MAT220 第2讲：数据可视化

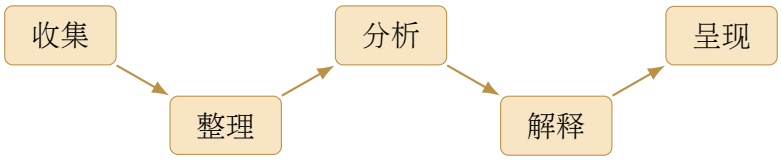
Contents

在第1讲中，我们介绍了统计学的基本语言：数据、变量、总体、样本，以及统计研究的基本流程。在本讲中，我们将开始讨论描述统计，重点放在数据的视觉呈现上。在课程的这一阶段，我们还没有引入任何公式。相反，今天我们要学习的是沟通这一思想：如何把一个数据集转化为一种视觉摘要，使别人能够快速浏览并理解。与上一讲的核心思想一致，不同的数据集可能具有不同的结构，因此“正确”的呈现方式将取决于这种结构。

---

1 作为呈现方式的数据可视化

上一讲中我们看到，在基本层面上，一个统计研究可以分解为五个关键组成部分：



我们强调过，这个流程更像是一条相互依赖的链条：如果其中任何一个环节以某种方式出现缺陷，那么就像是“破坏了有效性的链条”，研究结果也会变得不可靠。例如，如果我们一开始收集到的数据不准确或有偏，那么无论统计分析做得多好，结果都会和最初的数据一样不可靠。或者，如果你以无偏的方式收集了准确数据，但却使用了错误的公式，那么整个研究也会变得可疑。

在本课程后面，我们会花大量时间讨论统计研究中的分析和解释阶段。这是统计实践中极其重要的一部分：我们不仅需要知道应该正确使用哪些公式，还需要理解这些公式为什么正确，以及它们意味着什么。不过，在今天的讲义中，我们会假定这些更深层的考虑已经成立，然后直接进入研究过程的最后一步：呈现。

定义

数据可视化是指使用表格、图形、图表、示意图和地图等视觉对象来呈现数据。 其目标是让数据中的重要结构更容易被看见。

当然，统计研究有可能带来极其深刻的洞见。然而，我们的洞见始终取决于我们如何表示它：如果数据呈现得很糟糕，那么其中潜在的教训或叙事可能就无法被识别。换句话说，准确的数据如果不能让我们清楚地看到其中的模式，也没有什么用。

1.1 呈现更像艺术，而不是科学

一般来说，并不存在一种完美的数据集可视化方法。正如我们将看到的，不同的数据集允许许多不同的表示方式。因此，作为数据的呈现者，我们必须对自己想强调数据中的哪些要素作出有根据的选择。不同的呈现方式会对观众产生不同效果，因此也会传达略有不同的信息。由于这个原因，数据呈现更像是一种艺术形式，而不是严格的科学：我们并没有一个可以直接套用的“公式”来创造最佳的数据呈现，相反，我们需要结合实践判断和审美品味。正如我们将看到的，恰当的组合通常取决于我们的数据、语境和受众。

重要观点

数据呈现更像艺术，而不是科学。 图形并不是一个中立的信息容器。相反，它承载着呈现者在过程中作出的选择，而这些选择会影响观众注意到什么。

当然，这并不意味着什么都可以。正如我们将看到的，图形可能确实很差、很丑、具有误导性，甚至明显不诚实。在今天的讲义中，我们会逐步识别一些良好数据呈现的关键原则。在这个过程中，我们会概览一些基本的数据表示方法。我们将讨论：

1. 定量数据，
2. 分类数据，
3. 时间依赖数据，
4. 成对数据。

最后，我们将以一个专门为本讲设计的案例研究结束，这样你可以看到同一个数据可以用多种方式呈现。

2 表示定量数据

回忆一下，当一个数据集由数值测量或计数组成，并且算术运算对这些数值有意义时，我们称这个数据集为“定量”的。由于数值数据具有数学结构（顺序、算术等等），在视觉表示数据时，考虑这种额外结构通常很有用。

一般来说，直接呈现大量定量数据并不是特别有用。为了说明这一点，请考虑下表。它列出了42个人的鞋码，单位为厘米：

42个人的鞋码					
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

上表只是展示了原始数据。虽然这种呈现方式非常开放且诚实，但它并不特别有洞察力，因为我们无法轻易看出数据中的有用模式。例如，仅凭这些数字，我们无法快速回答以下可能有用的问题：

- 哪个鞋码最常见？
- 这组人的鞋码大致如何分布？
- 是否存在离群值？

这些问题的答案在某处包含于数据之中，但它们被隐藏起来了。更好的做法是通过另一种表示方式把这些隐藏结构揭示出来。

2.1 频数表与直方图

处理定量数据时，把数据分组到相同大小的区间中通常很有用。这些区间常被称为组距或组，更常见地也叫作箱。通过这种分组，我们可以创建如下所示的频数表。为了清楚起见，我们把同一组原始鞋码数据与其分组频数表并排重新画出。

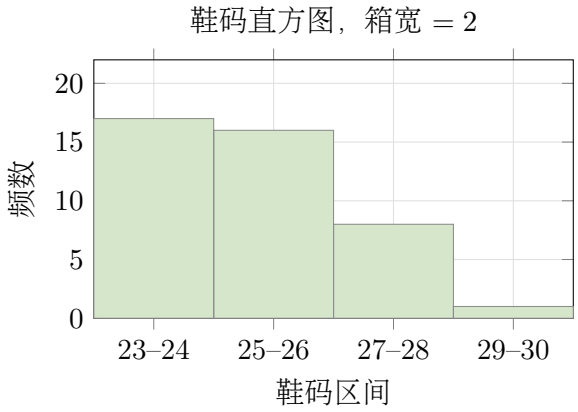
42个人的鞋码					
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

→

组距区间	频数
23–24	17
25–26	16
27–28	8
29–30	1

这里我们选择把42个数据点收集到宽度为2的箱中，也就是跨越数据范围的长度为2的区间。可以看到，频数表已经是一种改进，因为我们立刻能看出大量数据集中在23到26的范围内。

