

统计学B

第二章：数据总结

彭世喆
经济与管理学院
长沙理工大学

检验 数值数据

散点图

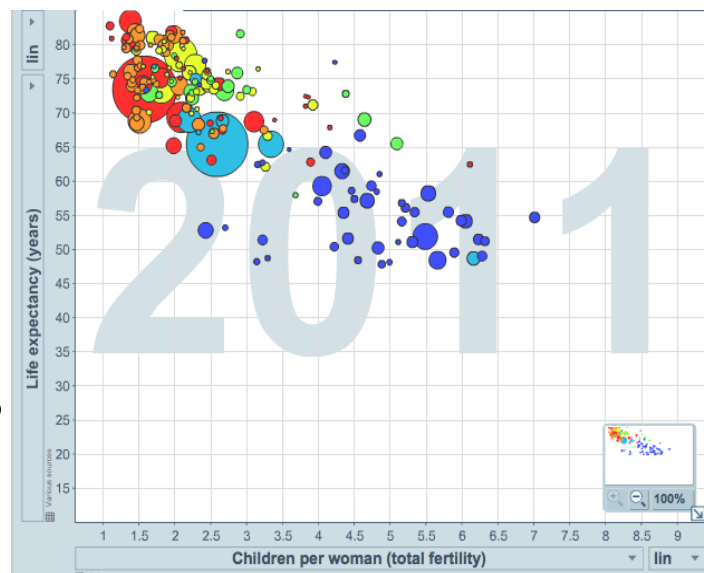
散点图有助于可视化两个数值变量之间的关系。

预期寿命与总生育率之间是相关还是独立？

它们呈现线性负相关：随着生育率上升，预期寿命下降。

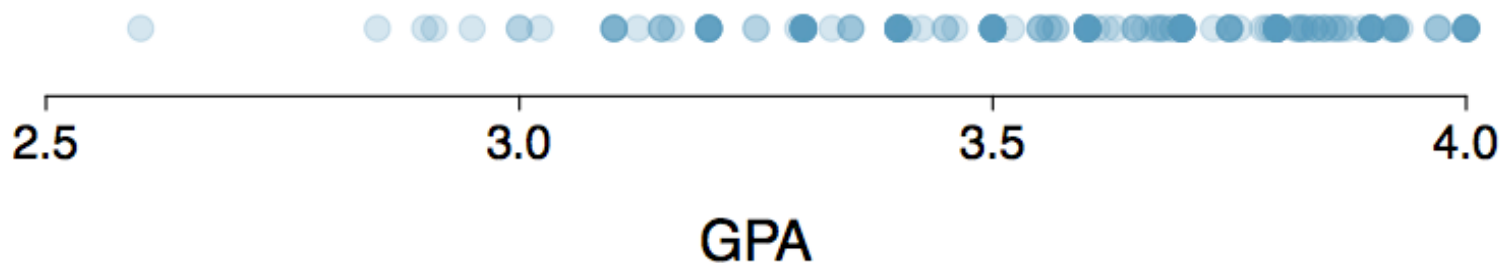
这种关系是历年来保持一致，还是发生了变化？

这种关系随着时间推移发生了变化。



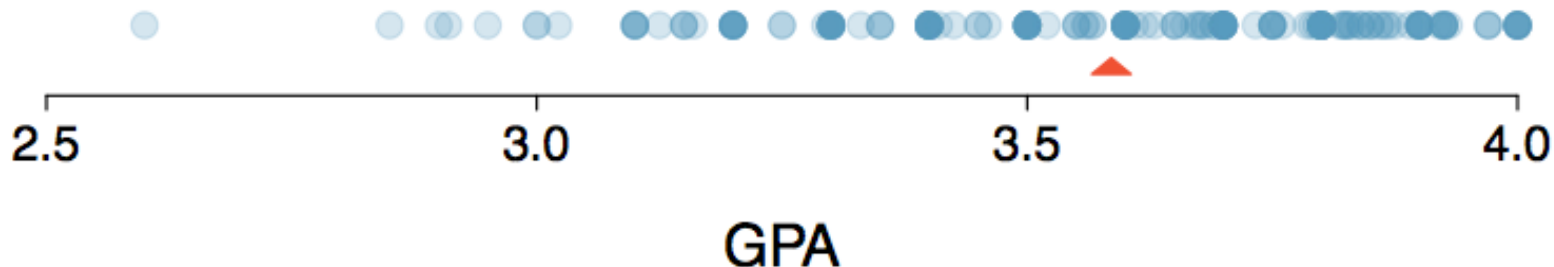
散点图（以点状图表示）

有助于可视化一个数值变量。颜色越深代表该区域观测值越多。



你如何描述该数据集中GPA的分布情况？请明确说明分布的中心、形状和离散度。

描点图与均值



均值也称平均值（上图中以三角形标记），是衡量数据分布中心的一种方法。

平均GPA为3.59。

均值

样本均值用 $\bar{x}$ 表示，计算公式为：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

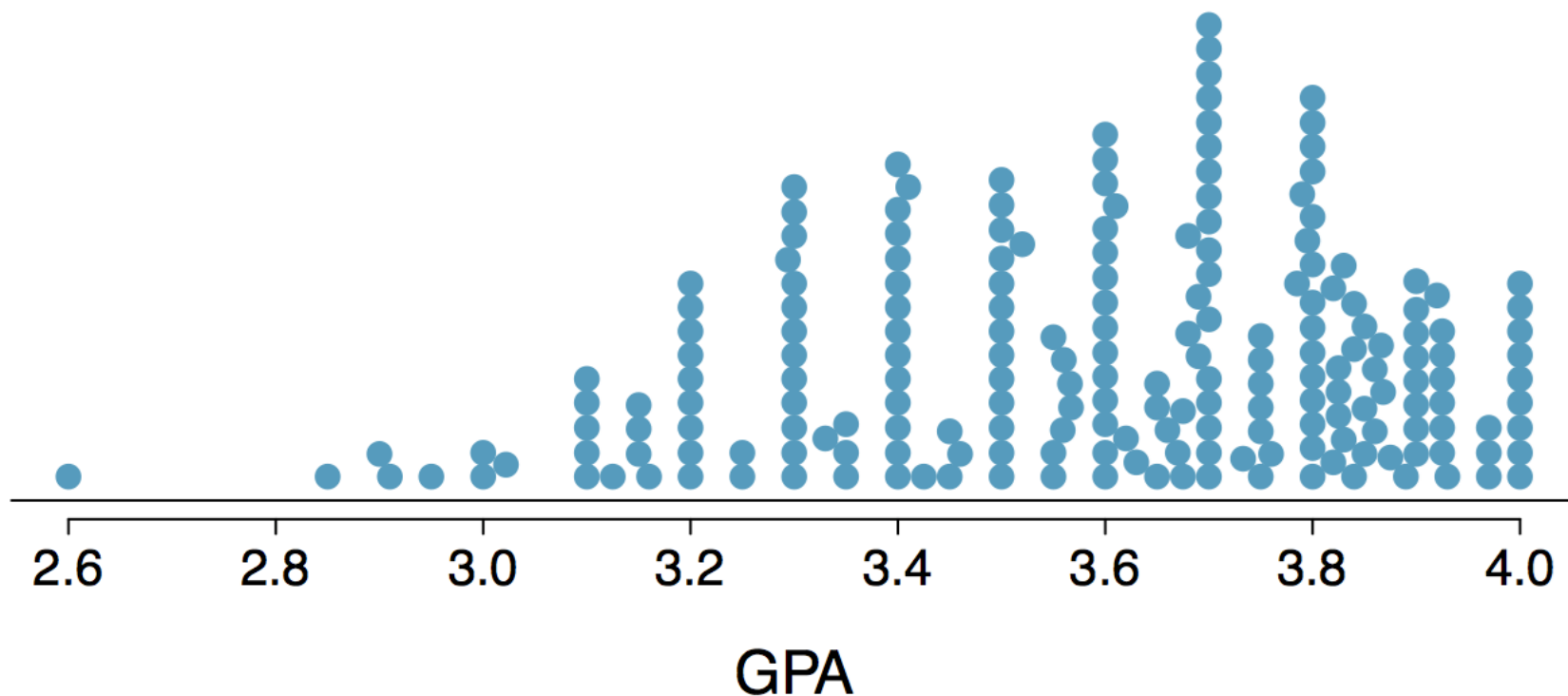
其中 $x_1, x_2, \dots, x_n$ 表示 n 个观测值。

总体均值的计算方法相同，但用 μ 表示。由于总体数据通常难以获取，因此通常无法计算 μ 。

样本均值是一个样本统计量，用作总体均值的点估计。这种估计可能并不完美，但如果样本良好（对总体具有代表性），通常是一种相当不错的估计。

