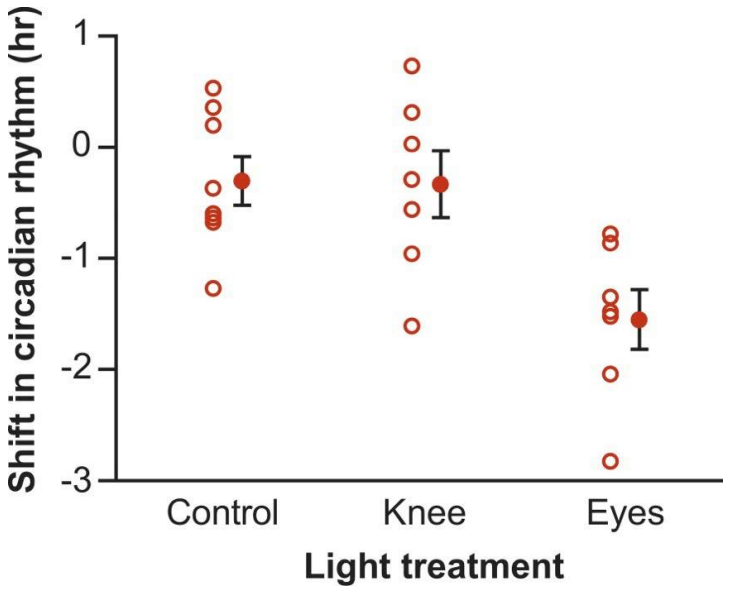




Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 原则3：如实表示大小
  - 重要的刻度涉及图形纵坐标轴上的最小值——基线（baseline）
  - 条形图的基线必须始终为零，因为人眼会本能地认为条形图的高度和面积与幅度成正比。
  - 如果主要目的是**显示不同处理之间的差异**，而不是比例大小，则不一定需要基线为零（如散点图）。

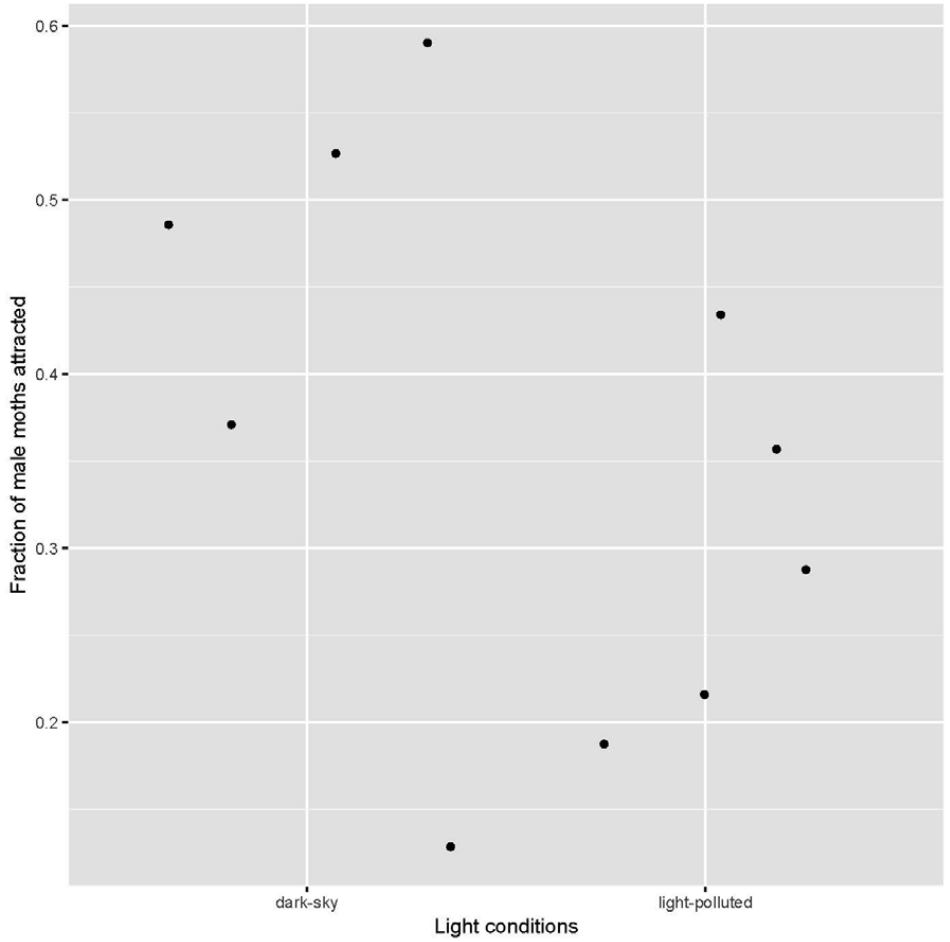
- 原则4：清晰绘制图形元素
  - 清晰的坐标轴和标签：X & Y（单位及刻度）
  - 可读性高的文本
  - 测量单位准确
  - 清晰可辨的图形符号（数据点或线）
  - 可区分的颜色
    - 例如，针对红绿色盲受众
  - 完整的图例（legend）
  - 误差条（error bar）
  - 网格线
  - 标题（title/caption）



- 原则4：清晰绘制图形元素

如何改进该图？

Altermatt 和 Ebert（2016 年）测量了来自 10 个不同种群的貂蛾（*Yponomeuta cagnagella*）的趋光性。其中5个种群位于人类灯光密集的城市地区。另外 5 个种群位于没有光污染的原始地区。



2. 数据展示

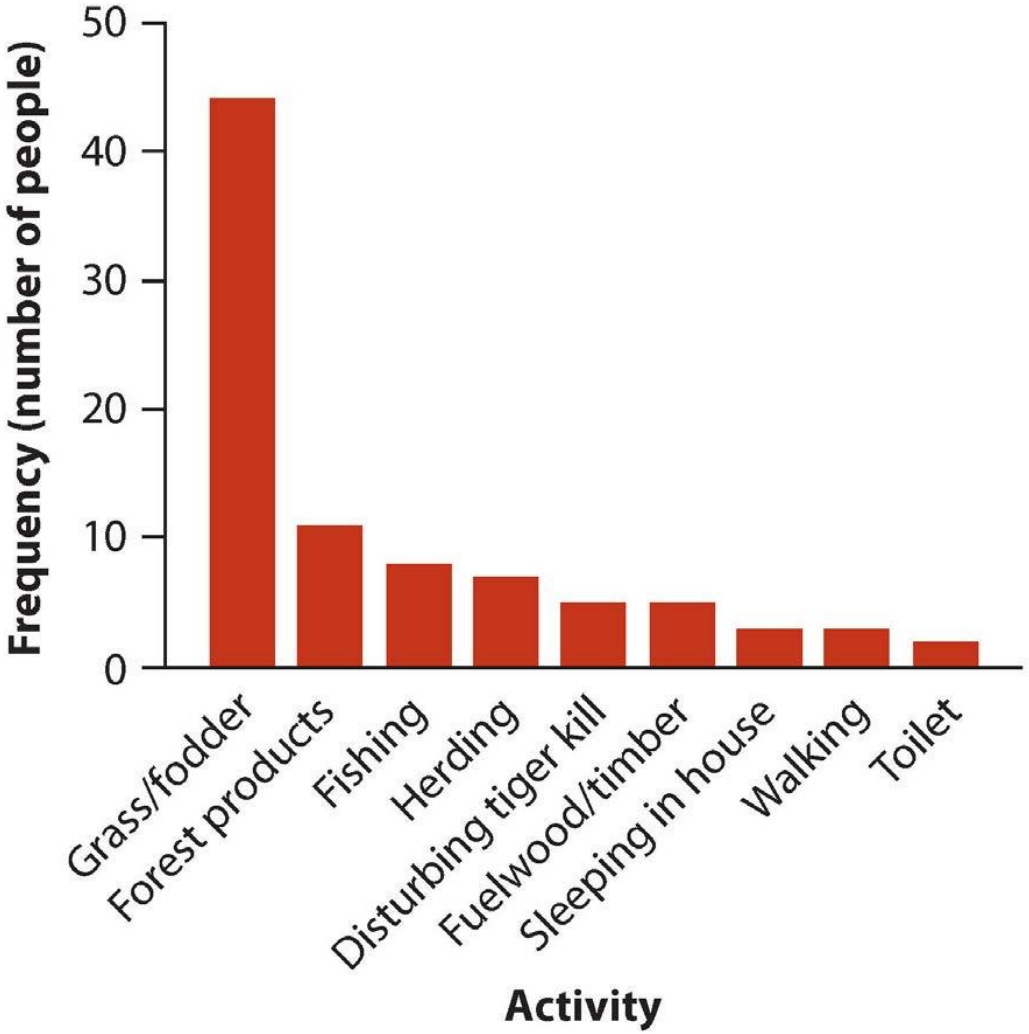
- 提高图表有效性（effectiveness）的四个有用原则（principle）
  - 显示数据 show the data
  - 使数据中的模式一目了然 make patterns in the data easy to see
  - 如实表示大小 represent magnitudes honestly
  - 清晰地绘制图形元素，尽量减少杂乱无章 draw graphical elements clearly