现在我们将使用频数表中的信息来绘制图形：x轴显示区间，y轴显示计数。这种图形称为直方图。



定义：直方图

直方图是一种用于表示定量数据的图形，它把数值分组到称为箱的区间中。每根柱子的高度表示落入该箱的数据值个数。

直方图展示的基本信息与频数表中的信息相同。然而，这些信息更加易于理解，因为我们可以快速看出各个频数之间的关系。

2.2 相对频数分布

频数告诉我们有多少观测值落入某个箱中。然而，有时候我们更关心比例，而不是原始频数本身。幸运的是，这些比例可以通过一个简单步骤计算出来。

定义：相对频数

一个组的**相对频数**是

$$\text{相对频数} = \frac{\text{该箱的频数}}{\text{总数}}.$$

它告诉我们数据集中有多大比例位于该箱中。

例如，使用我们的鞋码数据，第一组的频数是17，总共有42个人。因此，它的相对频数是：

$$\frac{17}{42} \approx 0.405 = 40.5\%.$$

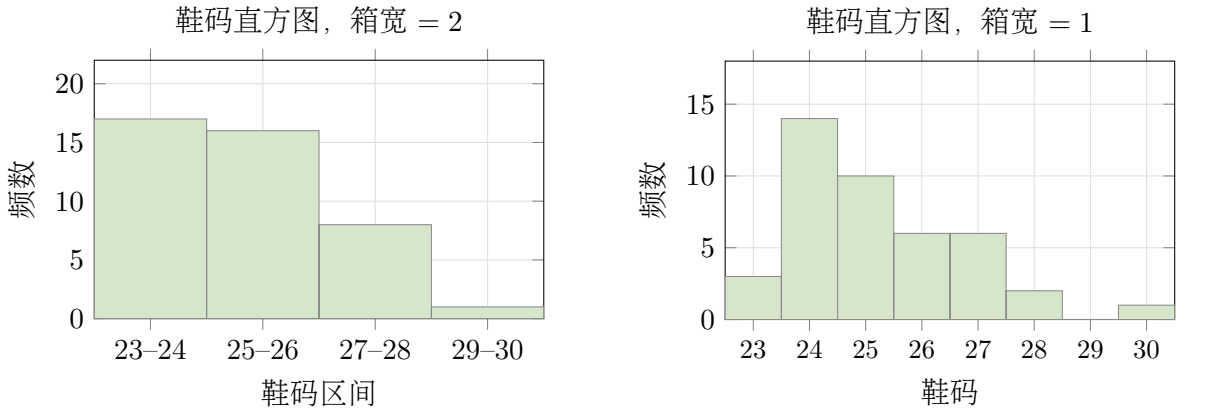
下表就是所谓的相对频数表。

组距区间	频数	相对频数
23–24	17	$17/42 \approx 40.5\%$
25–26	16	$16/42 \approx 38.1\%$
27–28	8	$8/42 \approx 19.0\%$
29–30	1	$1/42 \approx 2.4\%$

在比较大小不同的数据集时，相对频数很有用。 例如，如果一个班有42名学生，另一个班有120名学生，原始计数可能不能直接比较。 百分比通常可以让比较更清楚。

2.3 直方图的滥用：改变箱宽

直方图很有用，但也很容易被操纵。主要方法是改变箱宽，因为这可能导致数据被重新分配到不同箱中，从而出现不同形状。例如，我们的鞋码数据也可以按宽度1的箱来计数，这会得到一个看起来非常不同的直方图：



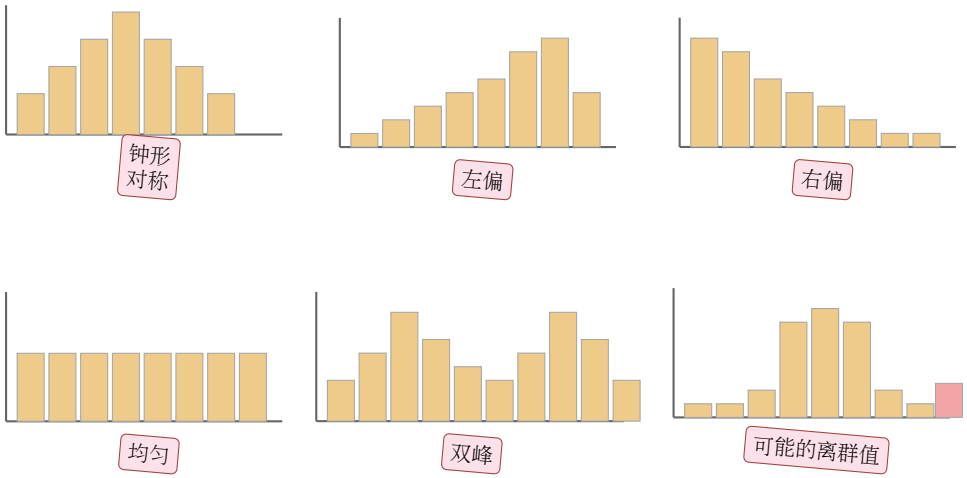
可以看到，右侧的直方图给出了更高分辨率的数据描述。这里我们可以看到每一个具体数值的分布。请注意，使用箱宽2时，一部分信息被遮蔽了：对于23–24这个鞋码箱，第一张直方图只是告诉我们其中有17个数据。然而，右侧的直方图传达出，鞋码为24的人比鞋码为23的人多得多。当我们把箱宽扩大到2时，这些额外信息就完全丢失了。

重要警告

改变箱宽可能改变直方图所讲述的视觉故事。 负责任的呈现者应该选择能够揭示数据结构的箱，而不是选择会夸大或隐藏某个预设结论的箱。

2.4 分布形状

有用的是，直方图的整体形状可以告诉我们数据大致是如何分布的。下面画出了一些常见的直方图形状。



我们现在逐一描述这些重要形状。

2.4.1 钟形对称分布

钟形对称分布是指直方图向中间升高，然后以大致平衡的方式向两侧下降的分布。两边不需要完全相同，尽管在许多情况下它们可能非常接近。核心思想是，中心部分是分布中数据最密集的部分，而数据从这个中心向两个方向扩散。

