第3章

描述、探索和比较数据

电子工业出版社
版权所有-盗版必究

本章问题　说到底，这就是一个小世界！

数据集 33“迪士尼乐园等候时间”中包含数个迪士尼游乐项目在上午 10 点和下午 5 点的等候时间（数据是作者通过迪士尼应用程序“我的迪士尼体验”采集的）。哪些游乐项目的等候时间最长？单个项目等候时间的离散程度有多大？不同时间段的等候时间有何不同？对于园区游客和负责园区管理及运营的工作人员来说，这些都是至关重要的问题。

队列是指一列等待被服务的人、交通工具或者物品。排队论是一门复杂但重要的学科，它的应用涉及我们每个人的日常生活。值得注意的是，这其中关于队列的心理学应用通常比统计学要重要得多。让我们来看以下三个例子。

- 到达休斯敦国际机场的旅客抱怨等候拿取行李的时间太长，但是通过增加行李传送带减少等候时间并没有消除旅客的抱怨。于是，旅客下机闸口被设置在远离行李提取处的地方，旅客需要走更远的路才能拿到行李。这样一来，旅客的抱怨反而显著下降了，原因在于，

现在他们把原本等待的时间用在了步行上。

- 现在很多车管局都会为访客提供排队号以及预期所要等待的时间。大屏幕上会清晰地显示当前排队状况。研究显示，为访客提供排队信息能够有效减少他们排队时产生的紧张和焦虑情绪。
- 奥兰多环球影城的“哈利·波特与逃离古灵阁”的排队等候区设有夺人眼球的设施和表演，比如可以看到有很多哥布林的古灵阁、以比尔·韦斯莱为特色的表演，以及带给人们深入地下金库错觉的电梯。这些眼花缭乱的体验让炎炎夏日中的排队等待变得惬意。

本章将着手研究关于等候时间的统计量，但应当谨记，许多行之有效的决策不仅仅需要统计知识，更需要依据常识。

让我们来考虑两个最受欢迎的迪士尼游乐项目“飞越太空山”和“阿凡达飞行历险”在上午 10 点的等候时间。

图 3-1 中的点图显示相比于“阿凡达飞行历险”，“飞越太空山”的等候时间更短。后者最大值与最小值之间的差值也小于前者。这种图表的解读过于主观，本章将介绍在任何统计研究中都至关重要的度量指标，其中非常重要的统计量有均值、中位数、标准差和方差。我们将使用这些统计量来描述、探索和比较“数据集 33”中列出的迪士尼著名游乐项目的等候时间。

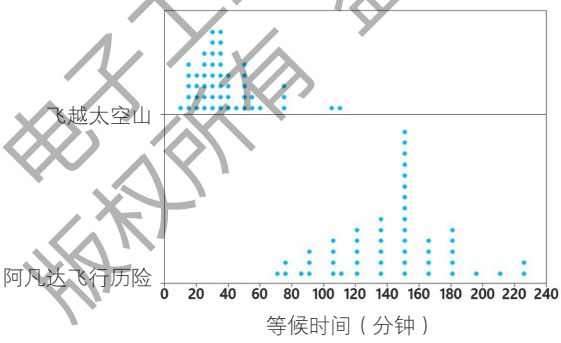

图 3-1：迪士尼乐园的等候时间

电子工业出版社 版权所有 盗版必究

本章目标

批判性思维和解读：超越公式与算术本身

现代统计学课程强调的不仅仅是记忆公式和进行复杂运算的能力。我们可以借助统计软件得到结果，以便把精力放在如何通过批判性思维给出合理的结论上。本章提供了相关重要流程的详细步骤，我们并不需要将其全部熟练掌握，但一般而言，在运用统计软件之前，进行一些手算可以帮助我们更好地深入理解。

本章中出现的方法和工具通常被称为描述性统计方法，它用于概括或描述数据的相关特征。在后面的章节中，我们将会运用推断统计对总体做出推断或泛化研究。以下是本章目标。

3-1：集中趋势的度量指标

- 通过计算均值、中位数、众数和中程数，掌握集中趋势的度量指标。
- 判断异常值是否对均值和中位数产生实质性影响。

3-2：离散程度的度量指标

- 通过计算全距、方差和标准差，掌握离散程度的度量指标。
- 运用范围经验法则，掌握标准差的解读，并学会判断某一数值是否显著低或显著高。

3-3：相对位置的度量与箱形图

- 掌握 *z* 分数，学会通过 *z* 分数的结果判断某一数值是否显著低或显著高。
- 掌握百分位数和四分位数。
- 掌握构建数据的箱形图。

电子工业出版社 版权所有 盗版必究

3-1 集中趋势的度量指标

核心概念：本节的重点是讲解数据集集中趋势的度量指标。具体来讲，我们将用均值和中位数来度量集中趋势。我们的目标不仅仅是求解度量集中趋势的数值，更重要的是如何解读。建议读者在学习第 2 部分之前，先充分理解本节第 1 部分介绍的核心概念。