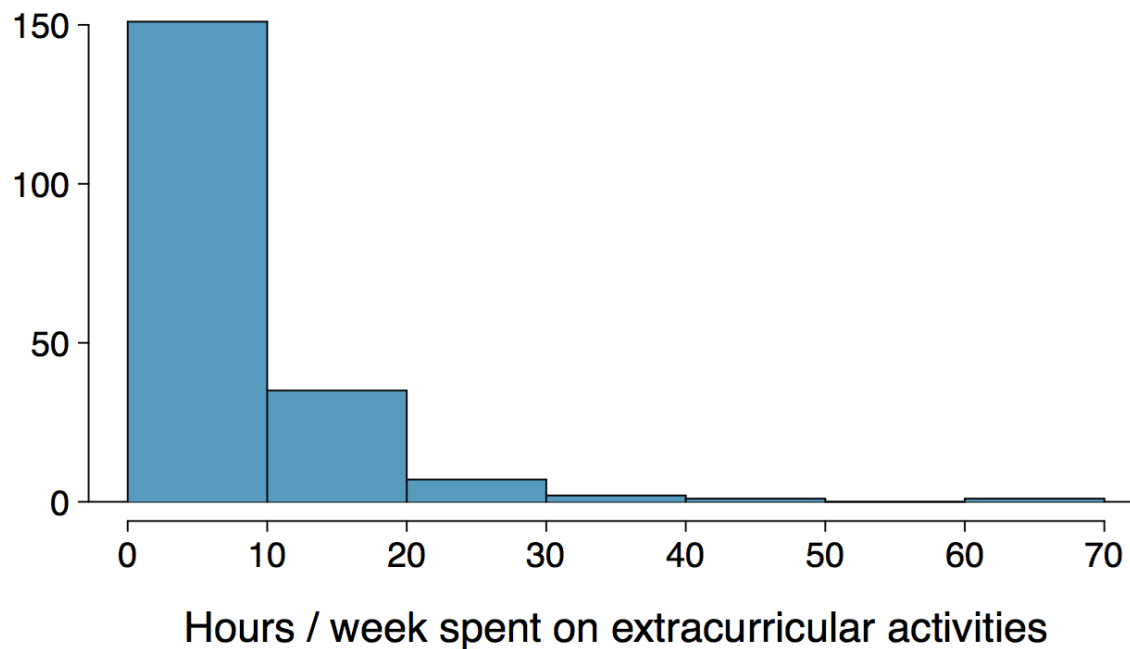
堆积描点图

更高的条形代表该区域观测值更多，便于判断分布的中心和形状。



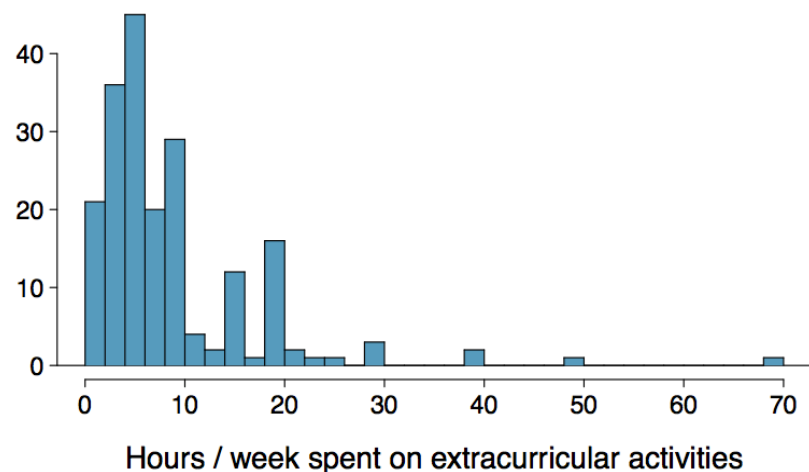
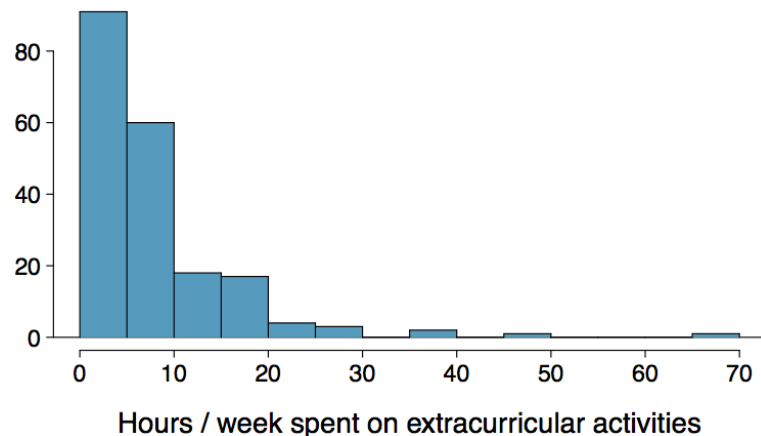
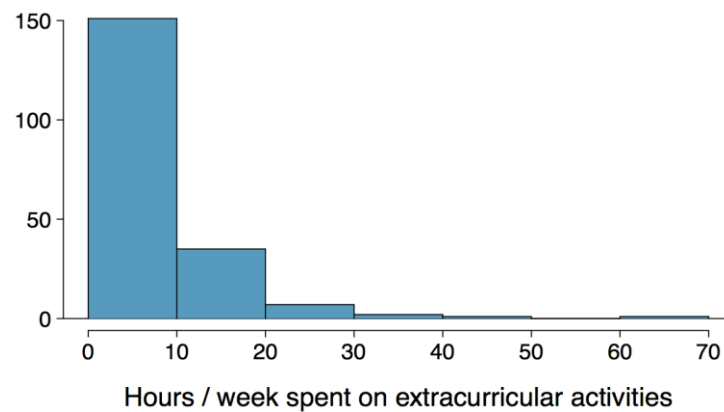
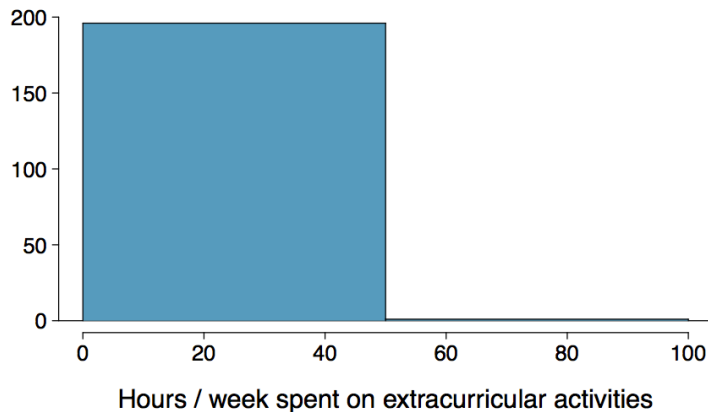
直方图——课外活动时间

- 直方图展现了数据密度。较高的条形代表数据分布较为集中的区域。
- 尤其适合描述数据分布的形状。
- 所选组距宽度可能会改变直方图呈现的信息。



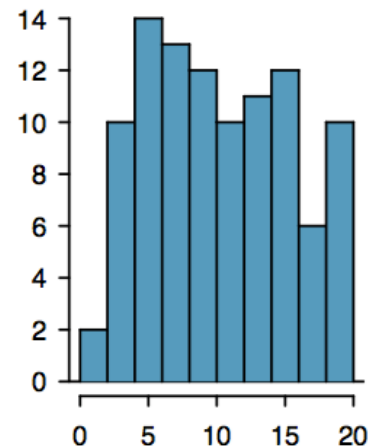
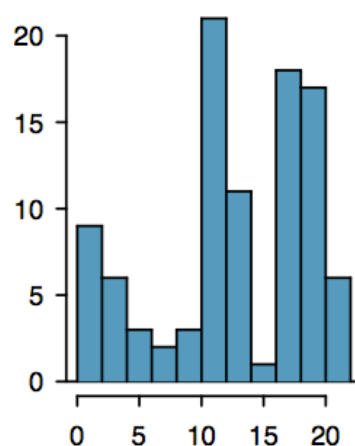
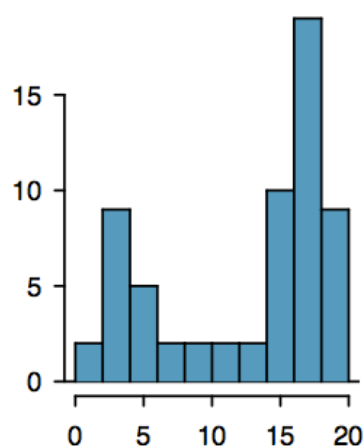
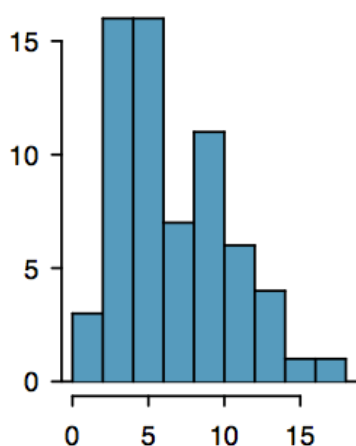
组距宽度

这些直方图中哪些有用？哪些透露了过多信息？哪些掩盖了太多信息？



分布形状：众数

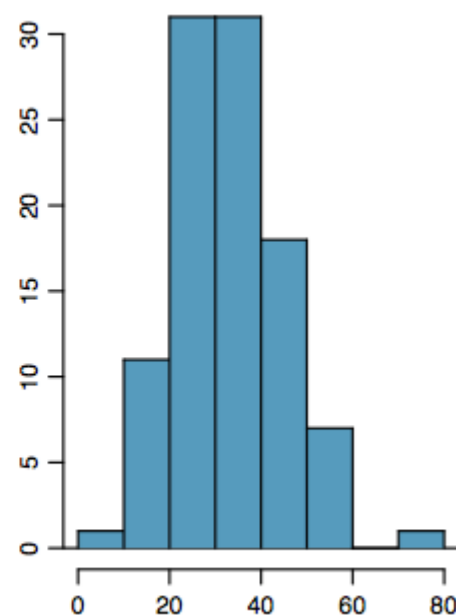
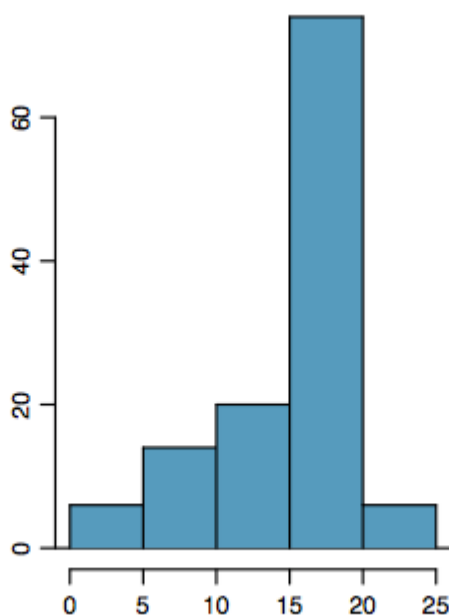
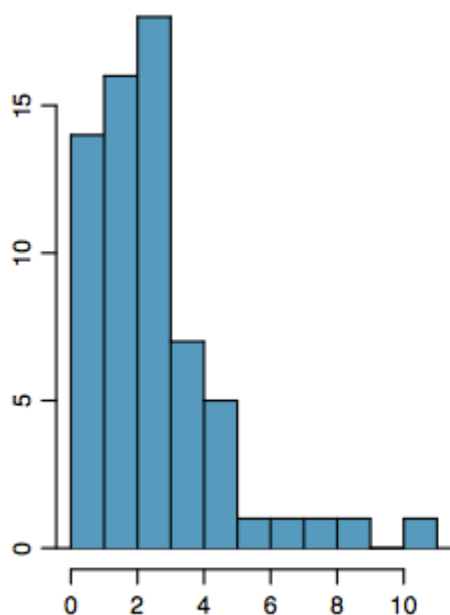
直方图是有一个显著峰值（单峰）、多个显著峰值（双峰/多峰），还是没有明显峰值（均匀分布）？