钟形对称分布

钟形对称分布是指直方图具有中央峰值，并且左右两侧大致平衡的分布。

许多自然变化的测量值，例如成年人身高，通常大致呈钟形，但并不是所有随机抽样数据都具有这种形状。

2.4.2 偏斜分布

偏斜分布是指一条尾巴比另一条尾巴更长的分布，也就是说它具有内在的不对称性。容易混淆的是，偏斜方向是按照更长尾巴的方向来命名的。也就是说，如果更长的尾巴在分布左侧，那么我们称该图形为“左偏”。

偏斜分布

当一个分布的一侧尾巴比另一侧更长时，这个分布称为**偏斜**。

- **右偏**： 更长的尾巴在右侧。
- **左偏**： 更长的尾巴在左侧。

长尾通常包含异常大的观测值或异常小的观测值。 例如，收入数据通常是右偏的，因为大多数人的收入处于普通范围，而少数人的收入极高。

2.4.3 均匀分布

均匀分布是指每个组具有相同频数，从而使直方图看起来平坦或呈长方形的分布。

均匀分布

均匀分布是指各组具有近似相等频数的分布。

精确的均匀分布在现实观察数据中并不常见，但在理论上很重要，尤其是在后面讨论概率论时。均匀分布表示这样一种情况：没有哪个区间相对于其他区间特别特殊。

2.4.4 双峰分布

双峰分布有两个明显的峰。这通常可能表明数据集混合了两个不同的总体，或者有两个不同的"解释性力量"作用于总体。其想法很简单：图形看起来像两个对称钟形分布的副本。单个峰可能暗示数据中的一个子群、机制或重复模式。对于双峰分布，有两个峰，这可能意味着有两种不同的性质或力量作用于数据。

双峰分布

双峰分布是指具有两个主要峰的分布，通常这两个峰之间至少有一个较低频数的组。

例如，我们可以测量经过新宿这样的热门城市车站的人流量。其相对频数分布会显示出两个明显的峰：一个集中在上午8点左右，另一个集中在下午6点左右。这是因为通勤者上班和下班时分别会形成两个交通高峰。

2.5 直方图告诉我们什么

直方图和相对频数直方图可以向我们展示定量数据的大致分布方式。通过观察直方图，我们通常可以判断：

- 数据分布是对称、偏斜、均匀还是双峰，
- 是否存在可能的离群值，
- 哪些数据区间包含最多数据，
- 数据有多分散。

3 表示分类数据

回忆一下，分类数据由标签、名称、组别或类别组成。分类数据不一定具有我们可以利用的内在数学结构。例如，如果我们的类别是血型，那么把A型阴性和O型阳性以某种有意义的方式“相加”是没有意义的。由于分类数据附带的数学结构非常少，我们能做的事情相当有限：主要操作是计算频数和比较比例。这两个操作引出分类数据的两种常见表示：

- **条形图**，表示频数或百分比；
- **饼图**，表示整体中的比例。

我们现在分别讨论它们。

3.1 条形图

条形图使用图中的独立条形来表示各个类别，类似于直方图。每个条形的长度或高度表示一个数值，通常是频数或百分比。

定义：条形图

条形图是一种用于分类数据的图形，其中每个类别由一个单独的条形表示。 条形的高度或长度表示该类别的频数、百分比或数值。

好的条形图遵循三个基本规则。

1. 条形可以是竖直的，也可以是水平的。
2. 条形应具有统一的宽度和统一的间距。
3. 每个条形都应使用相同的测量尺度。

条形之间的间隔很重要。 直方图表示数值尺度上的区间。 条形图表示不同的类别标签，因此条形通常在视觉上彼此分开。

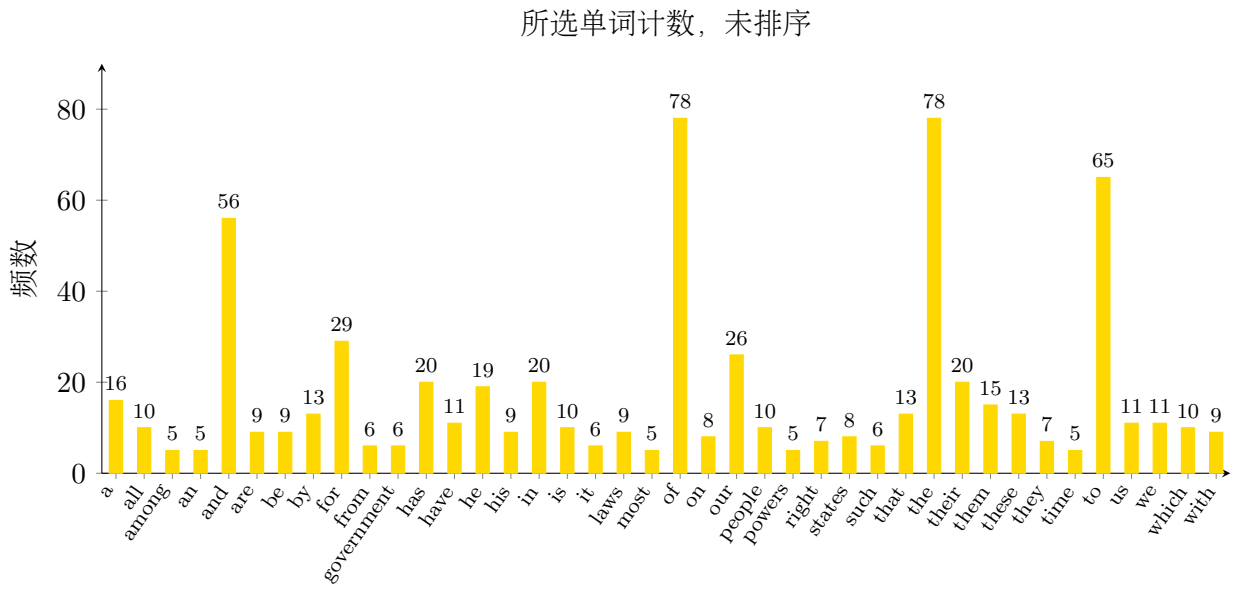
帕累托图是一种条形图，其中条形按递减顺序排列。 最常见的类别放在左侧，最不常见的类别放在右侧。