第 1 部分：集中趋势度量的基本概念

这一部分包含度量集中趋势的不同统计量：均值、中位数、众数和中程数。集中趋势的度量指标被广泛用于“总结”数据集的有代表性的数值。

定义

集中趋势的度量指标是位于一组数据中心或中间的数值。

度量集中趋势可以有不同的方法，对于不同的方法会有不同的定义。我们从均值开始介绍。

均值

通常来讲，在所有用来描述数据的数值型度量指标中，最重要的就是均值，也就是大家所说的平均数。

注意：在表示集中趋势的度量值时，千万不要使用“平均数”这个词。它通常被用来指代均值，但有时候也会被用来表示其他集中趋势的度量值。统计学家不会使用“平均数”这个词。在本书后面的章节中，但凡涉及有关具体集中趋势的度量值时，也不会使用该词。统计学界或者专业期刊同样不会使用该词。从现在开始，在提及集中趋势的度量值时，就不要再使用平均数这种说法了。

定义

数据集的**均值**（或**算术均值**）是针对其集中趋势的度量指标，即所有数据值之和除以所有数据值的个数。

均值的重要性质

- 如果从同一总体中抽取不同的样本并计算其均值（即样本均值），那么它们之间的差异会小于其他集中趋势的度量指标。
- 计算数据集的均值需要用到其中所有的数据值。
- 均值的一个缺点是，仅一个极端值（异常值）就可以大大改变均值的结果（根据以下定义，我们可以说均值不具备抗性）。

定义

如果一个统计量没有因极端值（异常值）的存在而发生很大的变化，那么该统计量就具备**抗性**。

均值的计算及其数学符号

均值定义的数学表达式为公式 3-1，其中希腊字母$\sum$为求和符号，所以$\sum x$为所有数据值之和。n 代表样本量，即所有数据值的个数。

公式 3–1

$$\text{均值} = \frac{\sum x}{n} \quad \begin{matrix} \leftarrow \text{所有数据值之和} \\ \leftarrow \text{所有数据值的个数} \end{matrix}$$

如果数据是来自总体的样本，则将均值记作$\bar{x}$（读作 x 杠）。如果数据即总体，则将均值记作μ。

电子工业出版社 版权所有 盗版必究

数学符号

提示：样本统计量通常用英文字母表示，比如$\bar{x}$，而总体参数通常用希腊字母表示，比如μ。

$\sum$	表示某个数据集的数据值之和
x	通常用作代表数据值的变量
n	样本量
N	总体大小
$\bar{x}=\frac{\sum x}{n}$	样本均值
$\mu=\frac{\sum x}{n}$	总体均值

课程人数的悖论

至少存在两种可以计算每门课程平均人数的方法，但结果却大相径庭。在一所大学里，如果我们得到737门课程所有学生人数的数据，那么可以计算出每门课程平均有40个学生。但如果创建一个列表，其中包含每个学生的课程人数，那么可以通过该列表计算出每门课程平均有147个学生。导致两者存在巨大差异的原因是大部分学生上的是大班课，只有少数学生上小班课。因为通常小班课出勤率更高，所以在不改变课程数量或教职员工的情况下，可以采取让每门课程的人数相对一致的方法来改善学生的上课体验，以提高出勤率。

电子工业出版社 版权所有 盗版必究

例 1：均值

数据集 33“迪士尼乐园等候时间”中包含 6 个大众游乐项目的等候时间（分钟）。试求在上午 10 点“飞越太空山”前 11 个等候时间的均值：

50　25　75　35　50　25　30　50　45　25　20

解答：

根据公式 3-1，先计算出所有数据值之和，再除以所有数据值的个数，即可得到均值。

$$\bar{x}=\frac{\sum x}{n}=\frac{50+25+75+35+50+25+30+50+45+25+20}{11}=\frac{430}{11}=39.1\text{ 分钟}$$

因此，“飞越太空山”的平均等候时间为 39.1 分钟。

轮到你了：试试 3-1 基础题的习题 5

中位数

中位数可以被大致理解为“中间的值”，即数据集中一半的值比中位数小，另一半的值比中位数大。以下为更加精确的定义。

定义

数据集的中位数是针对其集中趋势的度量指标，即将原始数据值按升序（或降序）排列后排在中间的值。

中位数的重要性质

- 如果数据中有一些极端值，中位数不会受到很大影响，那么中位数就是一个具备抗性的集中趋势的度量指标。
- 中位数没有直接用到所有的数据值（例如，如果将数据中的最大值改为更大的值，则中位数不会改变）。