注意：判断众数时，请后退一步，想象在直方图上描绘一条平滑曲线——想象条形是木块，信手将一根软面条抛在上面，面条所成的形状即可视为平滑曲线。

分布形状：偏度

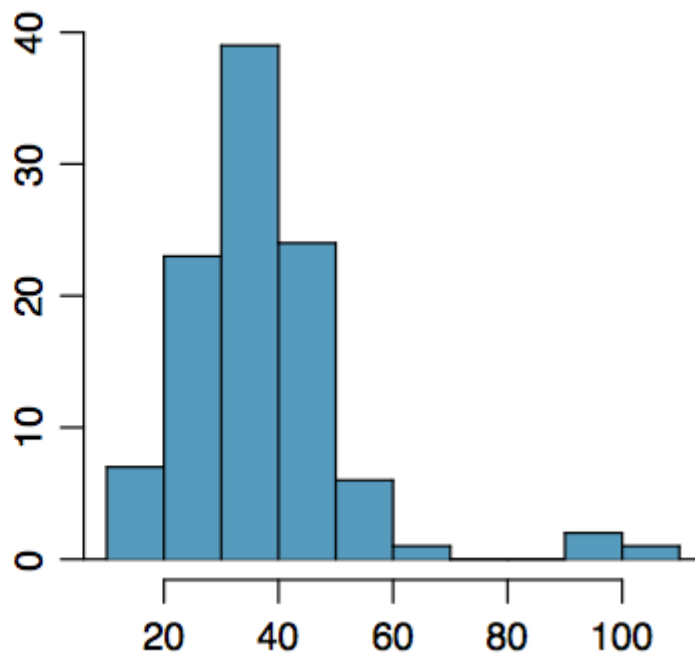
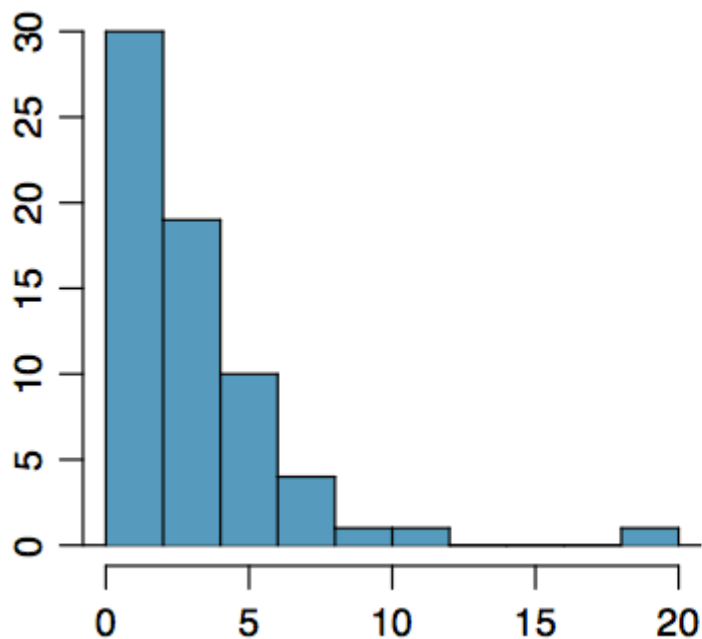
直方图是右偏、左偏还是对称的？



注意：直方图的偏斜方向以长尾所在侧来判定。

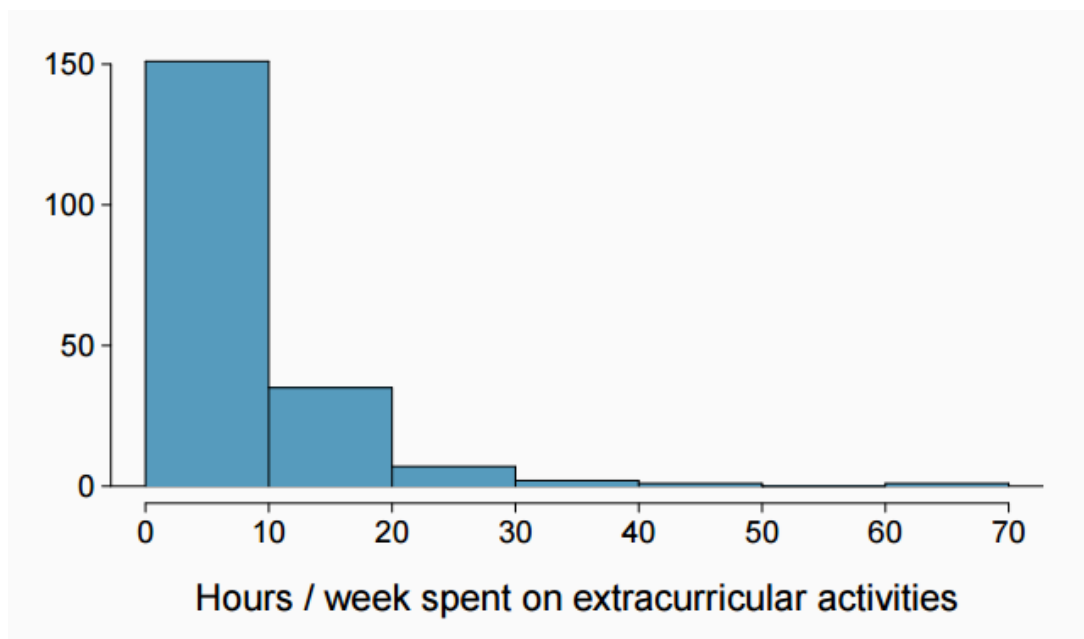
分布形状：异常观测值

是否存在任何异常观测值或潜在离群点？



课外活动

你如何描述学生每周课外活动小时数的分布形状？

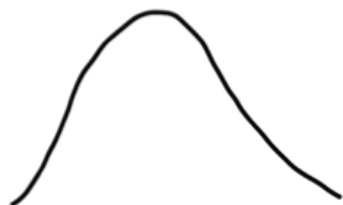


单峰且右偏，存在一个潜在异常观测值为60小时/周。

常见的分布形状

众数

unimodal



bimodal



multimodal



uniform



偏度

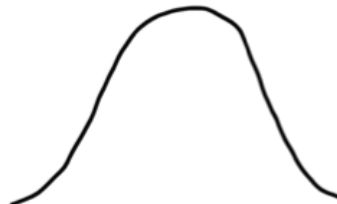
right skew



left skew



symmetric



众数在分布中表现为一个显著的峰值。

练习

你认为以下哪个变量可能呈均匀分布？

- (a) 成年女性的体重
- (b) 随机抽取的北卡罗来纳州人群的薪资
- (c) 房价
- (d) 同学的生日（按月中的日期计）

应用活动：分布形状

画出以下变量的预期分布草图：

- 穿孔数量
- 考试成绩
- 智商分数

提出一个简洁的方法（1-2句话），教人如何确定任意变量的预期分布。

你是典型数据吗？



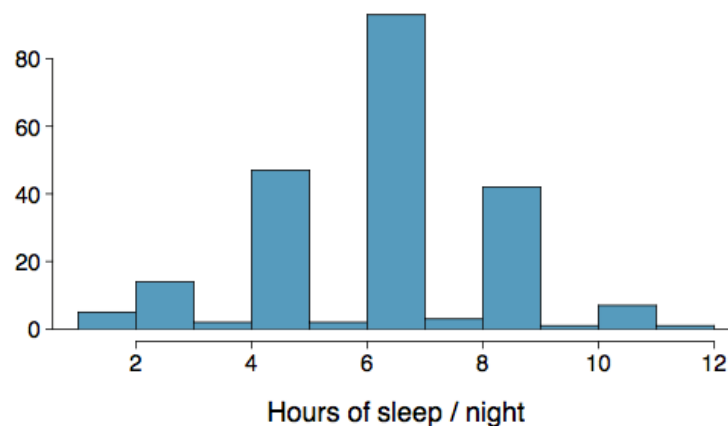
单纯依靠中心值对于传达分布的真实特征有多大用处？

方差

方差大致上就是各个数值离均差的平方的平均值。

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- 样本均值为 $\bar{x} = 6.71$ ，样本量为 $n = 217$ 。



- 学生每晚睡眠时数的方差可按如下方式计算：

$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \cdots + (7 - 6.71)^2}{217 - 1} = 4.11 \text{ hours}^2$$

方差（续）

为什么计算方差时我们要使用离差的平方？

- 为了消除负号，使得距离均值同样远的数据被平等对待。
 -
- 为了给较大的偏差赋予更高的权重。

标准差

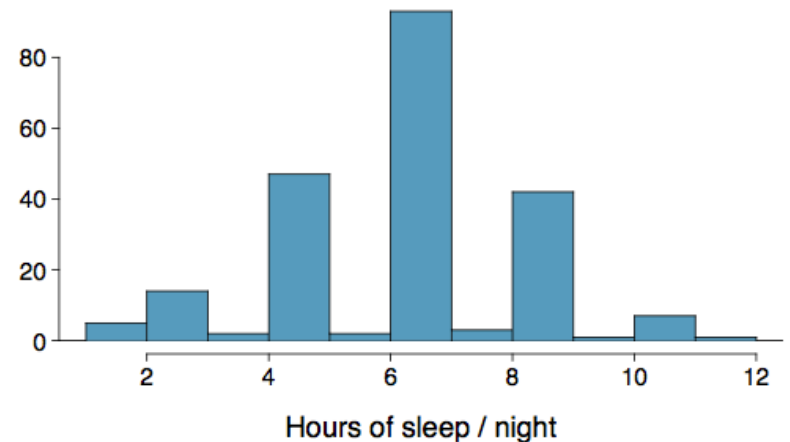
标准差是方差的平方根，大致描述了典型观测值与均值之间的距离。它的单位与数据的单位相同。

$$s = \sqrt{s^2}$$

- 学生每晚睡眠时数的标准差可按如下方式计算

:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



- 我们可以看出所有数据都在均值的三个标准差之内。

中位数

中位数是按升序排序后将数据一分为二的那个值。

0, 1, **2**, 3, 4

如果观测数量为偶数，则中位数是中间两个值的平均值。
。

$$0, 1, \underline{2}, \underline{3}, 4, 5 \rightarrow \frac{2 + 3}{2} = \mathbf{2.5}$$

由于中位数是数据的中点，50%的数值都在它之下，因此它也是第50百分位数。

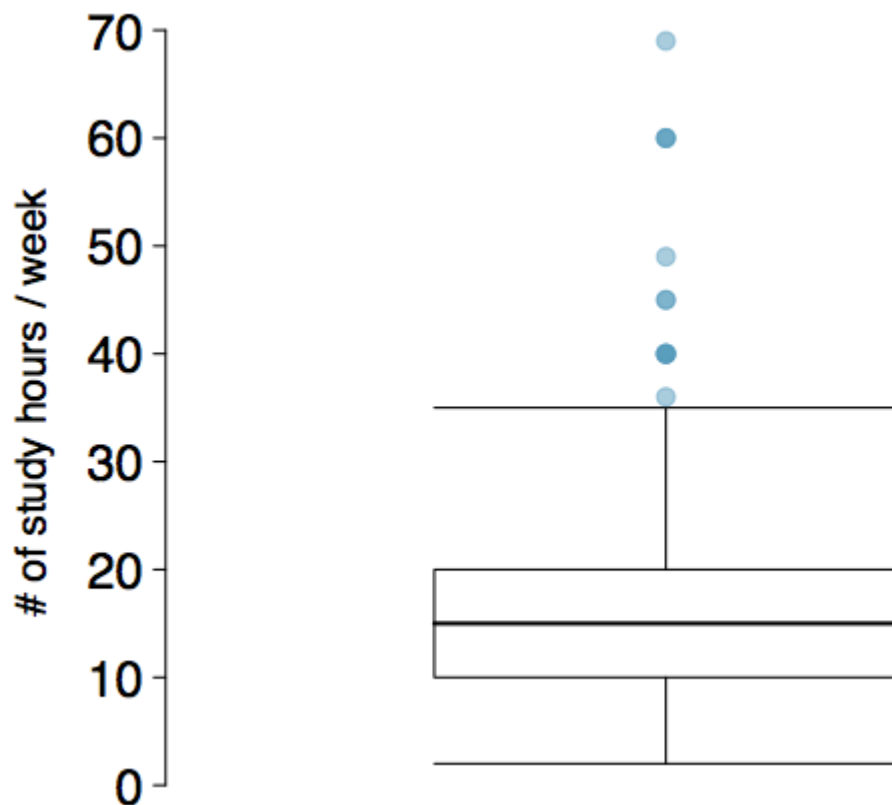
Q1、Q3和IQR

- 第25百分位数也称为第一四分位数（Q1）。
- 第50百分位数也称为中位数。
- 第75百分位数也称为第三四分位数（Q3）。
- Q1与Q3之间的区域是数据的中间50%。这两个数据跨度所形成的范围称为四分位距（Interquartile Range），或称IQR。