定义：帕累托图

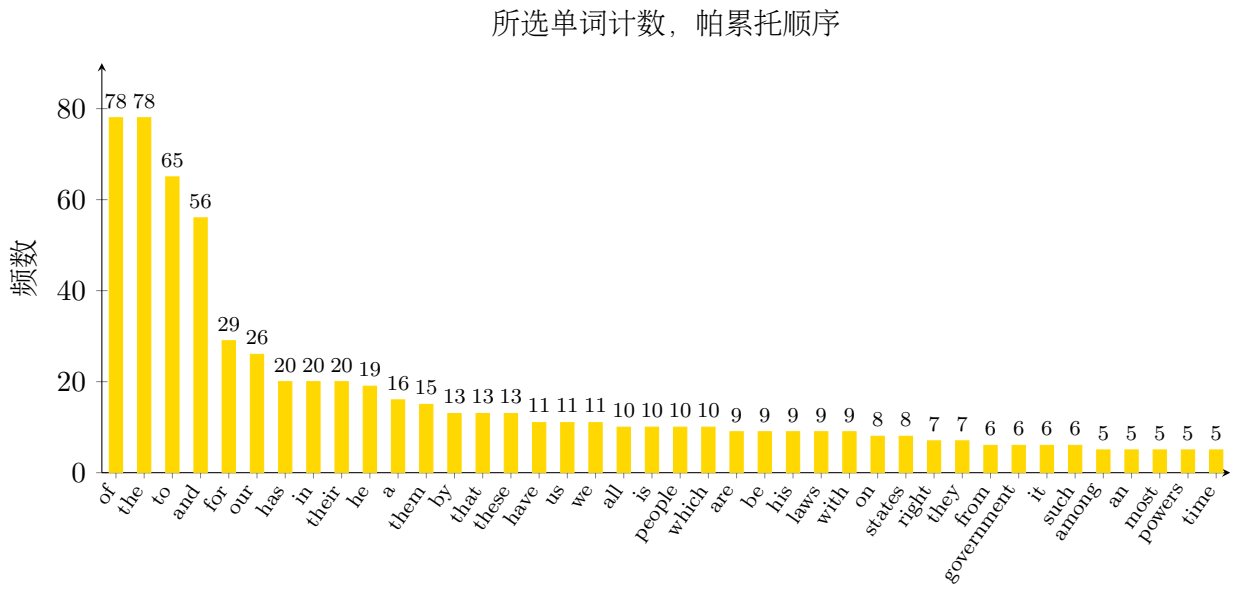
帕累托图是指条形从大到小排列的条形图。 当我们想快速识别最常见的类别时，它很有用。

3.2 例：美国独立宣言

假设我们统计《美国独立宣言》中频繁出现的单词。使用这一特定词频统计，正文包含1333个词，因此如果把每个词的频数都画成条形图，会非常拥挤。因此，我们将把所有至少出现5次的词画成条形图：



可以看到，这个图可以阅读，但仍然相当拥挤。由于有太多单独的条形，我们很难快速回答“哪些词属于最频繁出现的词”这样的问题。使用帕累托图可以改善这一点：



帕累托图包含相同的计数，但排序使主导类别更容易识别。换句话说，帕累托图中的顺序与视觉协调性更容易被眼睛接受。

3.3 饼图

饼图把比例表示为圆中的扇形。制作饼图时，我们首先通过计算每个类别出现的次数，然后除以总数来计算比例。

定义：比例

比例是一种部分与整体的比较：

$$\text{比例} = \frac{\text{频数}}{\text{总数}}.$$

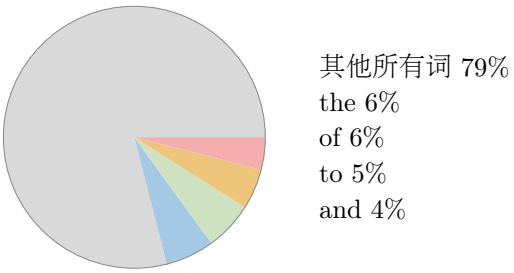
作为例子，我们使用上一节的数据。观察到单词“the”在《美国独立宣言》中出现78次，而总词数为1333。因此，单词“the”的比例是：

$$\frac{78}{1333} \approx 0.0585 = 5.85\% \approx 6\%,$$

也就是说，“the”大约出现了6%的时间。

定义：饼图

饼图是一种图形，其中类别被表示为圆的扇形。 每个扇形的大小表示该类别占总体的比例。



易读的饼图只有少数几个较大的扇形。

重要警告

饼图在只有少数几个扇形时效果最好。 如果饼图包含许多很小的切片，观众可能无法准确观察或比较各类别。

练习

假设一个40人的班级被问到他们最喜欢的便利店。 结果如下：

$$\text{7-Eleven} : 18, \qquad \text{FamilyMart} : 12, \qquad \text{Lawson} : 8, \qquad \text{其他} : 2.$$

(a) 选择7-Eleven的学生比例是多少?
(b) 这个数据集适合使用饼图吗?
(c) 这个数据集适合使用帕累托图吗?