3. 图形类型 Types of graphs

- 哪种类型的图表最能显示数据？最佳选择取决于：
  - 数据类型：数值还是分类变量
  - 目标：显示一个变量的测量结果还是两个变量之间的关联

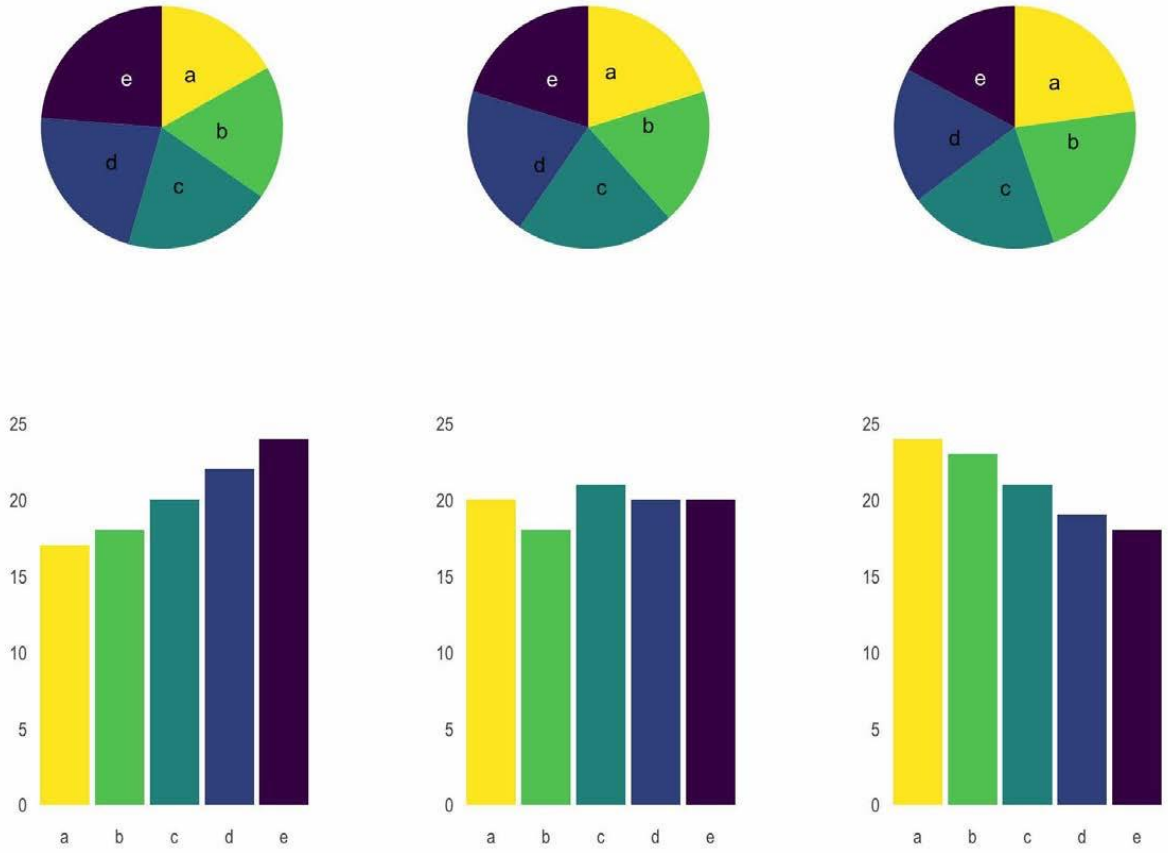
- 显示类别频数
  - 条形图（柱状图） Bar graph
  - 利用条形图的高度显示分类（分组）变量的频数分布
    - 零基线
    - 条形之间的空间用以强调高度
    - 类别顺序：通常从最常见到最不常见

（1979 年至 2006 年间，在尼泊尔奇特旺国家公园附近被老虎袭击并杀害的人类活动）



(Whitlock & Schluter 2020)

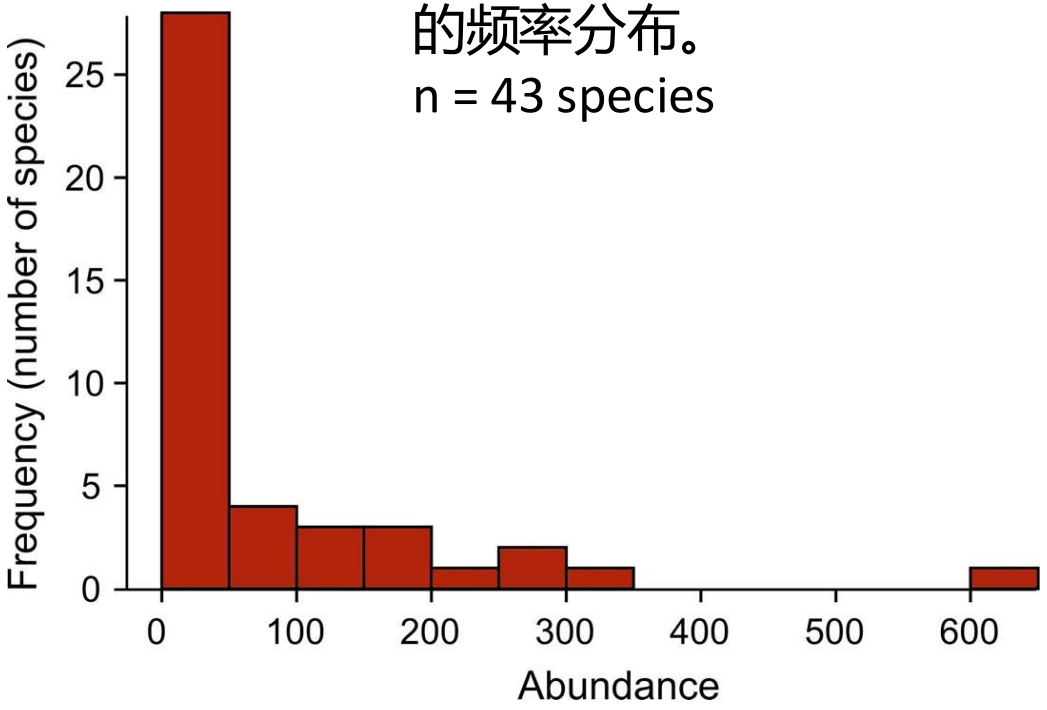
- 显示类别频数
  - 柱状图 Bar graph
  - 饼状图 Pie chart
  - 那种图形更有效?
  - 人类比较柱状图的相对面积比比较饼图的相对面积更有优势
    - 相对角度而言，人眼对长度更敏感



(www.data-to-viz.com from Schluter 2022)

