MAT120：论证　第13讲讲义

统计学导论

上一讲中，我们在学习概率，重点讨论了二项试验和概率分布。这一讲中，我们开始统计学单元，讨论什么是统计学、为什么抽样很重要，以及偏差如何悄悄混入数据之中。我们还会建立两个核心的描述性工具：集中趋势（寻找中间）和方差（衡量分散程度）。最后，我们会把这些思想应用到我们的Haribo 数据集上，作为一个完整例子。下一讲我们进入正态分布，再下一讲我们将学习假设检验。

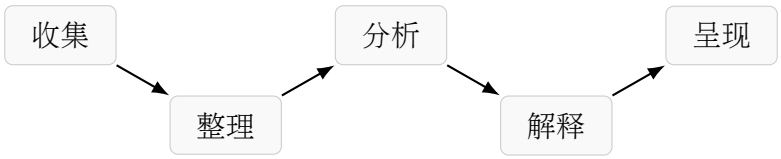
1 什么是统计学?

统计学是对数据进行数学研究的学科。我们可以把数据理解为“从现实世界中的某个对象那里收集来的一组信息”。因此，统计学的主要任务就是研究这些信息，并发现其中可能具有的结构，从而利用这种结构来帮助我们对现实世界中的事情作出判断和决策。统计学在科学与工业中无处不在：从鸟类的迁徙模式到热现象的基本理论，从微芯片的制造过程到经济体系的复杂运作，到处都会用到统计学。在入门层面上，可以说，统计学是最有用、也最重要的数学工具之一。

接下来的三讲中，我们将介绍统计学中的一些基本概念。今天，我们从最开始讲起，回答关于统计学本质以及数据集可能呈现出的基本结构的一些问题。在后面的课程中，我们将开始看到，统计学如何借助一些巧妙的数学工具，对系统作出有根据的检验。

1.1 统计研究的工作流程

一个好的统计研究，会把现实背景与数学方法结合起来，从而发现一个系统中的新信息。一个基本的统计研究流程可以写成如下形式：



每一步的含义

- **收集**，指对现实世界中的某个系统进行测量，并从中收集数据。
- **整理**，指把你的数据重新组织成一种有用的格式。
- **分析**，指使用合适的数学方法来描述数据集中的关键特征。
- **解释**，指把这些发现放回现实背景中，说明它们意味着什么。
- **呈现**，指用别人容易理解的形式，简洁地表达你的结果。

请注意，这个流程是现实考量与数学工具的有意识结合。这正是统计学的核心所在：统计学不仅仅是一套数学理论，相反，统计知识本身天然地涉及现实世界。

1.2 统计学的实践

统计学常常会让人感觉枯燥乏味，因为它看起来非常公式化。事实上，统计学中用到的许多公式背后都有很深的数学理论，但实际统计工作通常并不会直接使用那些理论。从数学复杂性的角度看，统计学有时会显得有些无聊，因为它非常朴素，而许多最常见的公式只涉及基础的算术。

但也正因为如此，统计学又非常有趣：仅凭最基本的数学运算，它就构造出了许多巧妙的公式，可以用来描述数据中的整体结构。统计学的魅力就在于，它从如此朴素的起点出发，却成为一种极其强大、极其有用的工具。统计学的实际价值怎么强调都不过分：世界本身就是由信息组成的，而关于你和社会的许多性质，都可以转化为数据形式并加以分析。

1.2.1 描述统计与推断统计

在实践中，统计学主要有两大支柱：描述统计和推断统计。

- **描述统计**处理的是对已经收集到的数据进行描述。我们容易想到的是新闻报道或科学论文中出现的图表和图形这类可视化描述。然而，大多数时候，我们也可以用数字来描述数据，比如平均数、变异程度，以及更复杂的方法如回归。
- **推断统计**处理的是我们可以根据数据得出的结论。通常，统计研究会受到资源限制，而收集"完美"的数据可能过于昂贵、过于耗时，甚至根本不可能。因此，推断统计使用一些巧妙的数学工具，利用少量信息来检验更大的系统。推断统计通常包括假设检验和估计等技术。

1.3 什么是数据?