解答

(a) 比例为 $18/40 = 0.45 = 45\%$。
(b) 是的。这里只有四个类别，因此饼图仍然可读。
(c) 是的。帕累托图也合理，尤其是当我们想快速显示最受欢迎的商店时。

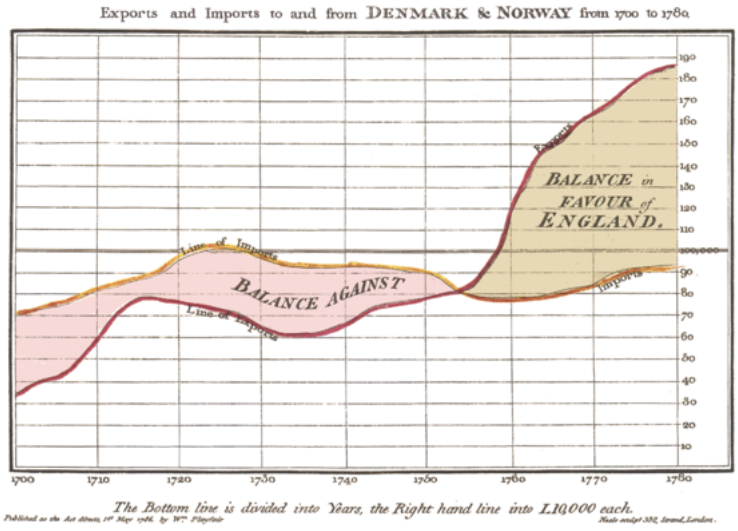
4 表示时间依赖数据

有时我们通过反复观察同一个对象来收集数据。这使我们能够追踪某个性质随时间的变化。例如，我们可能每周测量体重，每天记录向日葵生长的高度，记录每晚学习的小时数，或者追踪一个词在许多年代中的流行程度。这类数据称为时间依赖数据或时间序列数据。

定义：时间序列图

时间序列图是一种按照时间顺序绘制时间依赖数据的图形。横轴通常表示时间，纵轴表示正在追踪的数量。

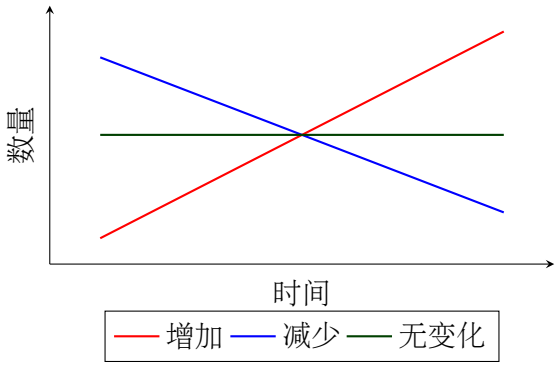
统计时间序列图的早期著名例子之一，是威廉普莱费尔在十八世纪末制作的：



可以看到，时间序列图非常直观：我们可以相当清楚地看到一个数量如何随时间变化。现在我们简要回顾几种有用的变化类型。

4.1 线性变化

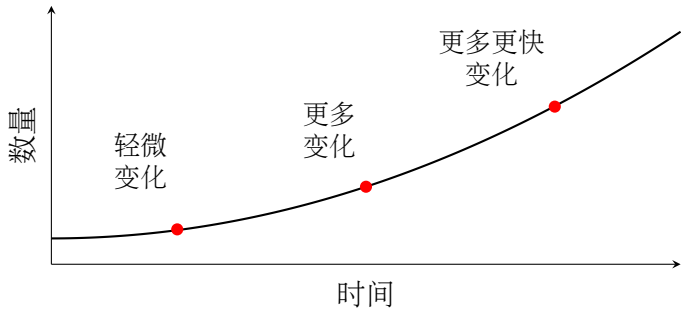
随时间变化的最简单类型是线性变化。这意味着当数据集被绘制在时间序列图上时，会形成一条直线。从几何上讲，变化率由直线的斜率表示：



上升的直线表示一个数量随时间增长。 下降的直线表示一个数量随时间减少。 水平的直线表示没有变化。 在每种情况下，主要的视觉思想都是相同的：线的形状告诉我们变化的方向和速度。

4.2 非线性变化

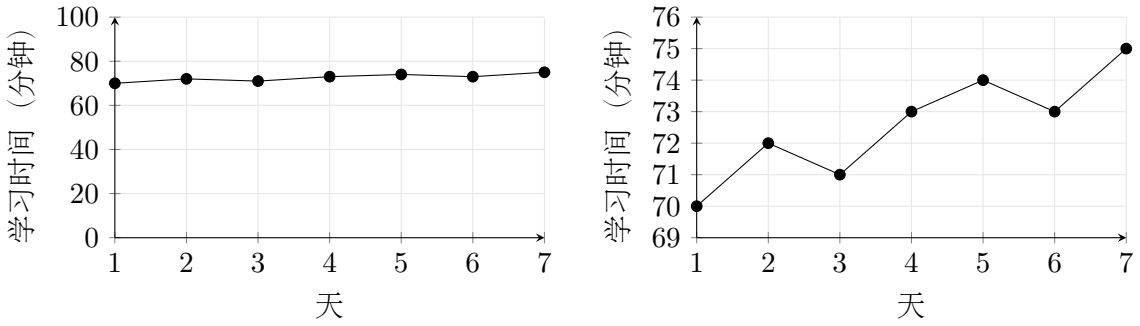
在大多数情况下，变化并不是以恒定速度发生的。相反，根据所测量的过程，变化本身可能会加快或放慢。考虑下面的图形：



在上图中，我们看到的是加速的例子，因为变化率本身正在随时间增加。

4.3 尺度与叙事

时间序列图对尺度特别敏感：改变纵轴可能使一个小变化看起来很剧烈，或者使一个大变化看起来很普通。为了说明这一点，请看下面两张图：



上面的两张图都是由同一个基础数据集构造出来的。然而，通过调整纵轴尺度，我们可以让数据显得非常不同。尺度的选择会微妙地嵌入不同的叙事：第一张图暗示数据变化很小，而第二张图暗示数据相当波动。这一观察引出下面的提醒。