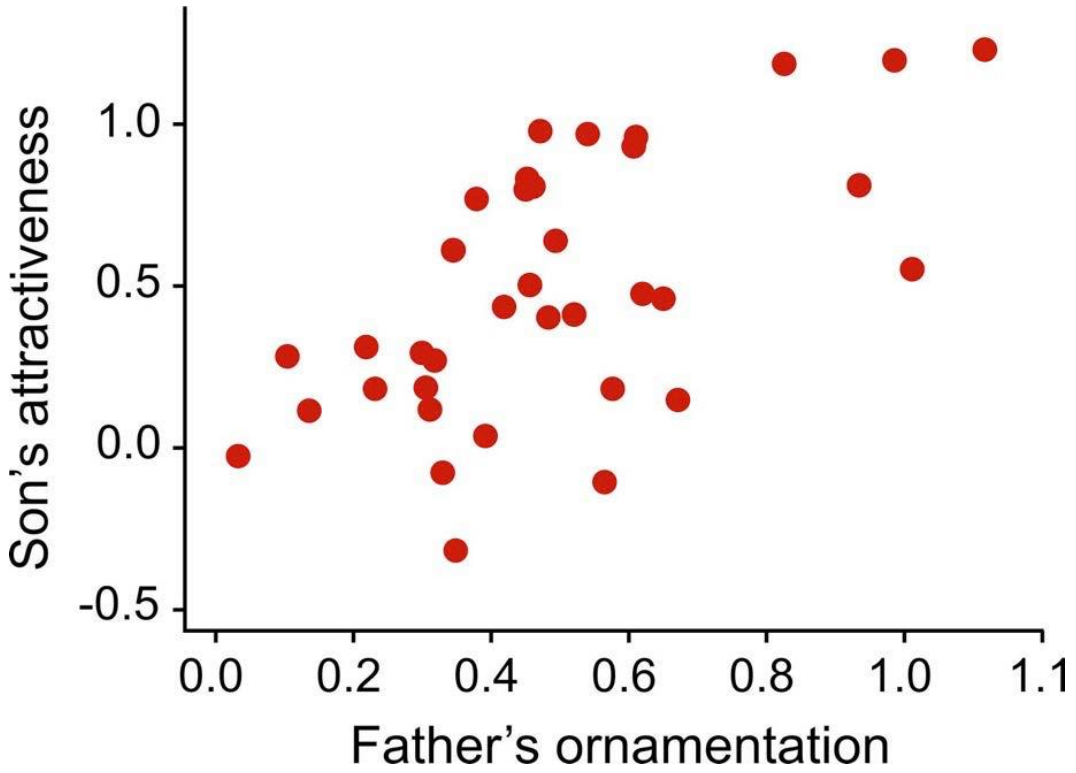
- 显示数值变量的频数分布
  - 直方图 Histogram
  - 使用直方图区域显示数值变量的频数分布
    - 基线为零
    - 柱状图无空格
    - 可选择区间宽度（bin breadth）分段区间（break）

Organ Pipe Cactus National Monument鸟类物种丰富度的频率分布。
n = 43 species



(Whitlock & Schluter 2020)

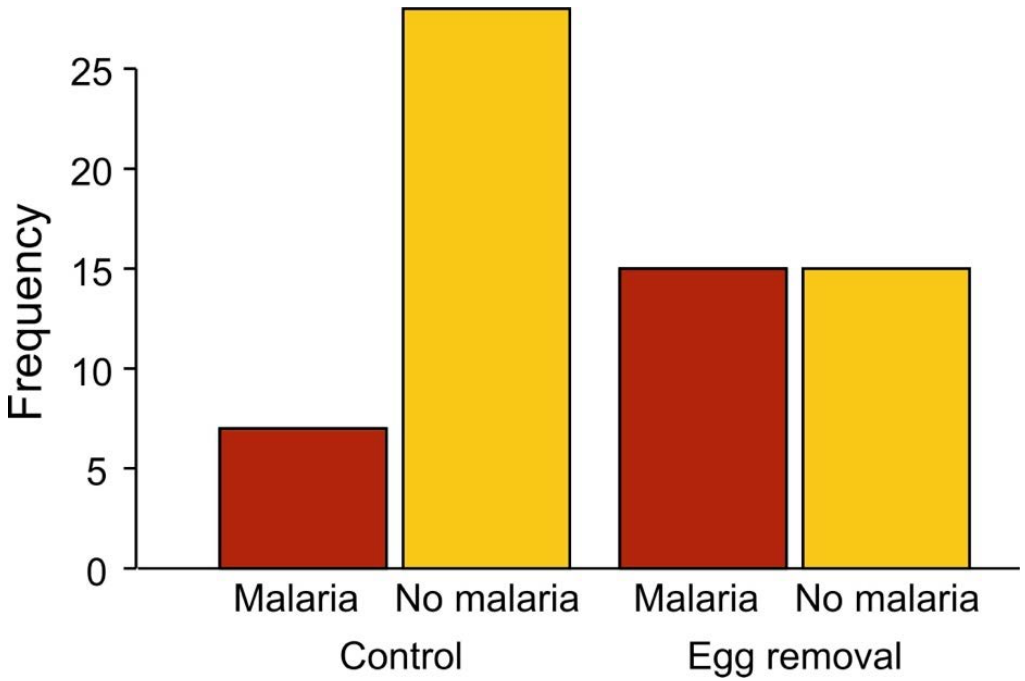
- 显示两个数字变量之间的关联
  - 散点图 Scatter plot
  - 通常可以使用非零基线
  - 目的是显示关联
    - 而不是0之上的高度
  - 数据点填充图形的可用空间



雄性孔雀鱼（guppies）的外表装饰（ornamentation）与其雄性后代平均吸引力的关系。
n = 36 families

(Whitlock & Schluter 2020)

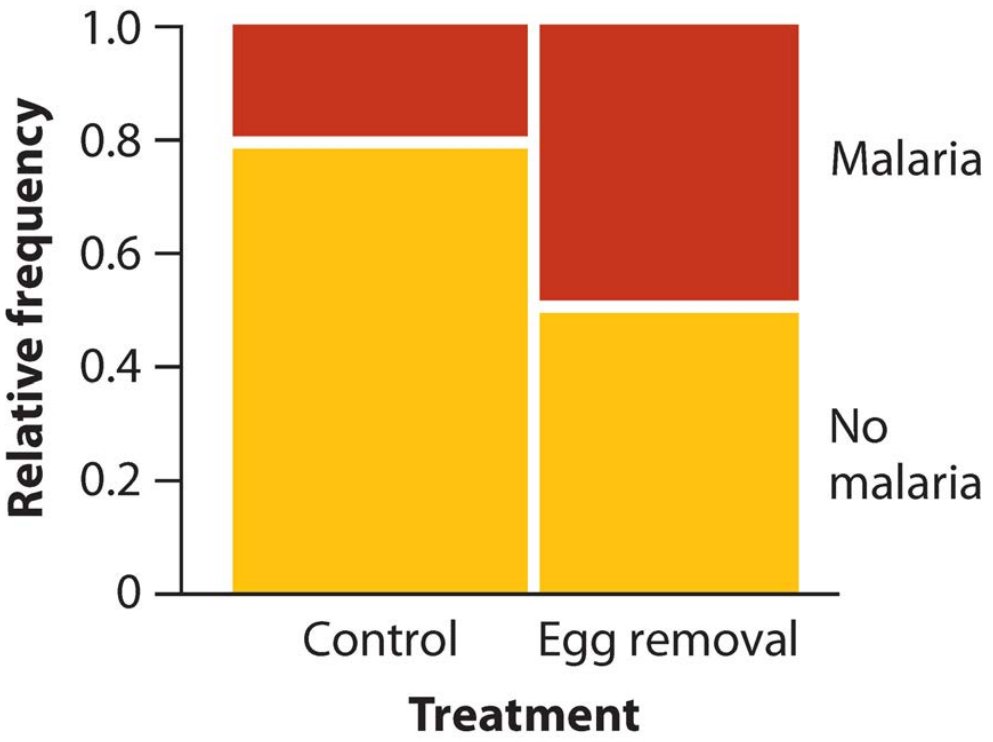
- 显示分类变量之间的关联
  - 分组条形图 Grouped bar graph
  - 利用条形图的高度显示两个（或多个）分类变量之间的关联
    - 解释变量 = outer groups
    - 响应变量 = inner groups
    - 基线为零（因此高度与频率成正比）
    - 外层组之间的条形间距更宽