现实中，我们并不总能获得一个能够完美描述某个系统的数据集。因此，我们只能接受手头现有的数据。统计学的任务之一，就是在并不完全了解总体的情况下，建立工具来推断关于总体的事情。

我们的数据质量，取决于我们获取数据的方法是否可靠。分析过程中，可能会不小心把一些不想要的影响带进去，而我们的任务就是识别这些影响，并尽可能减小它们。

1.3.1 数据的类型

数据主要分为两大类：定量数据和分类数据。

- **定量数据**是数值型数据。
- **分类数据**是非数值型数据。

定量数据具有某种数学结构，因此可以利用这些结构来帮助分析；而分类数据则更像是一份名称清单。

1.3.2 表示定量数据

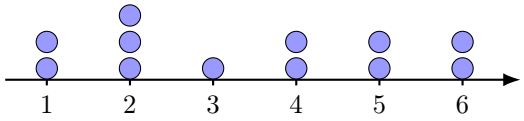
我们常常把数值数据写成$(x_1, x_2, \ldots, x_n)$ 这样的形式。这个记号和我们以前见过的集合论花括号{与} 很像。不过这里有一个重要区别：在书写数据时，我们还会考虑数据的顺序。这就是为什么我们用圆括号来写数据：这些括号提醒我们，要保留数据出现的顺序。[1]

示例数据： $(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$。

一种可视化重复取值的有用方法，是使用下面这样的图。在这个图中，我们把数值数据想成是画在数轴上的小球。每当有多个数据项取相同的数值时，我们就把另一个小球叠在前面的小球上方，这样高度就表示频数。为了完整起见，我们会总是在图的上方写出数据集本身。

[1]其实，我们在第8 讲讨论线性模型时已经见过这种记号。当时我们用"有序对"来表示方程的解，也就是形如(x, y) 的两个数。在数据集的情形中，我们是在推广这种记法。

数据：$(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$



当然，上面的图并没有保留数据原本的顺序。不过，因为大多数公式只使用加法和乘法这样的基本运算（它们都是可交换的），所以顺序并不总是那么重要。

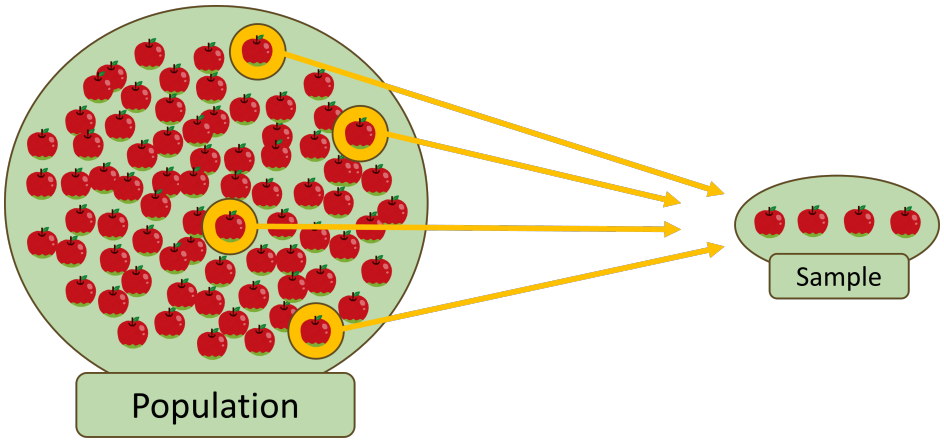
2 抽样

2.1 核心定义

很多时候，我们想研究的对象集合太大，无法直接完整研究。因此，在实际中，我们会取其中一个较小的子集来研究。这种现实考虑引出了下面这些术语。

定义

- **总体**：你想研究的全部对象所组成的集合。
- **样本**：从总体中选出的一个子集。
- **总体参数**：关于整个总体的某个事实、测量值或数量。
- **样本统计量**：从样本中计算出来的某个事实、测量值或数量。



从总体中获得样本的过程叫作抽样。抽样是统计学中极其重要的一部分，因为现实中我们几乎永远无法接触到完整总体。但是，在抽样时我们必须非常小心，因为有时候我们抽取数据的方式会不小心带入一种在总体中并不存在的叙事。在统计学中，这类错误叫作偏差。

2.2 例子：在东京抽样

假设Jenny 想做一个统计研究，以了解东京所有居民的平均工资。Jenny 并不在市政府工作，所以

她无法获得关于东京总体的人口普查数据。于是，她需要抽取一个样本。下面我们讨论Jenny 可以采用的一些抽样方法。不过在此之前，我们先说明一个重要观点。考虑下面两种抽样方式：

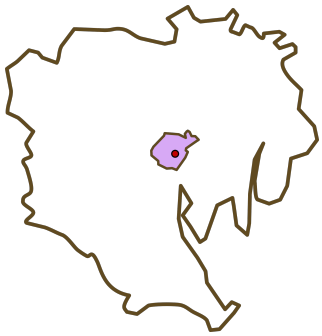
两种抽样方法

1. **街头抽样**：Jenny 在自己所在的街区随意走动，并在街上随机问30 个人的收入是多少。
2. **挨家挨户抽样**：Jenny 在自己所在的街区挨家敲门，询问人们的收入是多少。