重要警告

不同尺度会创造不同叙事。 阅读时间序列图时，在解释其故事之前，一定要先检查坐标轴。

这是数据可视化误导人们的最常见方式之一。 一张图可能在技术上是正确的，但如果选择尺度是为了夸大效果，它在视觉上仍然可能具有误导性。

5　表示成对数据

有时我们收集的数据自然成对出现。这通常发生在我们研究同一个对象并同时收集两个不同性质时，或者我们想比较每个观测对象的两个性质时。例如：

- 同一个人的左脚长度和右脚长度。
- 学生前往校园的通勤时间和距离。
- 住在东京的人们的身高和体重。

定义：成对数据

成对数据是指每个观测由两个相关测量值组成的情况。 我们经常研究成对数据，以调查两个变量之间是否存在关系。

5.1　例：通勤时间与距离

考虑下面这个由12名学生组成的数据集。

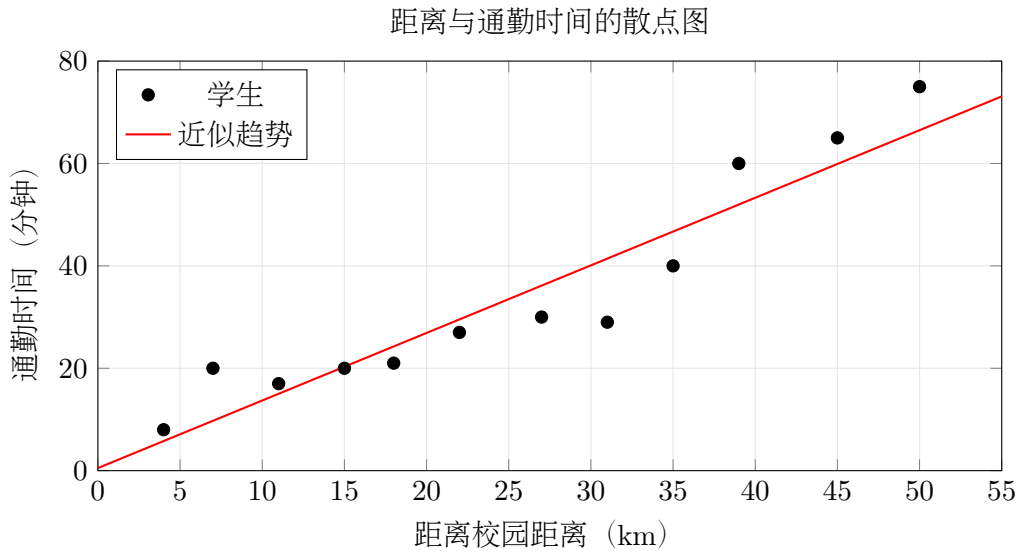
学生	距离校园距离（km）	通勤时间（分钟）
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75

这个数据集自然是成对的，因为每一行包含同一名学生的两条相互关联的信息。我们可以把这一对信息分成两个变量：第一个变量是距离校园的距离，第二个变量是通勤时间。这种情况下自然的图形是散点图。

定义：散点图

散点图是一种用于成对定量数据的图形。 每个观测被表示为一个点，其中横坐标给出一个变量，纵坐标给出另一个变量。

为了创建散点图，我们把一个变量画在x轴上，另一个变量画在y轴上：

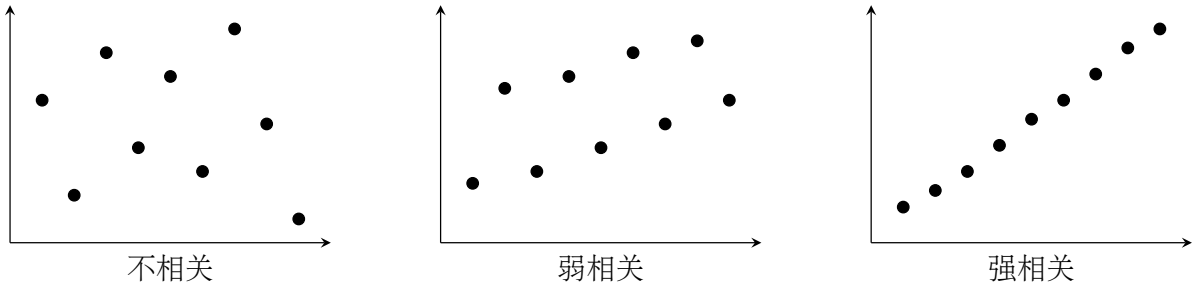


5.2 相关

上面画出的散点图显示两个变量之间存在明显的正关系，也就是说，住得离校园更远的学生往往通勤时间更长。当然，这并不是一个完美关系，因为通勤时间通常不仅仅取决于距离。不过，两个变量之间仍然存在清晰的一般模式。当成对数据显示出一般模式时，我们说这些变量是相关的。

定义：相关

相关意味着两个变量倾向于一起变化。 散点图可以显示成对数据是不相关、弱相关还是强相关。



课程后面，我们会学习一些如何用数值描述相关的方法。但现在，只需要能够识别基本思想即可：

- 如果点没有形成清晰模式，变量可能不相关。
- 如果点大致朝一个方向移动，但离散程度很大，变量可能弱相关。
- 如果点接近一条直线，变量可能强线性相关。如果点接近一条曲线，变量可能强关联。

负相关会表现为向下的趋势。

在通勤时间的例子中，我们画了一条穿过数据的近似直线。 寻找最佳拟合直线的行为称为**线性回归**。 我们将在课程后面更正式地学习它。

线性回归

线性回归是一种寻找直线的方法，该直线用来概括两个定量变量之间的关系。 在本讲中，我们只是在视觉上使用它。 正式方法将在后面介绍。

5.3 相关不等于因果

正如我们在第1讲中讨论过的，相关并不自动意味着因果。 散点图可以暗示两个变量有关，但它本身并不能证明一个变量导致另一个变量。

例。 假设距离校园距离和通勤时间强相关。 认为住得更远通常会导致更长通勤时间是合理的。 然而，准确的通勤时间还可能取决于铁路线、换乘站、自行车使用条件、交通状况、步行速度，以及学生是否住在方便的车站附近。

好的散点图可能会暗示因果关系，但解释仍然需要语境。

6 案例研究：日本的便利店

假设有一家公司生产红豆面包。他们想在日本某地的便利店（konbini）中销售自己的产品，并且正在做市场调研，以便为决策提供依据。他们让可怜的实习生Jenny研究日本全国便利店的分布情况，然后把结果呈现出来。简单地说，Jenny的任务是：