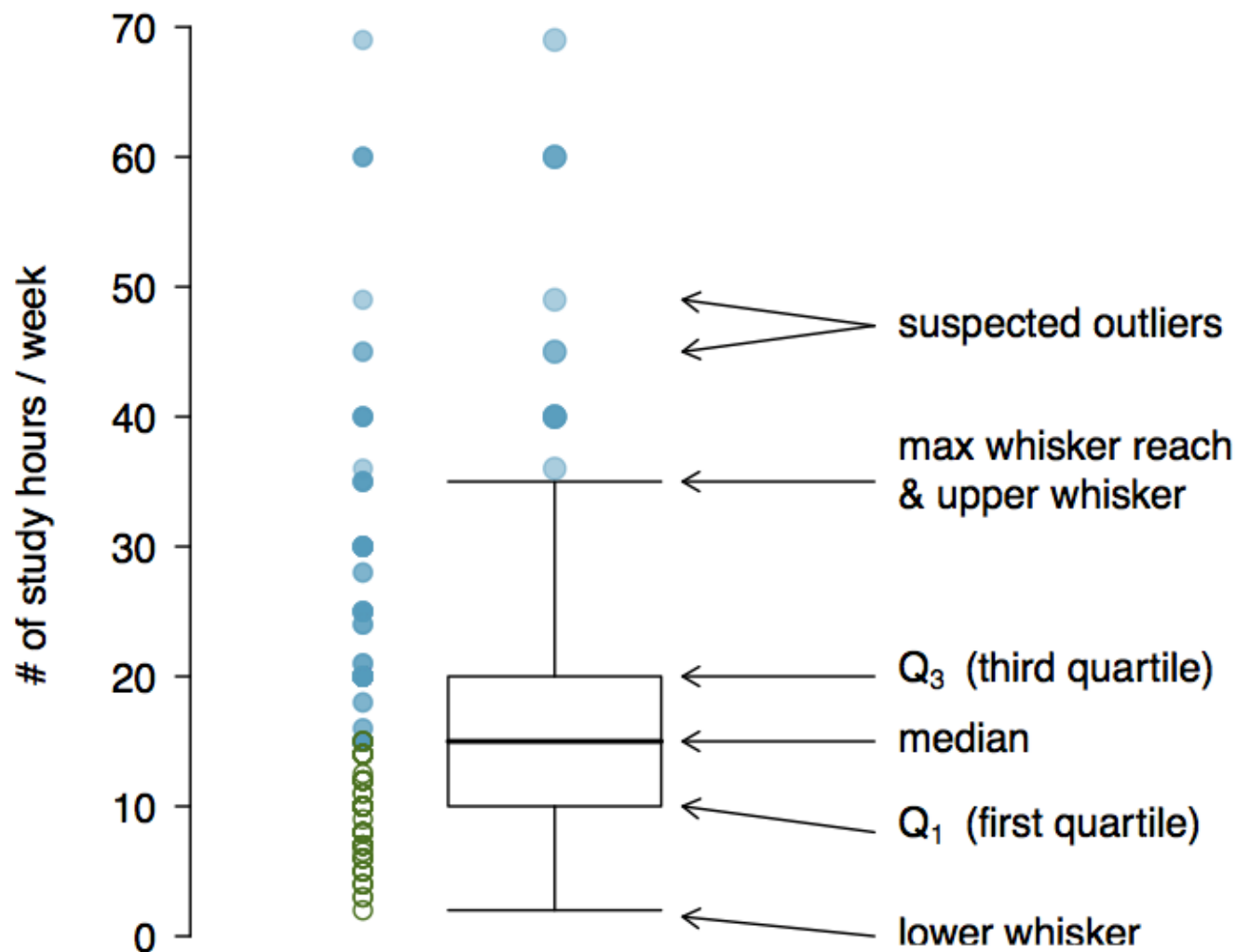
$$\text{IQR} = \text{Q3} - \text{Q1}$$

箱线图

箱线图通过样本统计量和极端值来总结数据集。箱线图箱子代表数据的中间50%，而箱子内的粗实线即为中位数。



箱线图的结构解析



须线与异常值

箱线图的须线最多可以向四分位数外延伸1.5倍IQR的距离。我们向中位数移动，直到达到最近观测值的数值。

$$\text{上须最大延伸长度} = Q3 + 1.5 \times \text{IQR}$$

$$\text{下须最大延伸长度} = Q1 - 1.5 \times \text{IQR}$$

$$\text{IQR: } 20 - 10 = 10$$

$$\text{上须最大延伸长度} = 20 + 1.5 \times 10 = 35$$

$$\text{下须最大延伸长度} = 10 - 1.5 \times 10 = -5$$

潜在异常值被定义为超出须线最大延伸范围的观测值，即相对于其余数据显得极端的观测值。

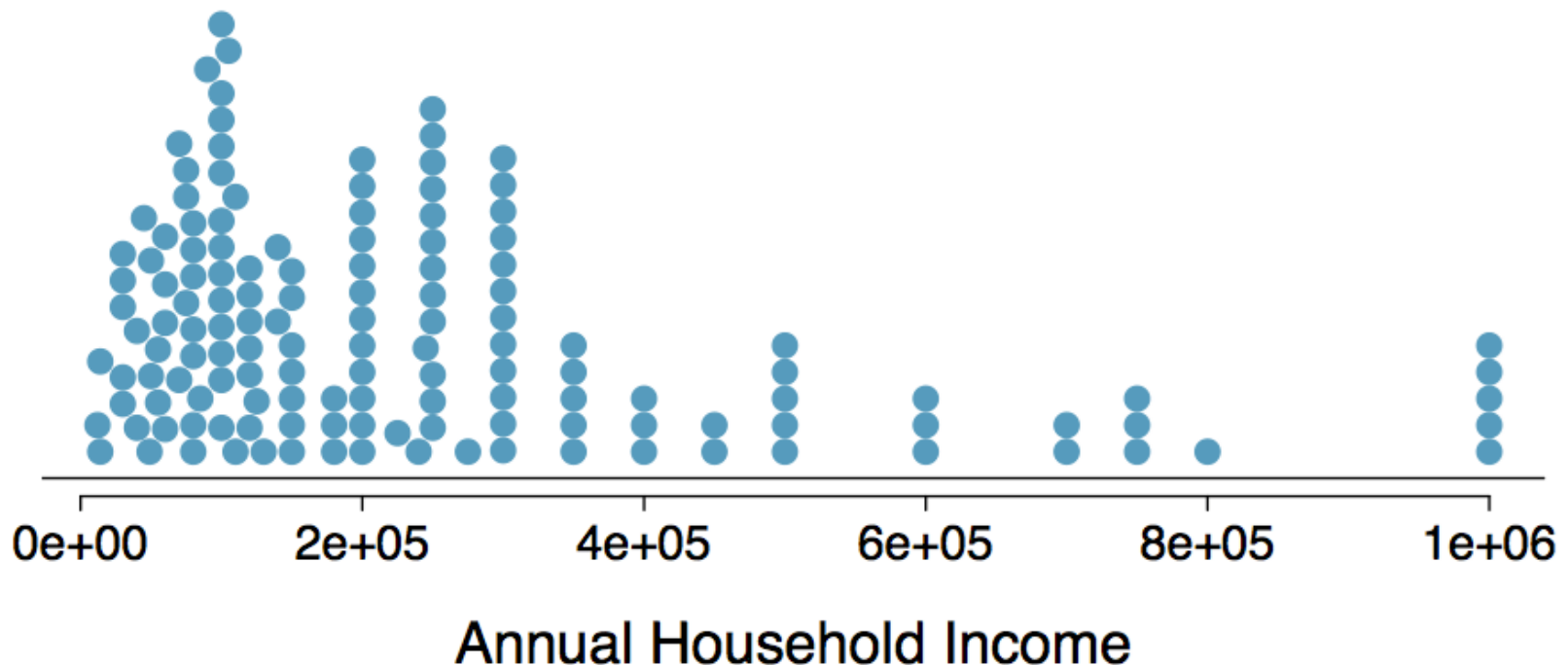
异常值（续）

寻找异常值为什么很重要？

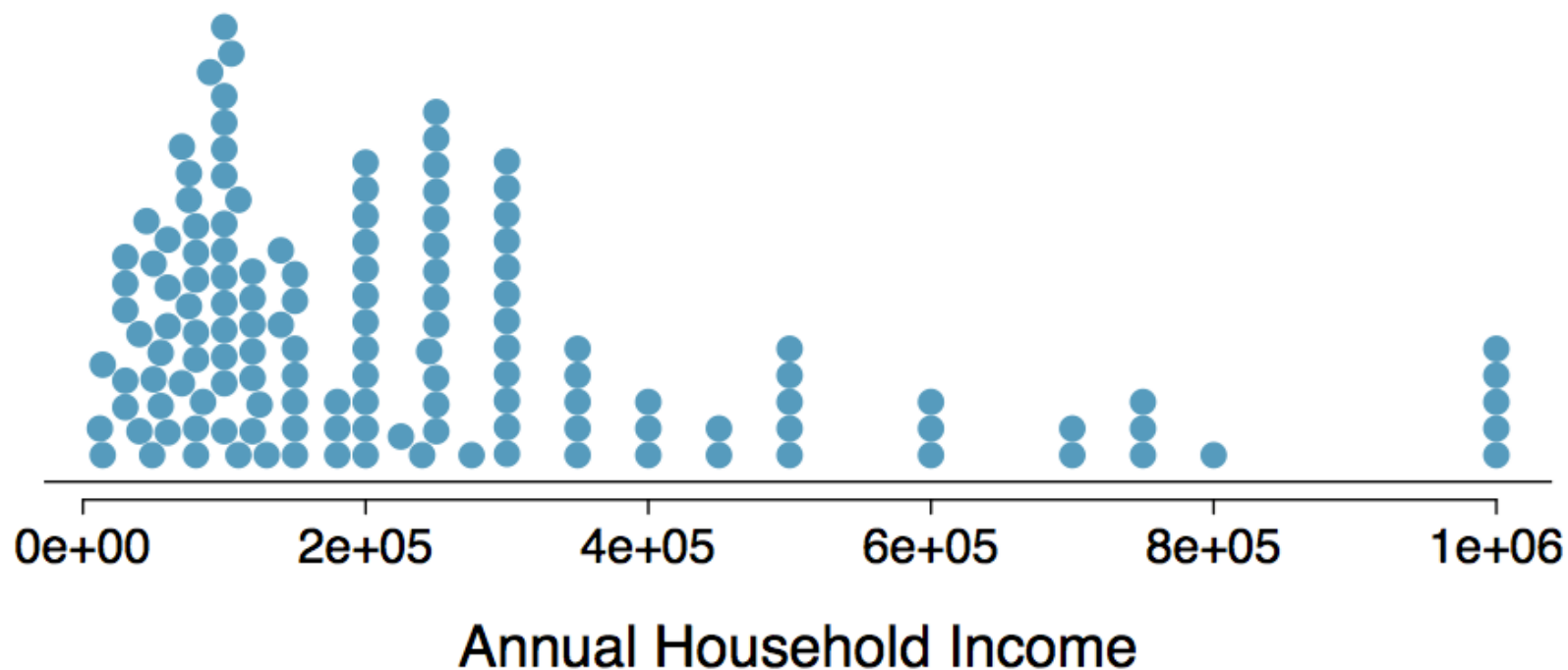
- 识别分布中的极端偏斜。
- 发现数据收集和录入的误差。
- 洞悉数据中有趣的特征。

极端观测值

如果将最大值替换为1000万美元，样本统计量（如均值、中位数、标准差及四分位距IQR）会受到怎样的影响？如果将最小值替换为1000万美元呢？



稳健统计量



scenario	robust		not robust	
	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to \$10 million	190K	200K	309K	853K