雌性大山雀（great tits）的疟疾发病率与试验性治疗的关系。n = 65 birds

(Whitlock & Schluter 2020)

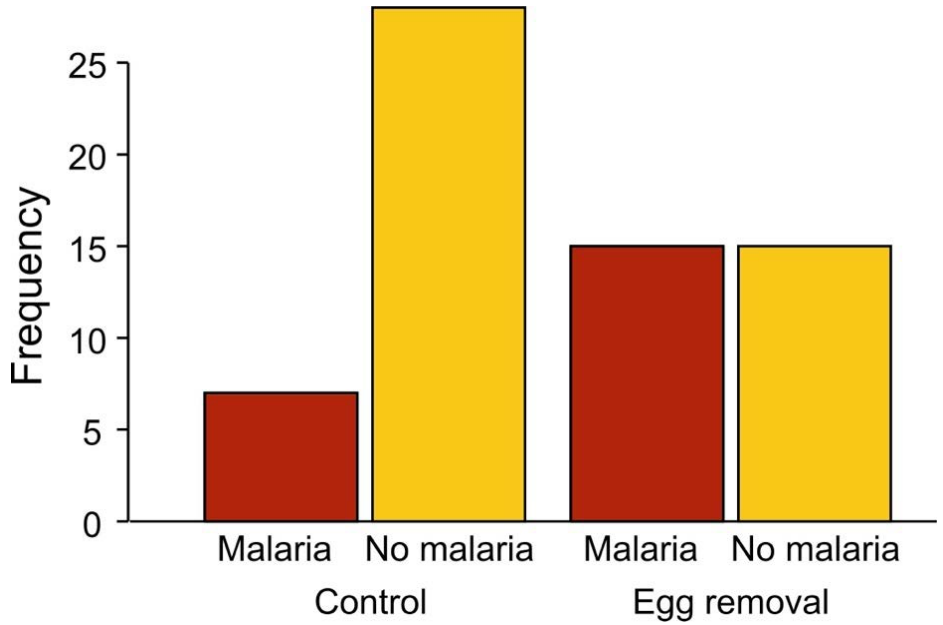
- 显示分类变量之间的关联
  - 马赛克图 Mosaic graph
  - 利用矩形面积显示两个（或多个）分类变量之间的关联。
    - 解释变量沿X轴排布
    - 响应变量沿Y轴叠加
    - 面积与频数成正比
    - 注意Y轴范围[0, 1]为相对频率
    - 类似于或然率表的图形表示法

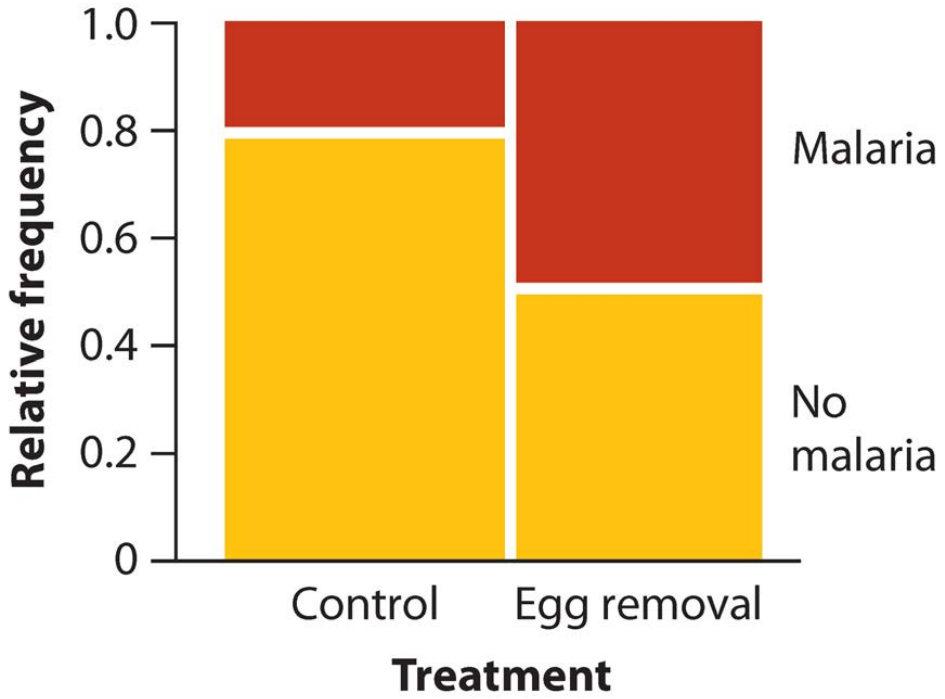


雌性大山雀（great tits）的疟疾发病率与试验性治疗的关系。n = 65 birds

(Whitlock & Schluter 2020)

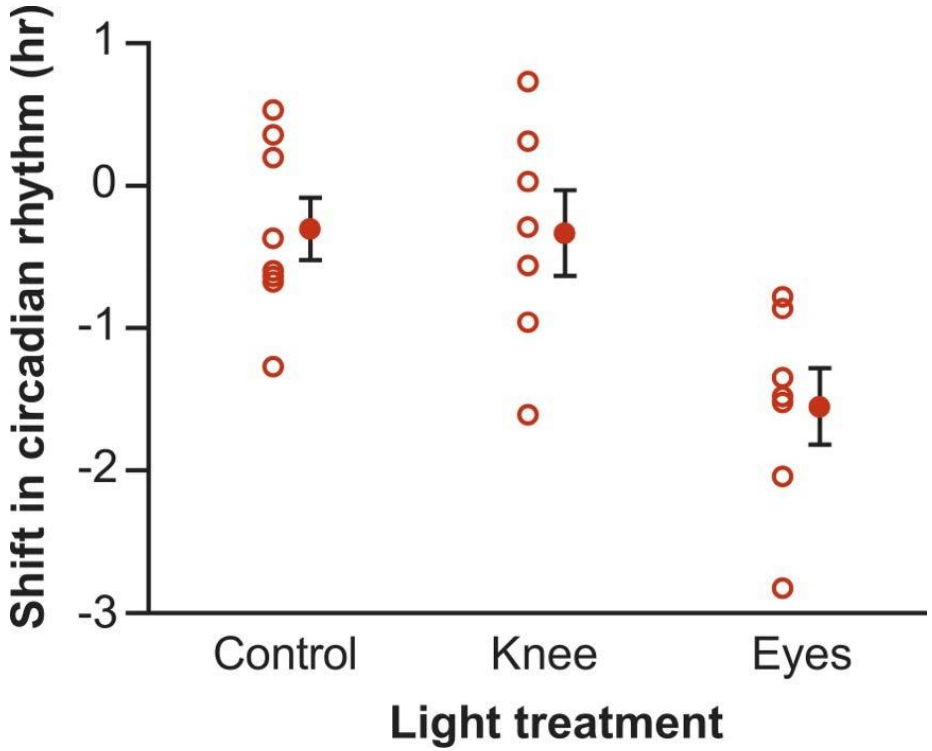
- 显示分类变量之间的关联
  - 分组条形图 vs . 马赛克图
  - 哪个图形更具有效性?