- 弄清楚便利店在日本国内如何分布，
- 然后把数据呈现给她的上司。

于是，Jenny使用了一份按都道府县列出的便利店数据集：

都道府县	店铺数
北海道	3003
佐贺	360
山梨	459
福井	337
东京	7405
爱媛	583
宫城	1161
岐阜	854
茨城	1432
广岛	1205
福岛	904
千叶	2701

都道府县	店铺数
青森	610
香川	406
爱知	3642
冈山	797
栃木	921
山口	565
富山	485
鹿儿岛	665
秋田	449
山形	447
群马	900
京都	1077

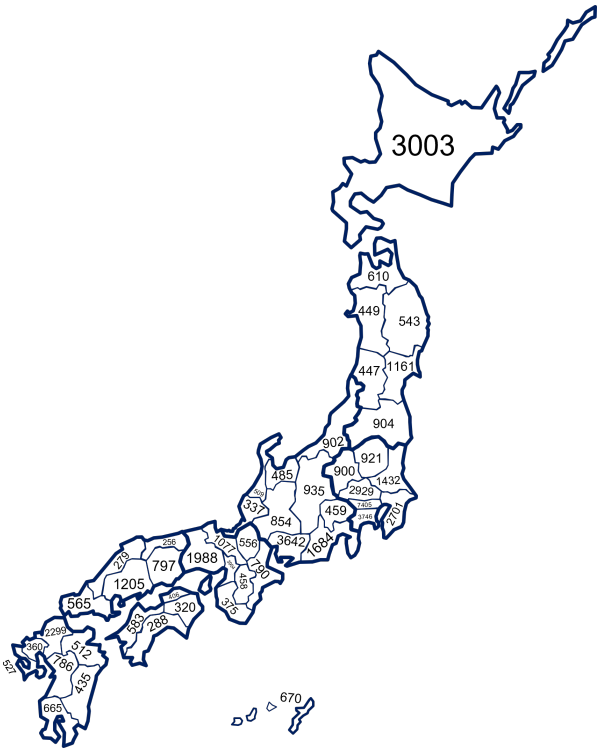
都道府县	店铺数
静冈	1684
高知	288
鸟取	256
岛根	279
冲绳	670
新	902
长野	935
宫崎	435
大分	512
和歌山	375
熊本	786
神奈川	3746

都道府县	店铺数
石川	509
长崎	527
岩手	543
玉	2929
福冈	2299
滋贺	556
大阪	3954
兵库	1988
三重	790
奈良	458
德岛	320

我们现在探索几种视觉呈现这组数据的方式。

6.1 地图表示

上面写出的数据是精确的，但并不容易解释。可以看到，它包含了所有相关信息，然而，要理解数据中的潜在模式非常困难。即使我们把它诚实地绘制在日本地图上，结果也并不具有启发性：



可以看到，这张地图并不是表示该数据的好方法。虽然它有一个好想法的雏形，但主要存在三个问题：

1. 数字太小，难以阅读；
2. 仍然难以看出数据中的任何模式；
3. 即使我们能够看清所有数字，它们的精确值也可能并不真正重要。

解决这个问题的一种方法是降低地图的分辨率，把数据按地区而不是都道府县汇总。首先，我们可以把计数改写为日本主要地区的数据：

地区	便利店数量
北海道	3003
东北	4114
关东	20034
中部	9807
关西	9198
中国	3102
四国	1597
九州	5584
冲绳	670

可以看到，这些数据讲述了一个更清楚的故事：关东的便利店数量遥遥领先，其后是中部和关西。然而，在让数据更易读的过程中，我们失去了细节。接下来我们会解决这一点。

6.2 颜色表示

在这个呈现语境中，Jenny很可能并不需要展示研究中的精确数值数据。换句话说，原始数据可能展示了过多信息，因而并不真正有用。我们可以通过颜色渐变来编码数量，在保留模式的同时压缩这些信息。下表中，较高的计数更接近蓝色，较低的计数更接近红色，中间为白色。

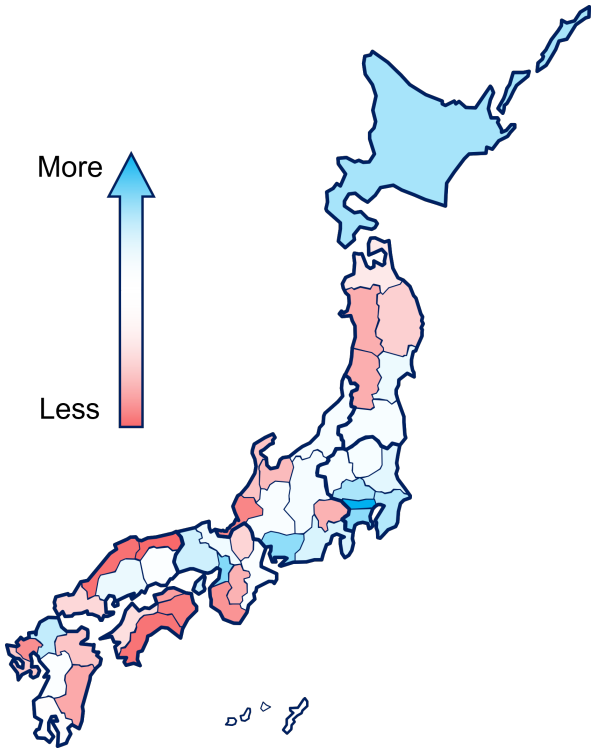
都道府县	店铺数
北海道	3003
佐贺	360
山梨	459
福井	337
东京	7405
爱媛	583
宫城	1161
岐阜	854
茨城	1432
广岛	1205
福岛	904
千叶	2701

都道府县	店铺数
青森	610
香川	406
爱知	3642
冈山	797
栃木	921
山口	565
富山	485
鹿儿岛	665
秋田	449
山形	447
群马	900
京都	1077

都道府县	店铺数
静冈	1684
高知	288
鸟取	256
岛根	279
冲绳	670
新	902
长野	935
宫崎	435
大分	512
和歌山	375
熊本	786
神奈川	3746