move smallest to \$10 million	200K	200K	316K	854K

稳健统计

相较均值与标准差，中位数和四分位距对于偏斜和异常值更具稳健性。因此，

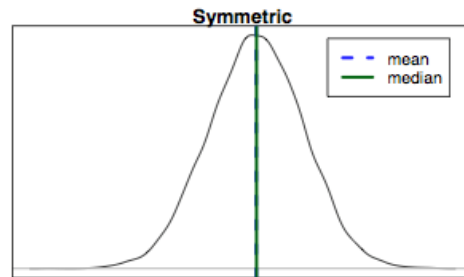
- 对于偏斜分布而言，使用中位数和四分位距来描述中心趋势和离散程度往往更为有益。
- 对于对称分布而言，使用均值与标准差来描述中心趋势和离散程度往往更为有益。

若你希望估算一名学生家庭的典型收入，你会对均值更感兴趣还是中位数？

中位数

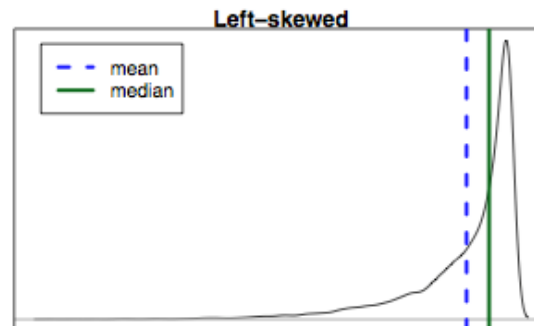
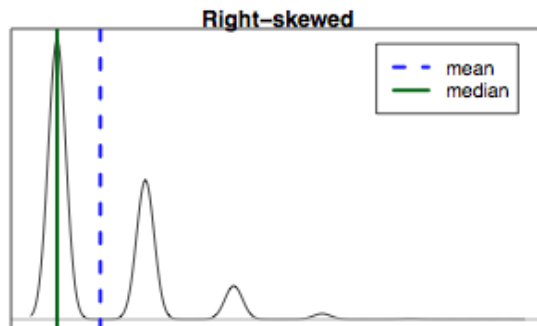
均值对比中位数

若分布对称，中心常被定义为均值：均值约等于中位数。



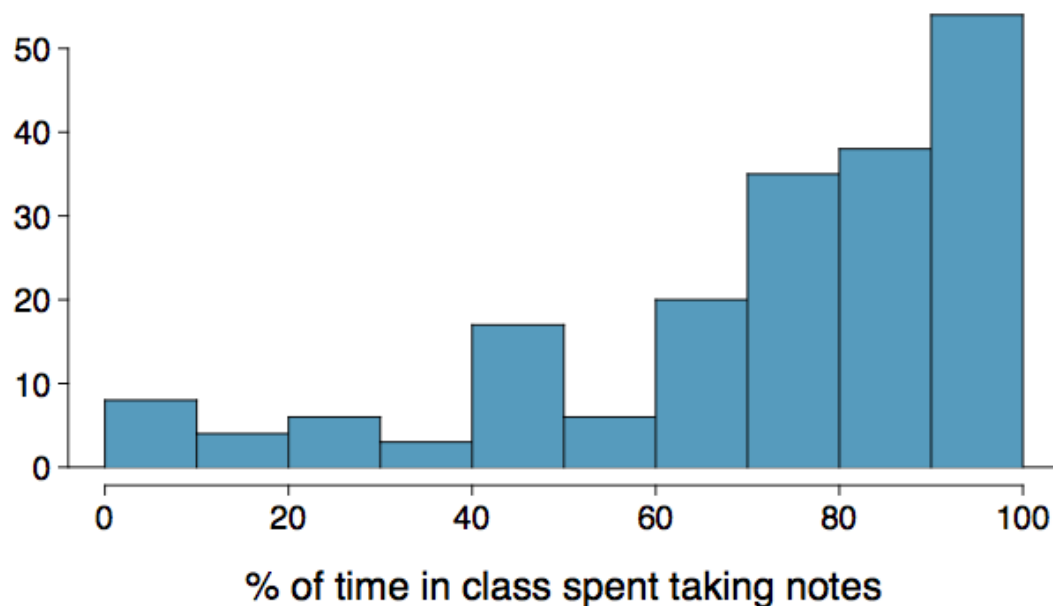
若分布偏斜或存在极端异常值，中心常被定义为中位数。

- 右偏观测值：均值 $>$ 中位数
- 左偏观测值：均值 $<$ 中位数



练习

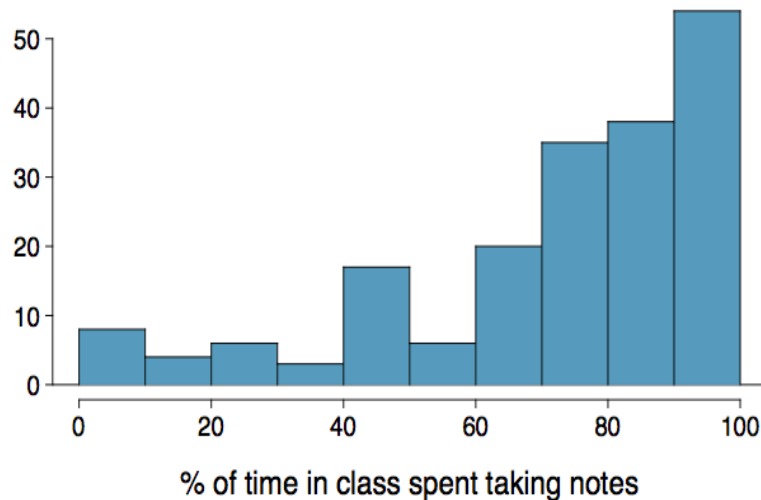
对于在课堂上实际用于记笔记的时间与使用Facebook、Twitter等软件时间的占比分布，以下哪种说法最有可能是正



- (a) 均值 > 中位数
- (b) 均值约等于中位数
- (c) 均值 < 中位数
- (d) 无法判断

练习

对于在课堂上实际用于记笔记的时间与使用Facebook、Twitter等软件时间的占比分布，以下哪种说法最有可能是正确的？



中位数：80%

均值：76%

(a) 均值 > 中位数

(b) 均值约等于中位数

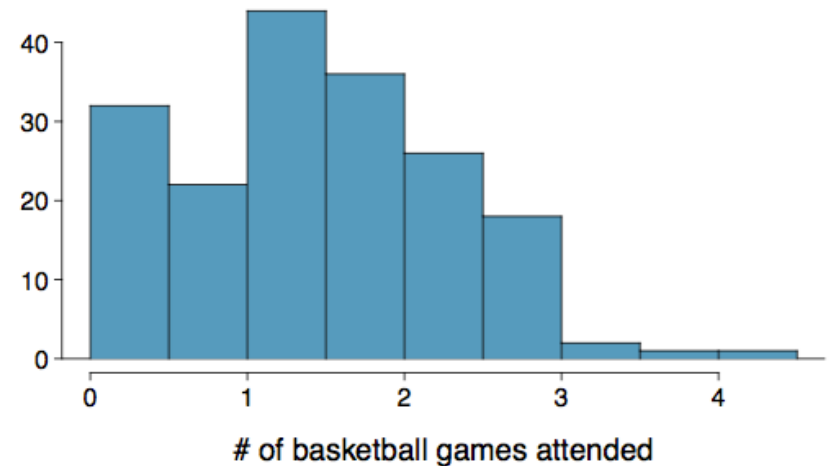
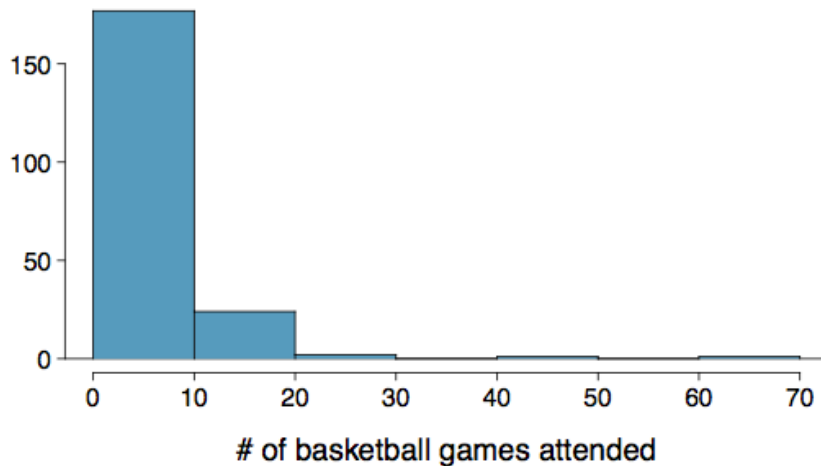
(c) 均值 < 中位数