(Whitlock & Schluter 2020)

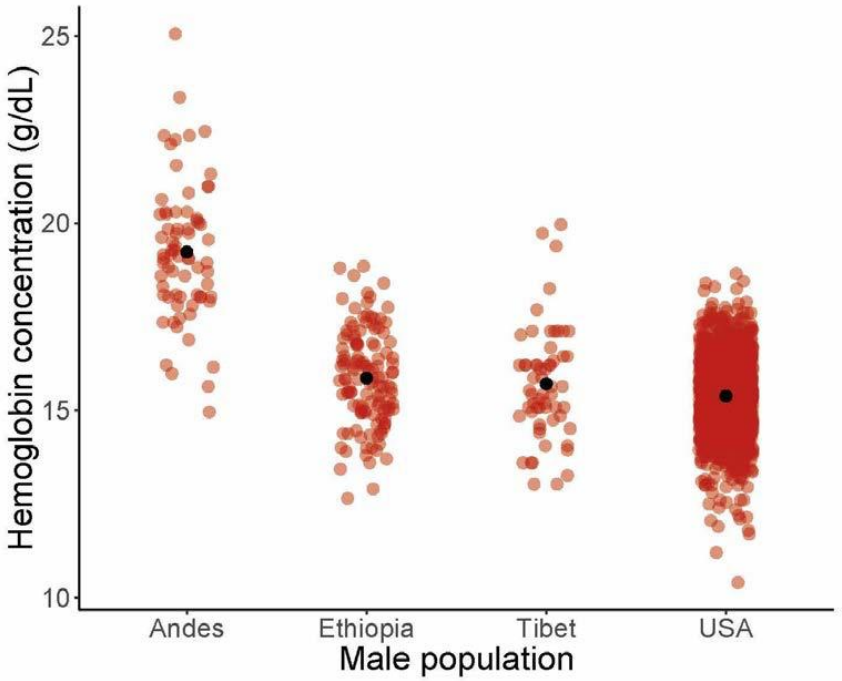
- 显示数值变量和分类变量之间的关联
  - 带状图 Strip graph
  - 显示组间差异
    - 显示数据点
    - 非零基线通常可以
      - 目标是变量间关联而非幅度或频数
    - 数据点填充可用空间



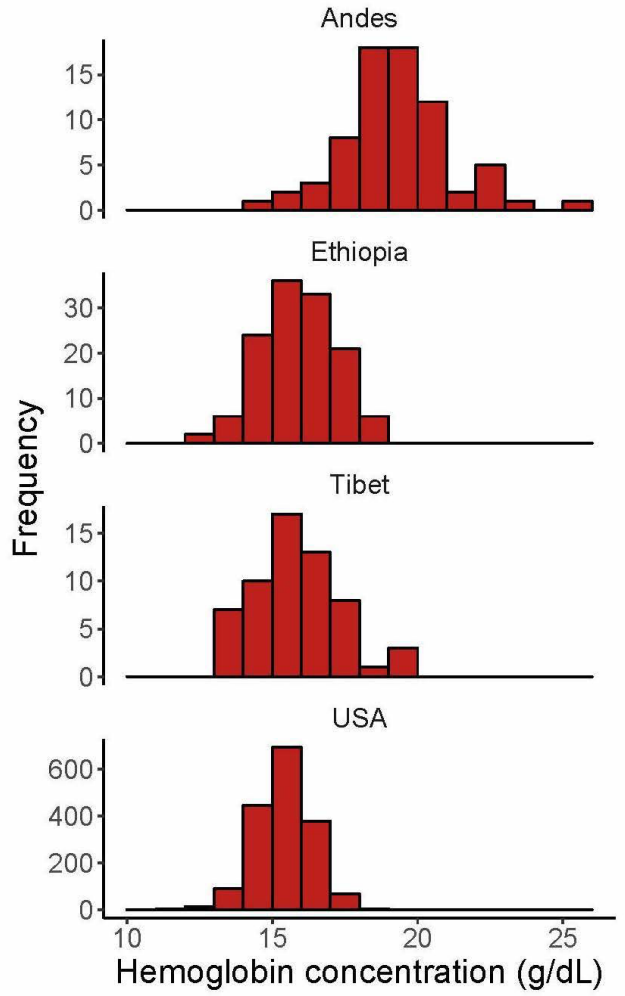
22 名受试者褪黑激素分泌昼夜节律的相位变化 (open circles)，及各组均值±1 SE（标准误）

(Whitlock & Schluter 2020)

- 显示数值变量和分类变量之间的关联
  - 带状图 vs. 多个直方图
  - 带状图的数据点太多
  - 将直方图垂直堆叠以进行最佳比较

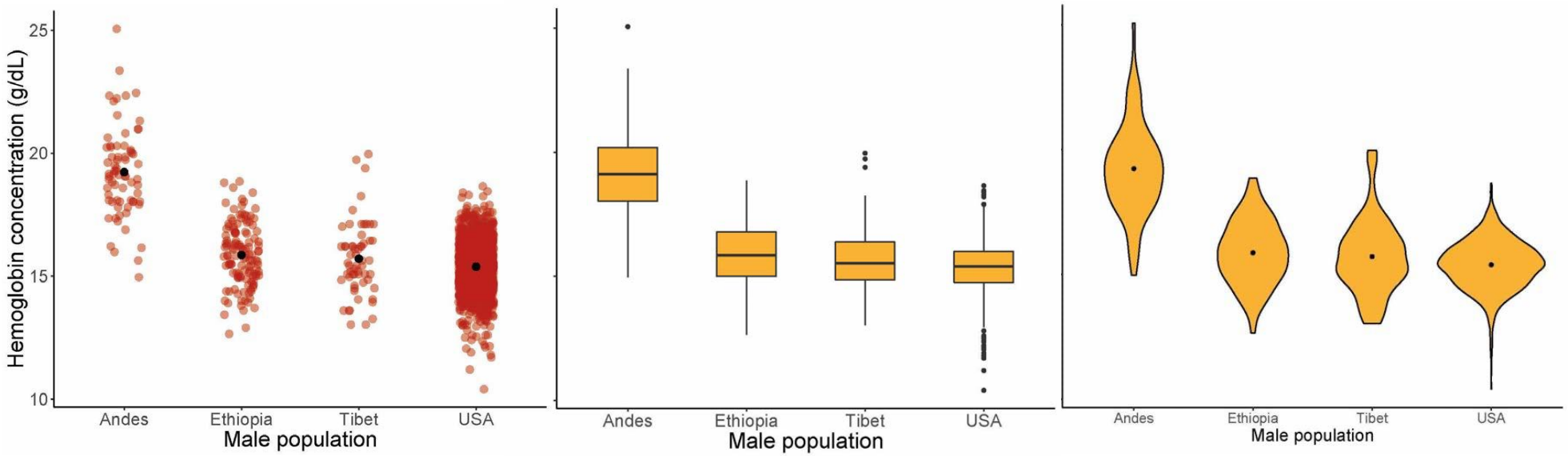


与美国海平面对照组相比，生活在高海拔地区的男性血液中的血红蛋白浓度



(Whitlock & Schluter 2020)