现在，这两种方法看上去似乎没什么区别。然而，我们必须记住，数据收集里存在人的因素。例如，有些人可能会因为一个陌生人突然来到家门口、未经邀请就询问金钱问题而感到威胁。在这种情况下，人们也许会倾向于低报自己的财富，因为Jenny 可能是一个在踩点找目标的小偷。于是，在最终得到的样本数据中，平均工资也许会低于街头抽样得到的平均值。

2.3 可能的抽样方法

在东京这个背景下，Jenny 可能采用如下四种抽样方法。

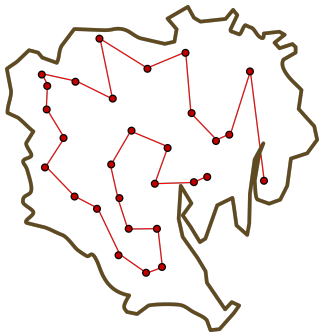


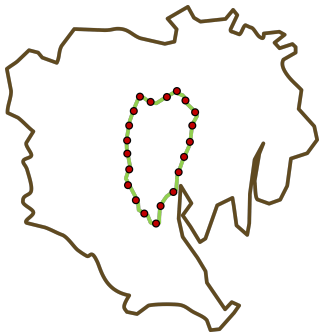
方法1：在她居住的千代田区选出30 名参与者。

缺乏多样性：从自己所在的街区抽样当然很方便。然而，她得到的数据可能并不能代表整个东京。有些街区更富有，有些更贫穷；有些更老龄化，有些更年轻；不同地区的人从事的工作类型也可能不同。因此，即使她询问了30 个人，她也可能只是在测量这个街区本身的特征，而不是东京整体的特征。对于千代田区而言，她的样本很可能会高估平均工资。

方法2：在东京全市范围内随机选取30 名参与者。

后勤困难：从整个城市随机抽样，在纸面上看起来似乎最不带偏差，但真正实施起来会既昂贵又耗时。她不仅需要前往各个随机选中的街区，而且在决定具体问谁的时候，还会遇到其他潜在偏差。尽管如此，一个看似在理论上无偏的方法，在实践中仍可能因为资源限制而失败。

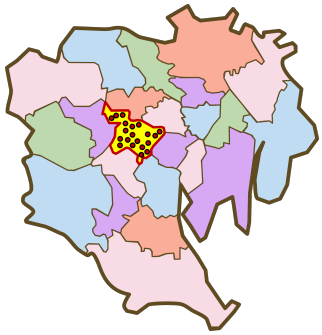




方法3：沿山手线出行，在30 个车站分别停下，并在每个车站询问一人。

可接触性不足：按“每站一人”来抽样，可能会把游客和并不住在东京的人也包含进来。它还可能无意中对旅游车站、通勤枢纽或商业区产生偏向，从而使样本偏向某类特定工作人群。换句话说，这种样本可能偏向某一类东京市民，也就是那些在山手线附近工作或活动的人。

方法4： 先从东京23 个区中随机选一个区，再在这个区内随机选取30 名参与者。



缺乏多样性： 虽然这种方法比跑遍整个东京更方便，也比只在Jenny 自己所在的千代田区抽样偏差更小，但这种样本仍然会缺乏多样性。随机性并不能保证一个小样本就能完整呈现总体中的多样性。随机性所能保证的，只是选取规则本身没有偏向。

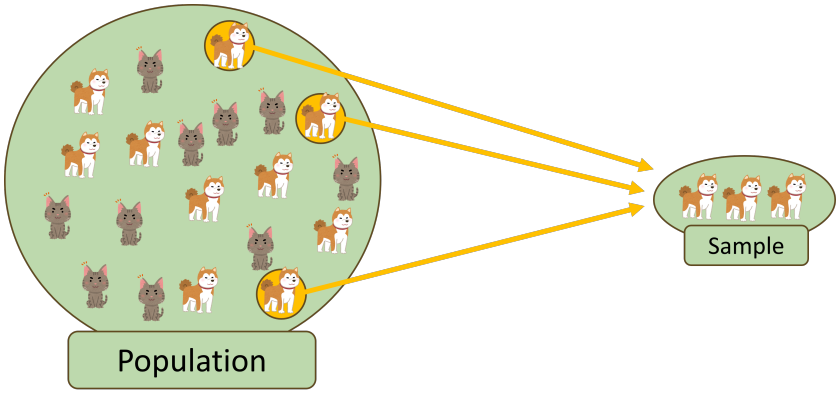
那么，这里最好的抽样方法是什么呢？事实上，这取决于Jenny 想要做什么。在实践中，往往并不存在唯一的“最佳方法”。相反，抽样通常是在资源与误差之间进行平衡，有时我们必须作出妥协。理想情况下，进行研究的研究者应当理解这些偏差，并且在传达结果时尽可能明确地说明这一点。