(d) 无法判断

极度偏斜的数据

当数据极度偏斜时，对其进行变换可能使建模更简便。一种常用的变换为对数变换。

左侧直方图显示了学生观看篮球赛次数的分布。右侧直方图显示了观看次数的对数的分布。



变换的利弊

- 偏斜数据经过变换后更容易建模，因为经恰当变换后，异常值往往变得不那么突出。

比赛次数	70	50	25	...
------	----	----	----	-----

比赛次数	4.25	3.91	3.22	...
------	------	------	------	-----

- 然而，分析结论可能变得难以解读，因为对已测变量取对数通常没有实际意义。

你预期还有哪些变量可能极度偏斜？

薪资、房价等。

强度映射图

从2000年至2010年的人口变化中可看出哪些模式？

View More Maps | ▼

Change in population since 2000

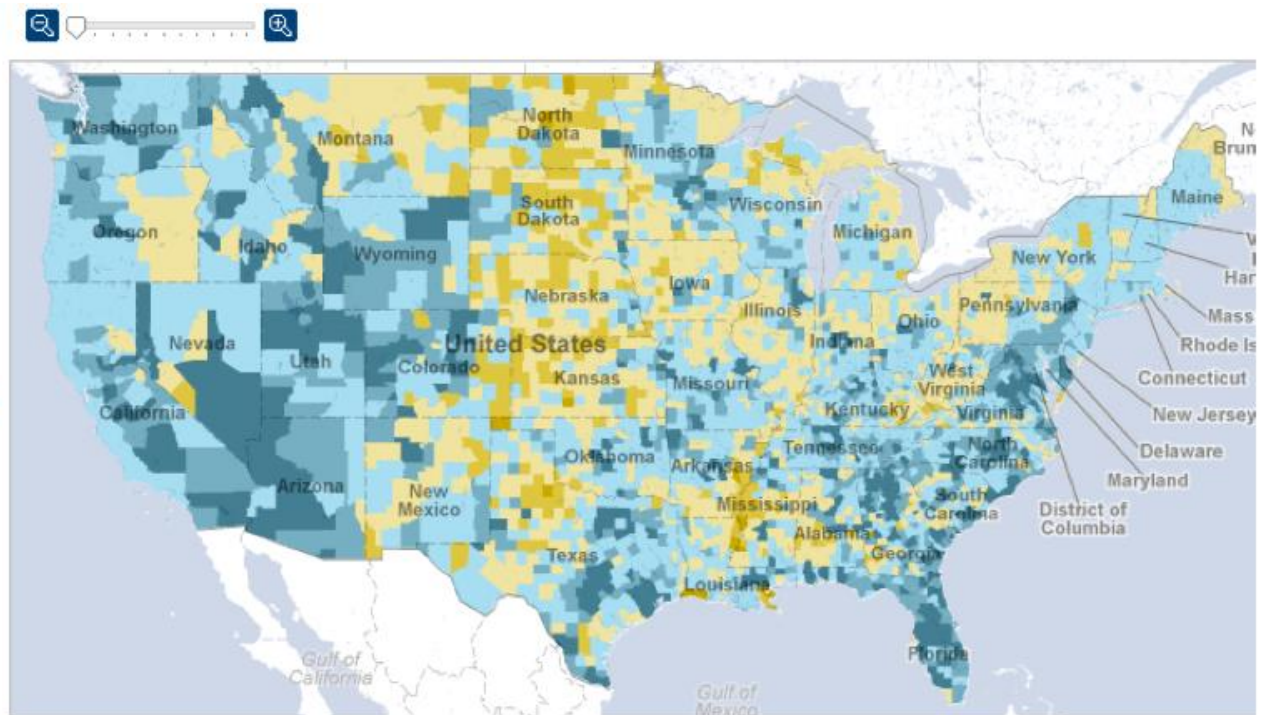


Zoom to a State | ▼

New Mexico

2010 POPULATION	CHANGE FROM 2000
2,059,179	+13.2%

RACE/ETHNICITY	SHARE OF POP.	CHANGE FROM 2000
Whites:	40%	+2%
Blacks:	2%	+16%
Hispanics:	46%	+25%



思考 分类数据

列联表

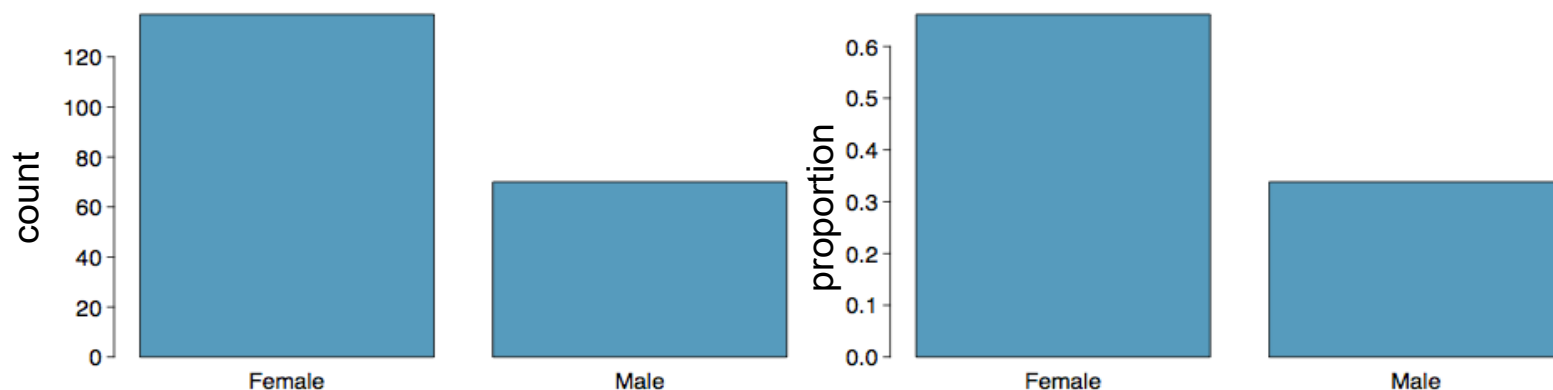
汇总两个分类变量之间数据的表格称为列联表。单变量的表格称为频数表。

下表呈现了学生的性别以及他们是否在大学期间寻找伴侣的分布情况。

		looking for spouse		Total
		No	Yes	
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

条形图

条状图是展示单一分类变量的一种常用方法。若展示的比例而非频数，则此类条形图被称为相对频数条形图。



条形图与直方图有哪些区别？

条形图用于展示分类变量的分布，而直方图用于数值变量。直方图的x轴是数值轴，因此条形顺序不可更改；但在条形图中，类别的排序可以任意调整（尽管某些排序比其他排序更具合理性，尤其对于有序变量而言）。

选择合适的比例

性别与学生在大学期间是否在寻找配偶之间是否存在关系？

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

为了回答这个问题，我们考察行比例：

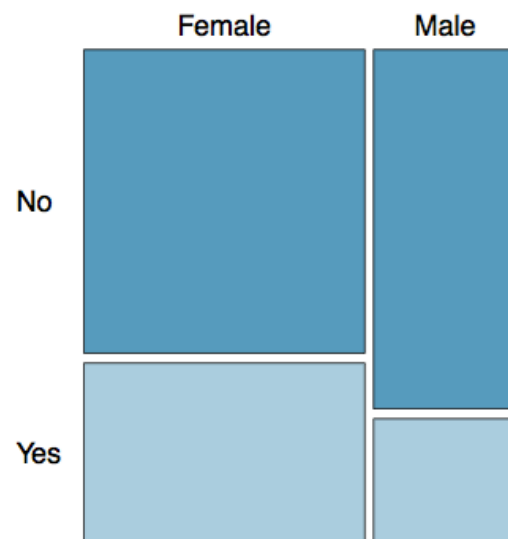
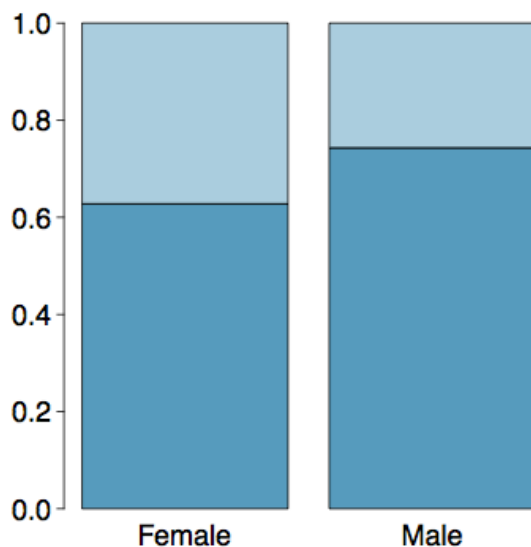
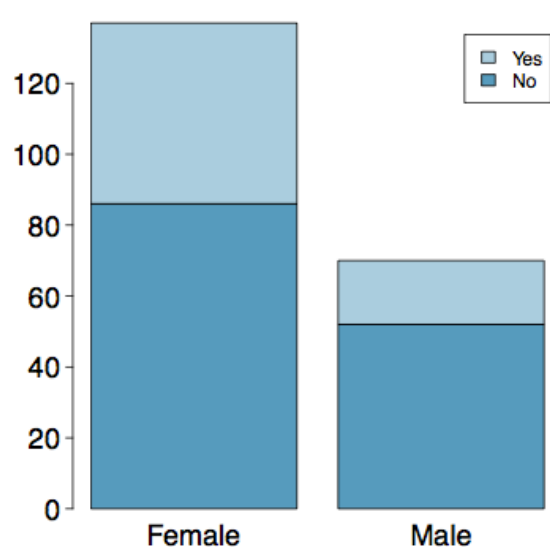
- 寻找配偶的女性百分比： $51 / 137 \sim 0.37$
- 寻找配偶的男性百分比： $18 / 70 \sim 0.26$

带有两个变量的条形图

- 堆叠条形图：列联表信息的图形展示，以计数形式呈现。
 -
- 并排条形图：将条形并排而非堆叠放置，展示相同的信息。
- 标准化堆叠条形图：列联表信息的图形展示，以比例形式呈现。