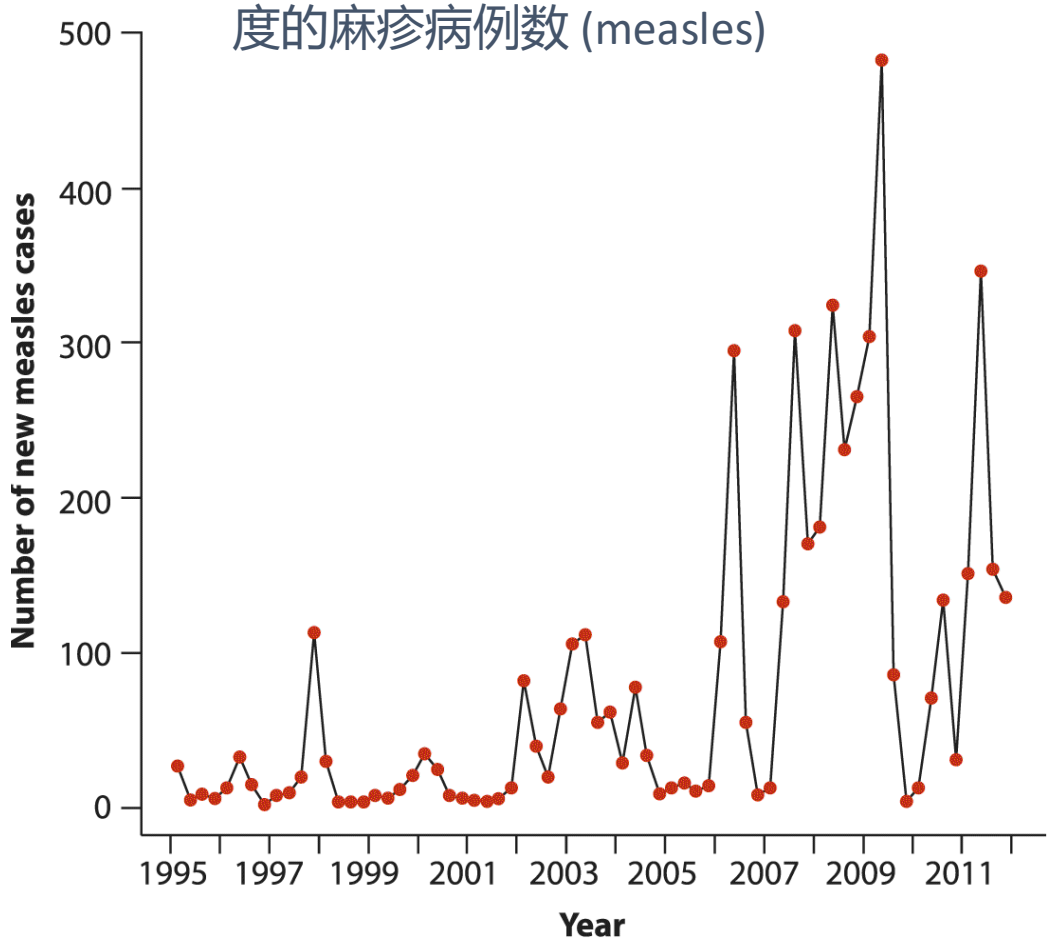
- 显示数值变量和分类变量之间的关联
  - 带状图 vs 箱形图 vs 小提琴图 (violin plot)
    - 箱形图展示了中位数、 第一和第三四分位数、数值范围和极端值
    - 小提琴图估算了各组的概率密度（kernel smoothing）
    - 非零基线通常是可以的
  - 哪种图形更有效?



(Whitlock & Schluter 2020)

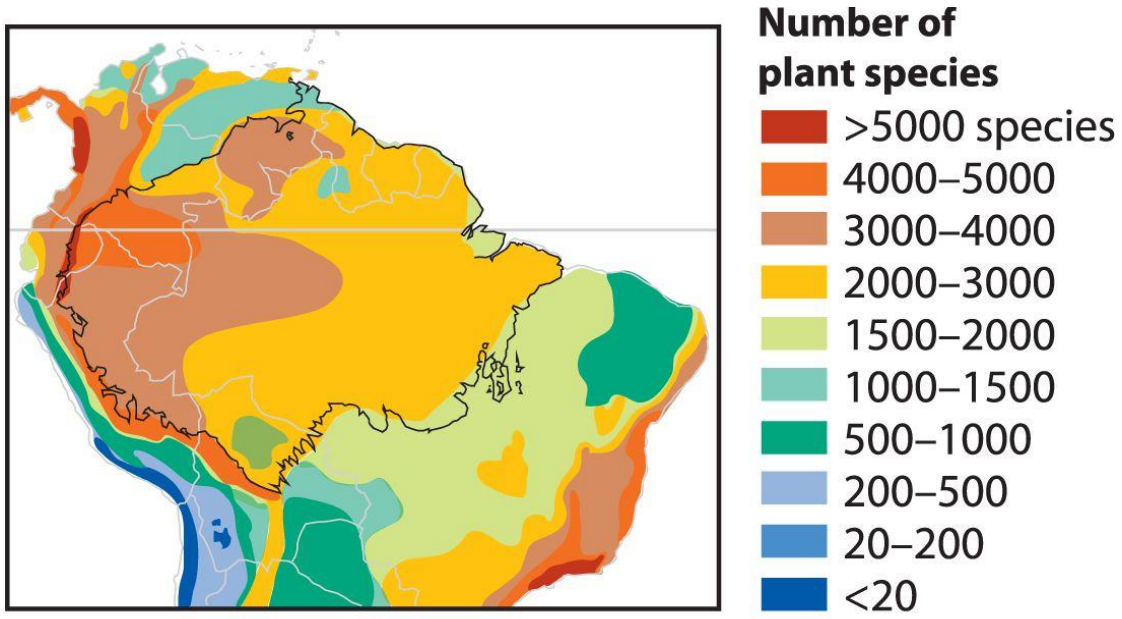
- 显示时间趋势
  - 折线图 Line graph
    - 用线段连接的点来显示随时间变化的趋势
    - 线段的陡峭程度反映了数字变化的速度。

995 年至 2011 年期间每个季度的麻疹病例数 (measles)



(Whitlock & Schluter 2020)

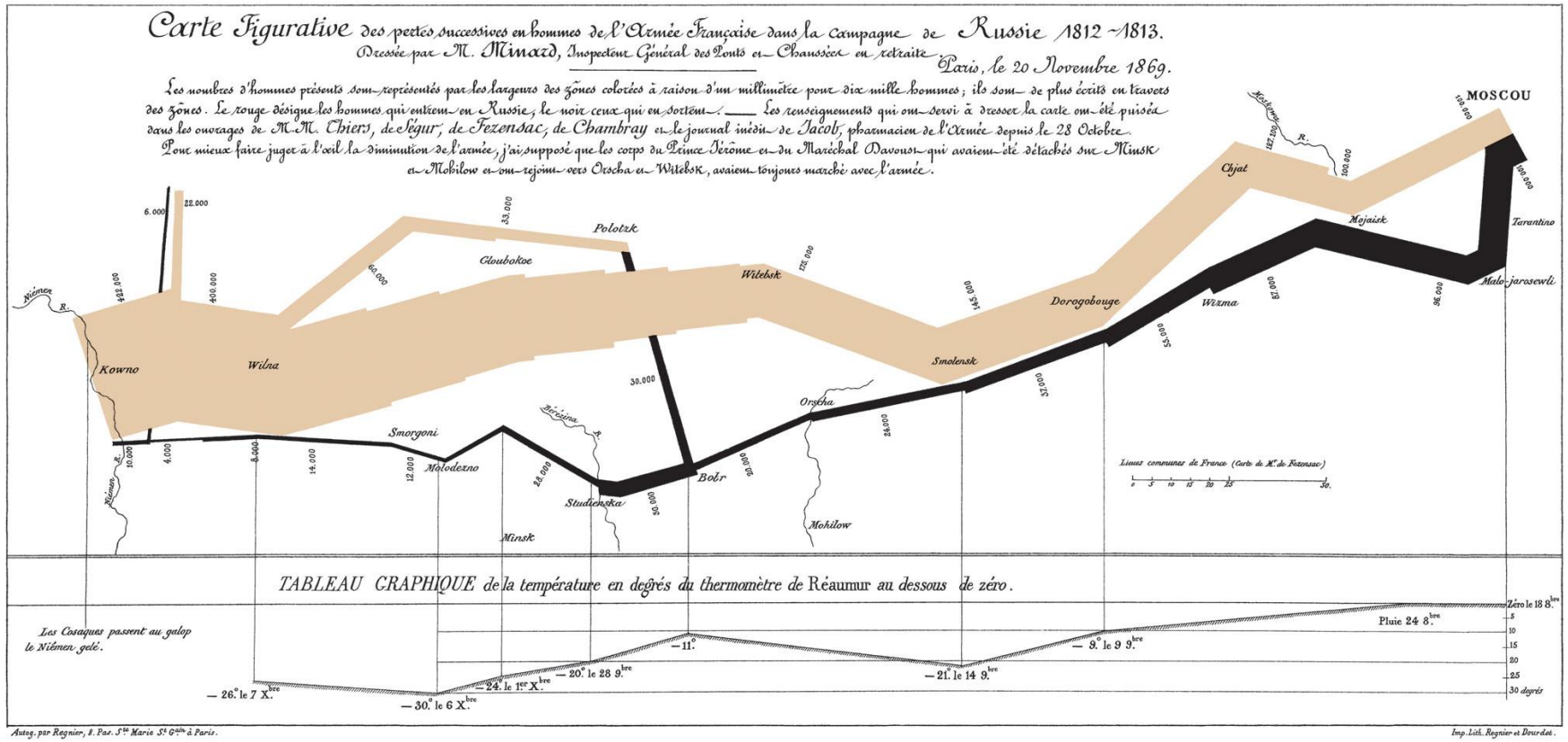
- 显示空间格局
  - 地图 Map
    - 使用颜色梯度显示地理上多个位置的数值响应变量
    - 相当于折线图的空间图
    - 解释变量 = 位置
      - 空间网格或政治边界
    - 可用于显示任何二维或三维物体（如大脑或人体）表面不同位置的测量值