都道府县	店铺数
石川	509
长崎	527
岩手	543
玉	2929
福冈	2299
滋贺	556
大阪	3954
兵库	1988
三重	790
奈良	458
德岛	320

我们可以把这些数据直接转化为地图上每个都道府县的对应颜色。当然，与直接呈现原始计数相比，这是一种信息损失。然而，这样做的收益更大：我们可以立刻看到数据中的地理模式。



可以看到，这些数据中的模式比之前按地区绘制的地图更加详细。这里，我们看到在城市地区附近存在明显的集中热点：东京、名古屋、大阪，以及在一定程度上的仙台、广岛和福冈。相反，在人口稠密地区之外，便利店频数明显下降：岛根、鸟取、四国和东北北部都呈现低于平均水平的计数。除了这两个一般观察之外，我们还注意到北海道是一个明显的离群值。

虽然上面的表示比之前的图形更好，但它仍然有可以改进的地方：在Jenny和她公司的语境中，管理层很可能已经知道城市地区的便利店数量高于全国平均水平。因此，这项研究并没有真正展示出什么新的东西。不过，这并不意味着数据中没有别的东西隐藏着。

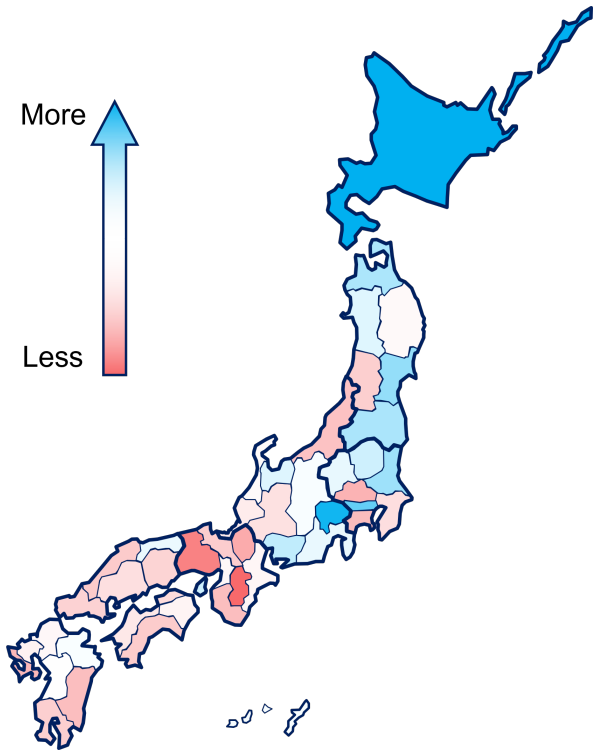
为了展示这组数据中的另一种模式，我们现在将店铺数除以人口，再乘以100,000，从而有效地创造出一个“每10万人便利店数量”的指标。下表中，我们再次使用颜色渐变来标记这一额外信息。

都道府县	店铺数	每10万人
北海道	3003	57.43
山梨	459	56.64
东京	7405	52.65
宫城	1161	50.40
茨城	1432	49.92
福岛	904	49.29
青森	610	49.24
爱知	3642	48.26
栃木	921	47.62
富山	485	46.83
秋田	449	46.77
群马	900	46.38
静冈	1684	46.33
鸟取	256	46.22
冲绳	670	45.63
长野	935	45.62

都道府县	店铺数	每10万人
大分	512	45.53
熊本	786	45.19
石川	509	44.91
岩手	543	44.83
福冈	2299	44.74
大阪	3954	44.72
三重	790	44.60
德岛	320	44.46
佐贺	360	44.33
福井	337	43.91
爱媛	583	43.65
岐阜	854	43.14
广岛	1205	43.01
千叶	2701	42.96
香川	406	42.69
冈山	797	42.18

都道府县	店铺数	每10万人
山口	565	42.07
鹿儿岛	665	41.85
山形	447	41.83
京都	1077	41.75
高知	288	41.62
岛根	279	41.54
新	902	40.96
宫崎	435	40.65
和歌山	375	40.63
神奈川	3746	40.54
长崎	527	40.13
玉	2929	39.87
滋贺	556	39.31
兵库	1988	36.35
奈良	458	34.56

这张表讲述的故事与原始总数完全不同：一些都道府县，尤其是山梨和东北北部的一些县，突然从红色变成蓝色，这意味着它们突然变得过度代表。相反，之前的一些城市地区在按人口缩放之后变成了红色。不过，东京仍然高度突出，北海道也仍然是离群值。更有趣的是，把这一模式画到日本地图上之后所显示的图形。请将下图与原始计数地图进行比较。



我们看到，在人均便利店数量方面，日本存在明显的东西分裂：日本东部通常得分较高，日本西部通常得分较低。在Jenny汇报的语境中，她因此可能建议进一步调查日本东部，具体取决于公司重视的是店铺密度、总店铺数，还是其他因素。

6.3 关于数据可视化的一些观察

最后，我们以关于数据可视化的一些重要观察结束。

重要观点

数据可视化通常是在追求洞见的过程中有意进行的信息损失。除了诚实和准确之外，你还应该遵循这样一个原则：你的呈现应当有用且有趣，理想情况下，它应当向观众展示他们原本可能看不到的东西。

数据呈现始终依赖于语境。正如我们在本讲中看到的，呈现者作出的选择总是可能扭曲叙事。今后当你观察数据时，最好问自己以下问题。

1. **数据类型：** 数据具有什么结构：定量、分类、时间依赖、成对，或者这些类型的某种组合?
2. **目的：** 这个图形试图回答什么问题?
3. **尺度：** 坐标轴和尺度是否诚实且易读?
4. **分组：** 如果数据被分组，箱或类别是否公平且合理?
5. **信息损失：** 哪些细节可能被隐藏或简化了?
6. **利益冲突：** 谁创造了这些数据? 他们为什么可能以这种方式呈现数据?
7. **叙事：** 这种可视化鼓励观众得出什么结论?
8. **其他视角：** 另一种表示方式是否会讲述一个不同但同样合理的故事?