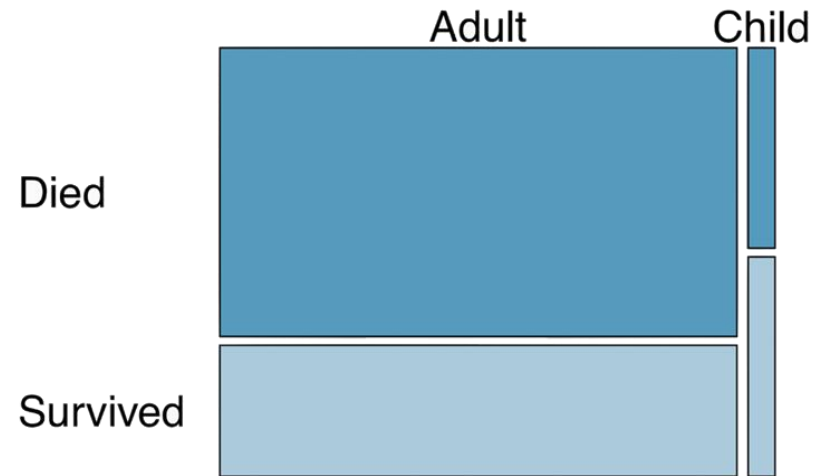
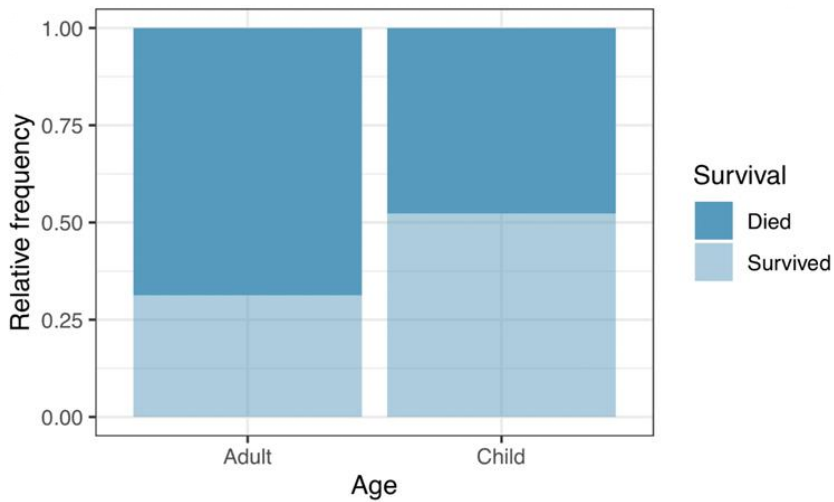
分段条形图和马赛克图

下面三种可视化呈现方式有什么不同？



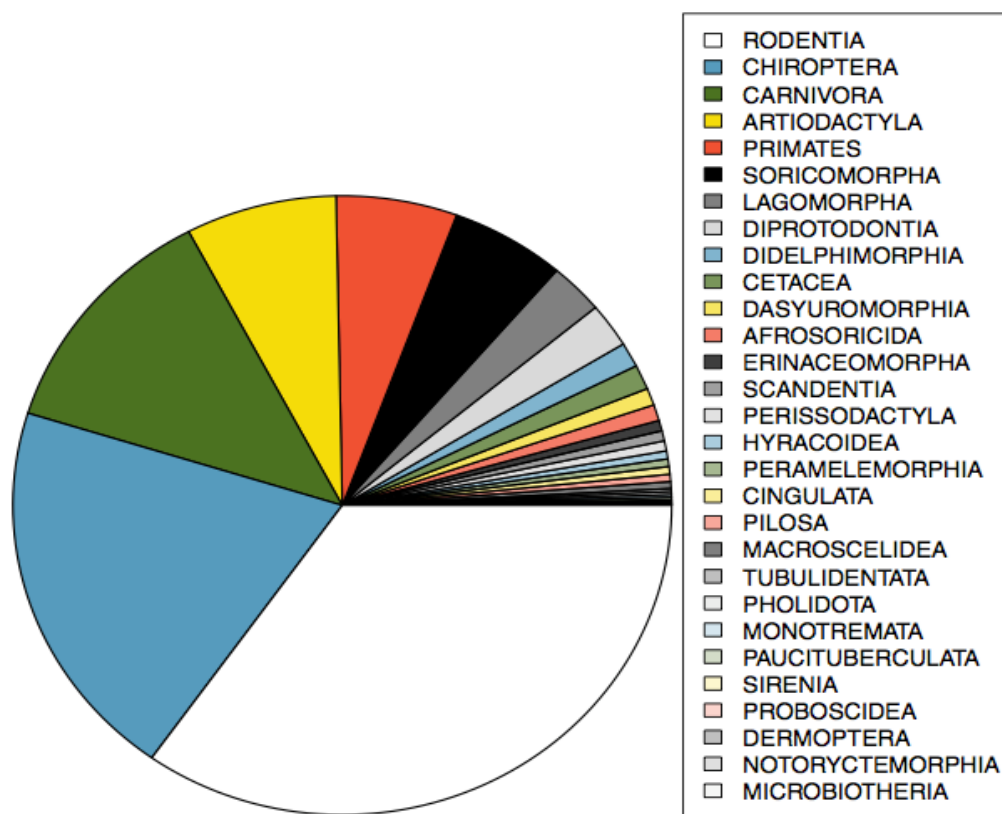
马赛克图

下面两种可视化呈现方式有什么不同？



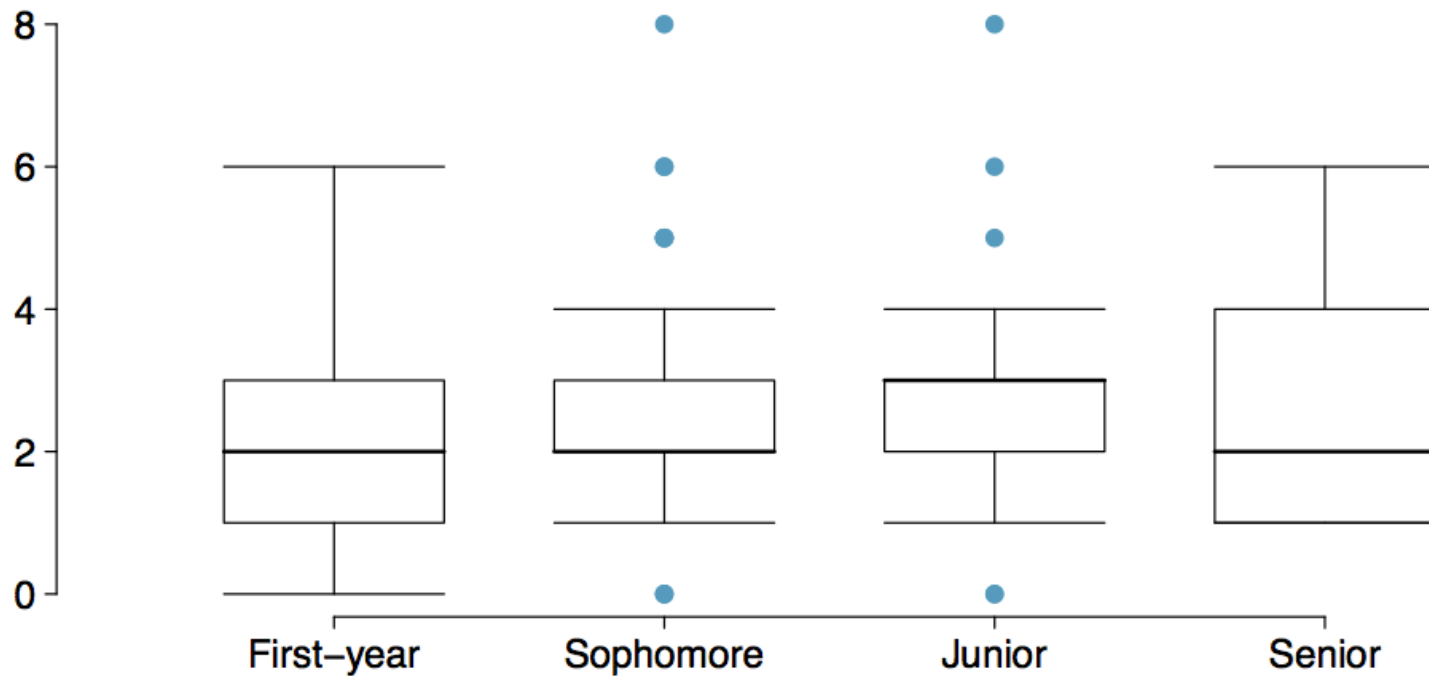
饼图

你能辨别出哪个顺序包含最少比例的哺乳动物种类吗？



比较组间数值数据

年级与学生参加社团数量之间是否存在关系？



案例研究： 性别歧视

性别歧视

- 1972年，作为一项关于性别歧视研究的一部分，48名男性银行主管每人收到同一份人事档案，并被要求判断该人是否应被提拔至被描述为“常规”的分行经理职位。
- 文件完全相同，除了一半主管的档案显示申请人为男性，而另一半主管的档案显示申请人为女性。
- 哪些主管收到“男性”申请、哪些收到“女性”申请是随机决定的。
- 在审查的48份档案中，有35人被推荐晋升。
- 该研究旨在测试女性是否受到不公平的歧视。

这是一项观察性研究还是一项实验？

数据

乍一看，晋升与性别之间是否存在关系？

		<i>Promotion</i>		Total
		Promoted	Not Promoted	
<i>Gender</i>	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

男性晋升比例： $21 / 24 = 0.875$

女性晋升比例： $14 / 24 = 0.583$

观察到某些差异并不一定意味着存在歧视，这些差异可能仅仅源于偶然或随机性。

练习

我们注意到男性档案与女性档案被推荐晋升的比例之间相差近30%（确切地说是29.2%）。根据这一信息，以下哪项是正确的？

- A. 如果我们重复这个实验，肯定会有更多女性档案被晋升。这一次只是偶然事件。
- B. 晋升取决于性别，男性更有可能获得晋升，因此在晋升决策中存在着对女性的性别歧视。可能
- C. 晋升男性和女性文件比例的差异是由于随机因素造成的，这并不能证明晋升决策中存在针对女性的性别歧视。可能
- D. 女性不如男性合格，这就是女性获得晋衔较少的原因。

两种对立的主张

1. “没有发生任何状况。”（没有区别）

晋升与性别无关，不存在性别歧视，观察到的比例差异只是随机因素导致的。

→ 原假设

2. “存在某种情况。”

晋升与性别之间存在依赖关系，存在性别歧视，观察到的比例差异并非随机因素导致。

→ 备择假设

作为假设检验的审判

假设检验非常像法庭审判。

- H_0 : 被告无罪
- H_A : 被告有罪
- 然后我们出示证据——收集数据。
- 然后我们依据证据进行判断——“如果原假设成立，这些数据是否有可能仅仅通过随机因素而产生？”
 - 如此巨大的差异，会有多罕见呢？是偶然，还是确凿的证据？
 - 如果这些情况发生可能性极低，那这份证据就让我们内心对于原假设产生了远超合理怀疑的态度。
- 最终，我们必须做出决定。多不可能才算不可能呢？



作为假设检验的审判（续）

- 如果证据不足以推翻无罪的假定，陪审团就会做出“无罪”的裁决。
 - 陪审团并不是说被告清白，而是说没有足够的证据来定罪。
 - 实际上，被告可能是清白的，但陪审团无法确定这一点。
- 用统计学的话来说，我们未能拒绝原假设。
 - 我们从不宣布原假设为真，因为我们根本不知道它是否为真。
 - 因此我们从不“接受原假设”。

作为假设检验的审判（续）

- 在审判中，举证责任在控方。
- 在假设检验中，举证责任在于不寻常的主张。
- 原假设代表常规状态（现状），因此我们视为不寻常且必须为其收集证据的是备择假设。

回顾：假设检验框架

- 我们从一个代表现状的原假设 (H_0) 开始。
- 我们还有一个代表我们研究问题（即我们正在检验的内容）的备择假设 (H_A) 。
- 我们在假设原假设为真的前提下进行假设检验，可以通过模拟（今天）或理论方法（课程后续内容）。
- 如果检验结果表明数据没有为备择假设提供令人信服的证据，我们就维持原假设。如果检验结果提供了这样的证据，那么我们就拒绝原假设，转而支持备择假设。

模拟这个实验.....