中位数的计算及其数学符号

样本的中位数通常用 $\tilde{x}$（或 M，Med）表示，但并没有一个被广泛认定的符号，也没有专门的符号来标记总体中位数。计算中位数，首先要对数据值进行排序，然后按照以下流程之一进行计算。

1. 如果数据值的个数为奇数，那么中位数为排序后排在中间位置的值。
2. 如果数据值的个数为偶数，那么中位数为排序后排在中间两个值的均值。

电子工业出版社
版权所有 盗版必究

测量单位的重新定义

1983 年，距离单位“米”被重新定义为光束在真空中用 1/299,792,458 秒走完的距离。1967 年，时间单位“秒”被重新定义为铯 -133 的原子基态跃迁 9,192,631,770 个周期的持续时间。

在 1889 年到 2018 年期间，质量单位“千克”被定义为国际千克原器的质量：一块存放在巴黎保险库中的铂铱合金。现在质量单位“千克”被重新定义为普朗克常数：一个自然界的物理常数。具体的定义较为复杂，但新的定义可以实现用自然常数而不是实物来定义国际单位制（SI）的基本单位。

截至撰写本书时，仅有利比里亚、缅甸和美国这三个国家还没有把国际单位制纳入官方的测量单位制。

例 2：中位数——奇数个数据值

试求在上午 10 点“飞越太空山”前 11 个等候时间的中位数。

50 25 75 35 50 25 30 50 45 25 20

解答：

先将数据值按升序排列，结果如下：

20 25 25 25 30 35 45 50 50 50 75

因为数据值的个数是奇数（11），所以中位数为排序后中间位置的数据值，即 35.0 分钟。需要注意的是，中位数 35.0 分钟和“例 1”中得到的均值 39.1 分钟不同。

轮到你了：试试 3-1 基础题的习题 17

例 3：中位数——偶数个数据值

重复“例 2”，但使用在上午 10 点“飞越太空山”前 12 个等候时间的数据。即计算以下等候时间（分钟）的中位数：

50 25 75 35 50 25 30 50 45 25 20 50

解答：

先将数据值按升序排列，结果如下：

20 25 25 25 30 35 45 50 50 50 50 75

因为数据值的个数是偶数（12），所以中位数为排序后中间两个数据值（35 和 45）的均值。因此中位数为 (35+45)/2=40.0 分钟。

轮到你了：试试 3-1 基础题的习题 7

众数

众数并不常用于定量数据，但是唯一可以被应用于定性数据（只包括姓名、标签或者类别的数据）的集中趋势的度量指标。

定义

数据集中的众数是出现频数最高的值。

众数的重要性质

- 对定性数据也可以计算众数。

电子工业出版社 版权所有 盗版必究

- 一个数据集可以没有众数，或者有一个众数，或者有多个众数。

计算众数

一个数据集可以有一个众数、多个众数或没有众数。

- 如果两个数据值具有相同的最大频数，那么每个数据值都是众数，这样的数据集被认为是双众数的。
- 如果两个以上的数据值具有相同的最大频数，那么每个数据值都是众数，这样的数据集被认为是多众数的。
- 如果数据值没有重复出现，那么就不存在众数。

中位数并不是全部

哈佛大学植物学家史蒂芬·古尔德曾经写道：“中位数并不是全部信息。”他描述了他是如何因患癌症（腹膜间皮瘤）而有此观点的。他去图书馆了解到该类型的癌症是不治之症，生存期中位数仅为 8 个月，为此他十分震惊。古尔德写道：“我怀疑大多数没有学过统计知识的人看到这段话时，会理解为‘我在 8 个月内必死无疑’。但是我们不能下这样的结论，因为这不是正确的，与癌症做斗争的心态才是最为重要的。”古尔德非常仔细地解读了中位数：因为他很年轻，且癌症被发现于早期，有最好的药物可以治疗，所以他的生存期要比中位数大很多。他推断一部分人的生存期应该远高于 8 个月，没有理由不相信他不在这些人里面。具备了对中位数深思熟虑的解读和积极的心态，古尔德在确诊后又活了 20 年。他最后死于一种和腹膜间皮瘤不相关的另一种癌症。

电子工业出版社
版权所有 盗版必究

CP **例 4：** **众数**

试求在上午 10 点“恐怖魔塔”前 11 个等候时间的众数：

35　35　20　50　95　75　45　50　30　35　30

解答：

对数据值进行排序，以便找到出现多次的数据值：

20　30　30　35　35　35　45　50　50　75　95

因为出现频数最高的值是 35（三次），所以众数为 35 分钟。

轮到你了：试试 3-1 基础题的习题 7