南美洲北部植物物种数量，每个点的面积为 100 km×100 km.

(Whitlock & Schluter 2020)

- 显示空间格局
  - 地图 Map（可整合更多数据）
    - 拿破仑的俄罗斯远征



(Minard 1869)

4. 表格 Tables

- 表格也用于比较不同组之间的测量结果并揭示变量之间的关系
- 对于某些类型的数据，表格可能是传达结果的最佳方式

- 制作表格时，要让读者产生 “哦！” 而不是 “咦？” 的感觉。
  - “Oh!” , not “Huh?”

- 将含大量数值的表格放入附录（Appendix or Supplement）

4. 表格

- 难以直观看出 *F* 和生存（survival）之间的关系
- 行间距不均的割裂感
- 空白太多
- 小数点后数字太多

Table 2.5-1 Inbreeding coefficient (*F*) of Spanish Habsburg kings and queens and survival of their progeny.

King/Queen	*F*	Pregnancies	Miscarriages & stillbirths	Neonatal deaths	Later deaths	Survivors to age 10	Survival (total)	Survival (postnatal)
Ferdinand of Aragon								
Elizabeth of Castile	0.039	7	2	0	0	5	0.714	1.000
Philip I								
Joanna I	0.037	6	0	0	0	6	1.000	1.000
Charles I								
Isabella of Portugal	0.123	7	1	1	2	3	0.429	0.600
Philip II								
Elizabeth of Valois	0.008	4	1	1	0	2	0.500	1.000
Anna of Austria	0.218	6	1	0	4	1	0.167	0.200
Philip III								
Margaret of Austria	0.115	8	0	0	3	5	0.625	0.625
Philip IV								
Elizabeth of Bourbon	0.050	7	0	3	2	2	0.286	0.500
Mariana of Austria	0.254	6	0	1	3	2	0.333	0.400

Source: Data are from Alvarez et al. (2009).

4. 表格

- 改进表格
  - 将你最想让眼睛/大脑比较的数字垂直堆叠，不要有空隙；
  - 将希望显示关联的列放在相邻的位置，并对其中一列进行排序。

Table 2.5-2 Inbreeding coefficient (*F*) of Spanish kings and queens and survival of their progeny. These data are extracted and reorganized from Table 2.5-1.

King/Queen	*F*	Survival (postnatal)	Survival (total)	Number of pregnancies
Philip II/Elizabeth of Valois	0.01	1.00	0.50	4
Philip I/Joanna I	0.04	1.00	1.00	6
Ferdinand/Elizabeth of Castile	0.04	1.00	0.71	7
Philip IV/Elizabeth of Bourbon	0.05	0.50	0.29	7
Philip III/Margaret of Austria	0.12	0.63	0.63	8
Charles I/Isabella of Portugal	0.12	0.60	0.43	7
Philip II/Anna of Austria	0.22	0.20	0.17	6
Philip IV/Mariana of Austria	0.25	0.40	0.33	6

5. 总结 Summary

- 图形显示必须清晰、真实、高效；

- 四原则：要努力显示原始数据，并使数据中的模式一目了然，真实地表示量级，并清晰地绘制图形元素（与表格的规则相同）；

- 条形图和直方图是显示分类变量和数值变量频数分布的推荐图形；

5. 总结 Summary

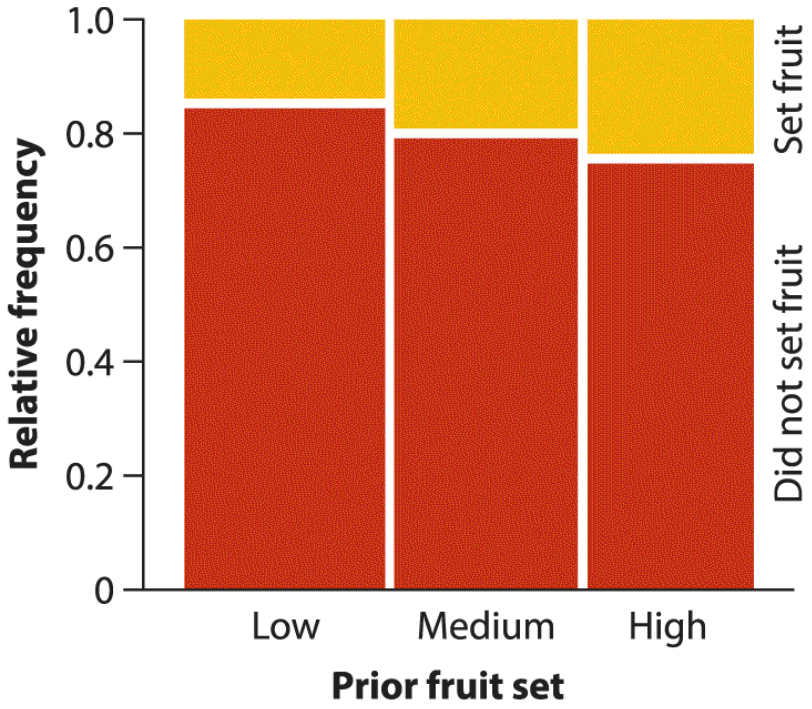
- 用于显示变量之间的关联和组间差异的推荐图形包括：

Types of data 数据类型	**Graphical method 图像类型**
One categorical variable 单个分类变量	Bar plot 条形图
One numerical variable 单个数值变量	Histogram 直方图
Two numerical variables 两个数值变量	Scatter plot 散点图
Two categorical variables 两个分类变量	Grouped bar graph 分组条形图 Mosaic plot 马赛克图
One numerical variable and one categorical variable 单个数值变量 ~ 单个分类变量	Strip chart 带状图 Boxplot 箱形图 Violin plot 小提琴图 Multiple histograms 多组直方图

6. Discussion

- 1. Each of the following graphs illustrates an association between two variables. For each graph, identify
  - (1) the type of graph,
  - (2) the explanatory and response variables,
  - (3) the type of data for each variable.
    - whether numerical or categorical

a. Observed fruiting of individual plants in a population of Campanula americana (美国风铃草) according to the number of fruits produced previously (Richardson and Stephenson 1991):