.....在独立的假设下，即一切归因于随机性。

如果基于随机化模型的模拟结果看起来与数据相似，那么我们可以判定：男性和女性员工晋升比例的差异仅仅是由偶然因素造成的（晋升与性别相互独立）。

如果基于随机化模型的模拟结果看起来与数据不符，那么我们可以判定：男性和女性员工晋升比例的差异不是由于偶然因素，而是由性别的实际影响造成的（晋升与性别存在依赖关系）。

应用活动：模拟实验

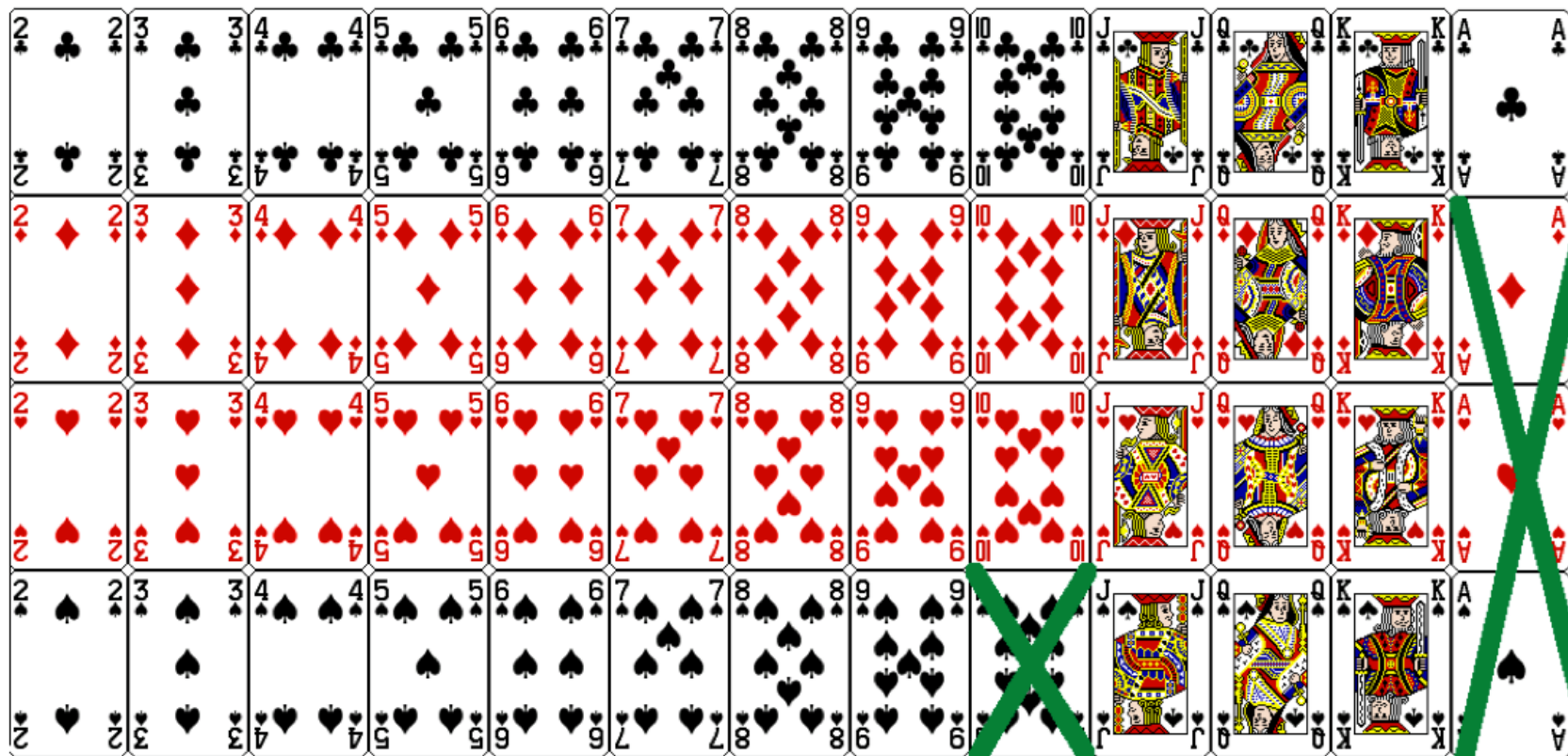
使用一副扑克牌来模拟这个实验。

1. 让花牌代表未提拔，非花牌代表被提拔。将A视为花牌。
 - 拿出小王和/or大王。
 - 取出3张A >> 至此牌堆中剩下恰好13张花牌（花牌：A、K、Q、J）。
 - 取出一张数字牌 >> 至此牌堆中剩下恰好35张数字（非花牌）牌（数字牌：2-10）。
2. 洗牌，并将牌分成两组，每组24张，分别代表男性和女性。
3. 数出并记录每组中有多少文件被提拔（数字牌）。
4. 计算每组被提拔文件的占比，并取其差值（男性-女性），记录此数值。
5. 重复步骤2-4多次。（一次模拟本身并不具有说服力）

步骤1

35 number (non-face) cards

13 face cards

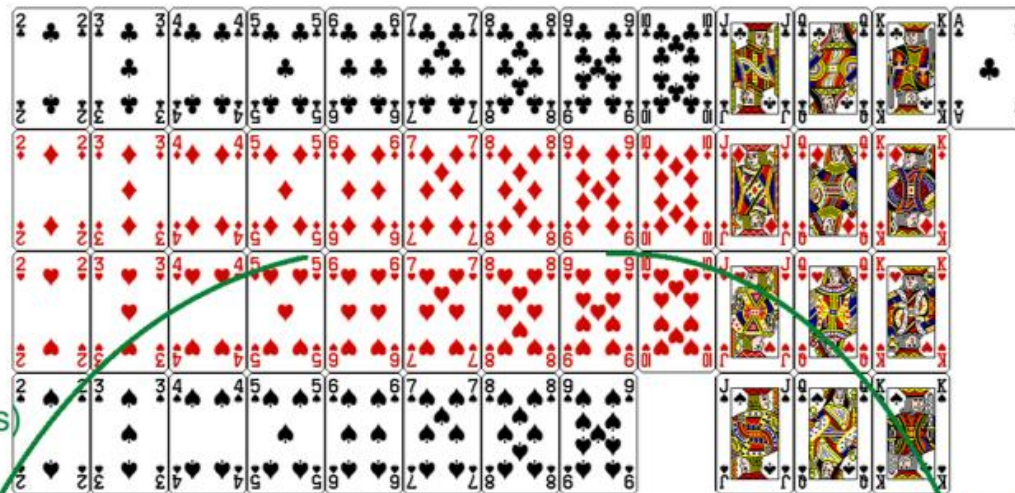


假设每位专家的回应与之前相同。

步骤2-4

Shuffle and
split into
two groups
of 24

(males and females)



Males
18 promoted
 $18 / 24 = 0.75$

Females
17 promoted
 $17 / 24 = 0.708$

Difference = $0.75 - 0.708 = 0.042$



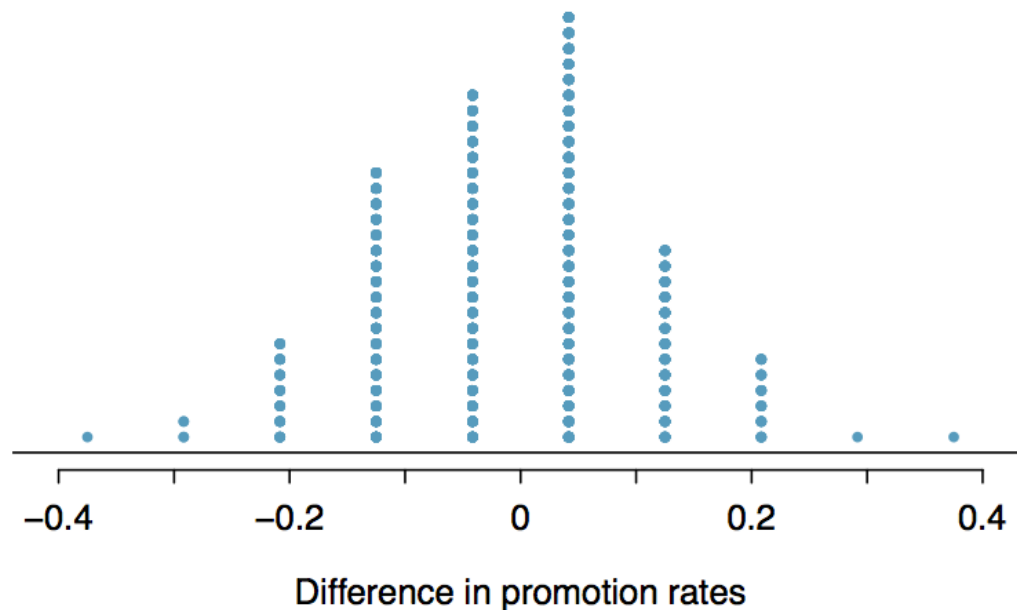
练习

你刚刚运行的模拟结果是否为性别歧视女性提供了令人信服的证据，即性别与晋升决策之间存在依赖关系？

- A. 不，这些数据没有为备择假设提供令人信服的证据，因此我们不能拒绝性别与晋升决策相互独立的原假设。观察到的两个比例差异是由随机因素造成的。
- B. 是的，这些数据为备择假设——在晋升决策中存在针对女性的性别歧视——提供了令人信服的证据。观察到的两个比例差异是由性别的真实影响造成的。

使用软件进行模拟

使用之前描述的方法运行这些模拟非常繁琐且缓慢。实际上，我们使用软件来生成模拟。下面的点图显示了基于100次模拟得到的晋升率模拟差异的分布。



我们多久才能看到至少相差29.3%的情况？可能但相当罕见。所以拒绝 H_0 。