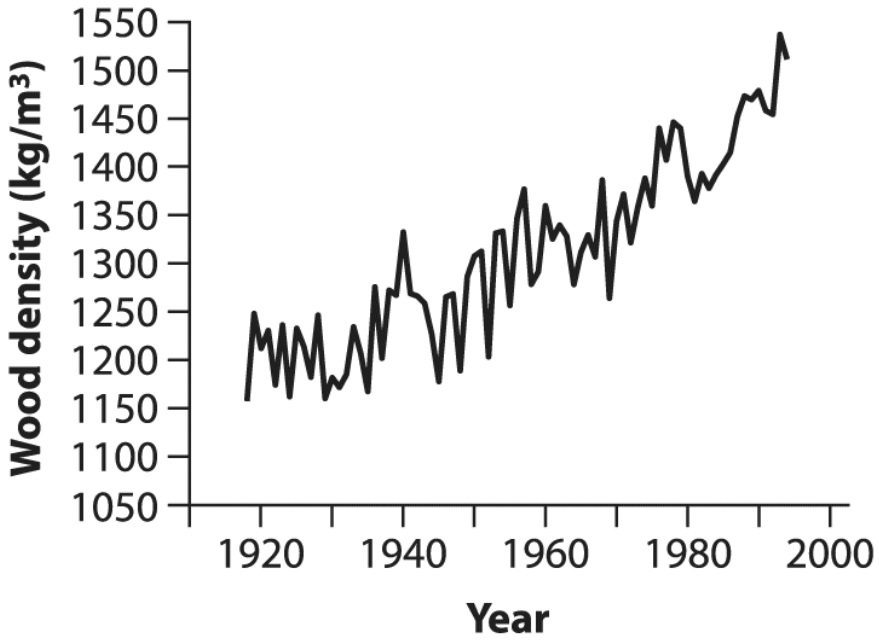


Whitlock & Schluter, *The Analysis of Biological Data*, 3e

6. Discussion

- 1. Each of the following graphs illustrates an association between two variables. For each graph, identify
  (1) the type of graph,
  (2) the explanatory and response variables,
  (3) the type of data for each variable.
    - whether numerical or categorical

b. The maximum density of wood produced at the end of the growing season in white spruce trees (白云杉) in Alaska in different years (data from Barber et al. 2000):



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

6. Discussion

- 2. Draw scatter plots for invented data that illustrate the following patterns:

  a. Two numerical variables that are positively associated

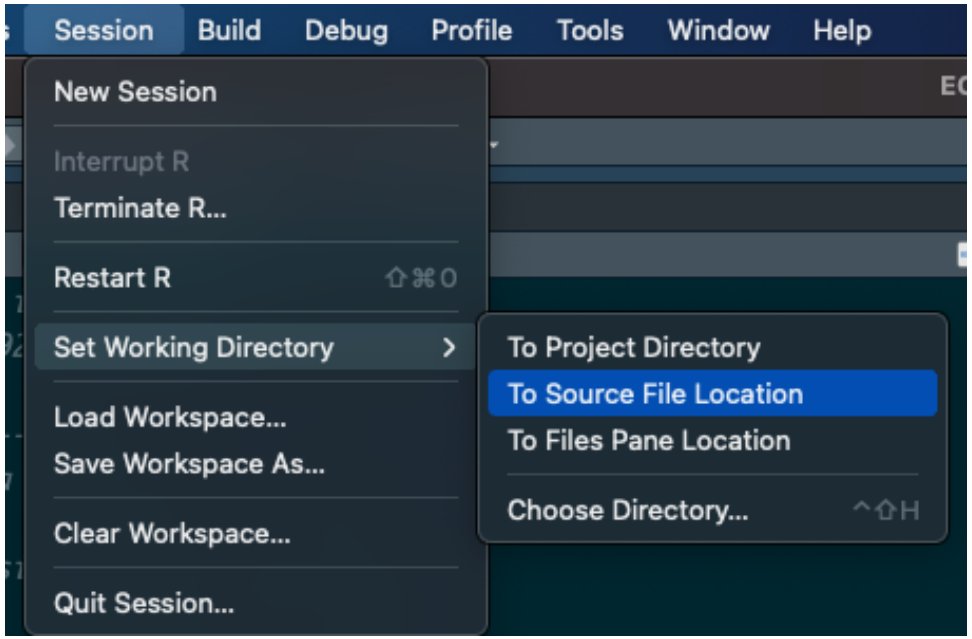
  b. Two numerical variables that are negatively associated

  c. Two numerical variables whose relationship is nonlinear

  - which type of graph to use?

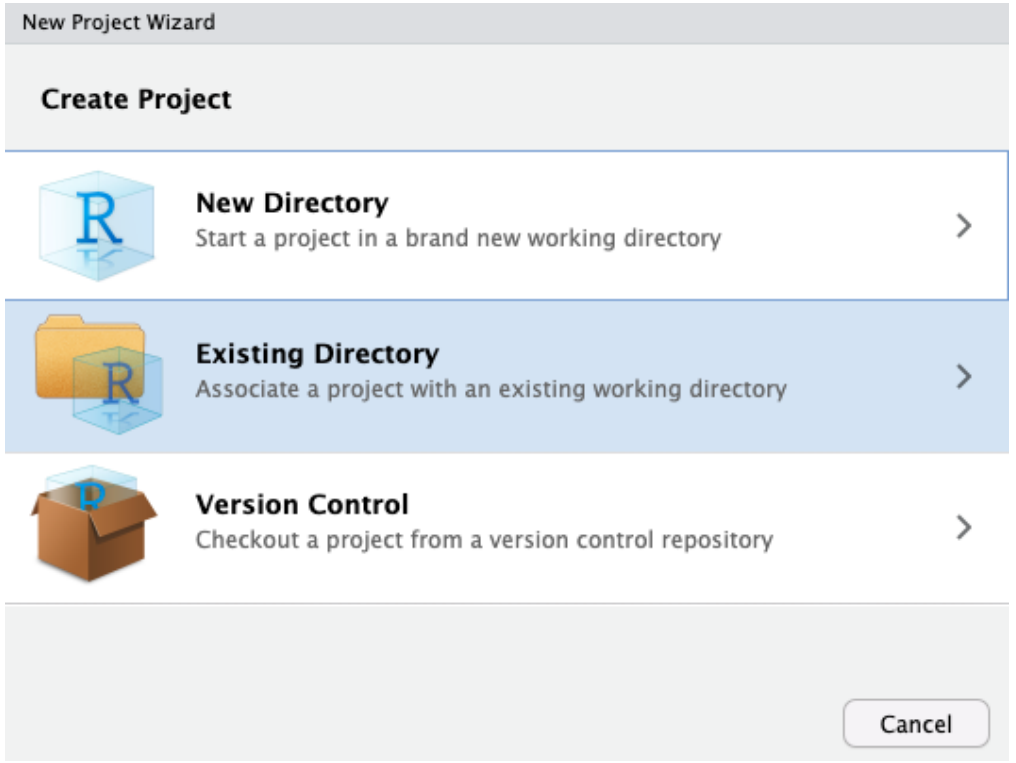
R Lab

- 如何设置工作路径 （working directory）以方便读取及存储文件
- 方法1：不设置，每次设定读取文件的路径
  - read.csv(file.choose())
  - read.csv(“path/to/your/file”)
- 方法2：每次开启Rstudio都手动设置路径
  - setwd(“path/to/your/directory”)
- 方法3：设置为R源代码文件所在路径
  - 在Rstudio右下角视图打开 “Files”中的R代码文件后
  - 工具栏：Session --> Set Working Directory
  - 之后选择 “To Source File Location”



R Lab

- 如何设置工作路径 （working directory）以方便读取及存储文件
- 方法4： Rstudio中新建并设置project
  - 选择Existing Directory
  - 浏览并选择读存文件的路径
    - 例如： “R-Lab”
  - 创建project
  - 之后可以在Rstudio右下角视图打开“Files”中的R代码文件
  - 每次开启Rstuido后在右上角选择对应的project



R Lab

- 文件管理的路径设置（文件夹设置）的建议
  - 方法1：代码文件和数据文件在同一个文件夹下
    - 适合Working Directory为Source File Location
      - read.csv(“chap02e2aDeathsFromTigers.csv”)
  - 方法2：代码和数据文件分属两个文件夹，上层文件夹为“Biostats”
    - 适合将Biostats设置为Project所在路径
    - 读取文件时需要输入数据文件夹名字
      - read.csv(“DataFiles/chap02e2aDeathsFromTigers.csv”)