2.4 简单随机抽样

定义：简单随机样本

一个容量为n 的简单随机样本，是指从总体中随机选出的一个子集，并且总体中的每一个成员都有相同的被选中机会。等价地，在不放回抽样的情况下，总体中每一个大小为n 的子集，都有相同的被选中机会。

即使抽样是随机的，也始终有可能抽到一个“不走运”的样本，它并不能诚实地代表总体。比如，一个总体中有10 只猫和10 只狗，如果你随机抽取一个大小为3 的样本，那么完全有可能抽到3 只狗，从而低估了猫在总体中的比例。和前一节中的第四种方法一样，简单随机样本的“无偏”含义是，没有某种特别偏向某类样本的选择规则。但是，这并不保证你最终抽到的任何一个具体样本，都会诚实地反映总体中可能存在的多样性。



2.5 抽样偏差的例子

三种常见偏差

- **幸存者偏差**：“过去的建筑真漂亮，现代建筑都很丑。”这种说法可能出现，是因为我们主要记住并保留了过去最好的那些例子。
- **选择偏差**：“我在歌舞伎町问了所有人，东京是不是喧闹混乱，他们都说是，因此东京确实喧闹混乱。”但歌舞伎町并不是东京的随机样本。
- **资助偏差（利益冲突）**：如果一项研究由会从某种结论中获益的团体资助，那么结果就可能带上偏差。

以上这些都是抽样偏差的例子，也就是由于样本的选取方式，意外地暗示出一个本来在总体中并不存在的结论。在幸存者偏差的例子里，说话者是根据那些现在仍然存在的旧建筑来得出结论的。但更可能的情况是，那些丑的建筑已经被拆除并被新的重要项目取代了，而历史上较美的建筑则因为其审美价值而被保存下来。如果你能神奇地穿越回1700 年代，大概不会看到每一栋建筑都像今天仍保存下来的那些建筑一样漂亮。

3 集中趋势

3.1 集中趋势的含义

当我们收集了大量数据时，常常希望用一个单独的代表性数字来概括它们。做到这一点的一种方式，是描述数据的“中间”在哪里。这正是集中趋势的目的：集中趋势的度量是一条规则，它接收整个数据集，并输出一个数字，用来表示某种意义下的数据“中间”。

事实上，并不存在唯一完美的“数据中间”概念。相反，有几种彼此竞争的不同概念，它们在不同情境下会显得更有用或更没那么有用。例如，如果你的数据中含有极端离群值（少数特别大或特别小的值），那么某些集中趋势的度量会发生很大变化，而另一些则几乎不会变化。

3.2 求和记号

在讨论集中趋势之前，我们先要引入一种新记号：求和记号。这只是把一大串数字之和写得更紧凑的一种方式。我们用符号Σ 来表示一大串数的和。

求和记号

$$\sum_{i=1}^{n} x_i \quad \text{表示} \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

字母i 称为指标（它告诉你当前加的是哪一项）。数字1 和n 告诉你从哪里开始数、在哪里结束。这里，n 是数据集中数据点的总数。

3.3 三种常见的“平均数”：均值、中位数、众数

在日常语言中，“平均数”这个词其实有歧义，它可能有多种含义。在基础统计中，最常见的三种平均数是均值、中位数和众数。这三种平均数都是描述数据集“中间”的不同方式，但它们回答的问题

略有不同。

- **均值**：数据的“平衡中心”。如果把每一个数据点想成数轴上的一个重量，那么均值就是平衡点。它使用了每一个数值，因此对极端离群值很敏感。
- **中位数**：把数据排好序之后的中间值。它主要依赖于数据的顺序，因此不会对离群值作出很强烈的反应。当数据偏斜时，它往往是一个很好的“典型值”。
- **众数**：出现频率最高的数值。这一点对于分类数据尤其有用（例如，“最常见的糖果种类”）。对于数值数据来说，众数有时很有信息量，有时则可能误导，这取决于数值的重复方式。

我该用哪一个？（经验法则）

- 如果你关心总量和平衡（例如，公平分摊费用），那么**均值**最自然。
- 如果你想要一个能够忽略极端离群值的“典型”数值（例如，工资），那么**中位数**通常更好。
- 如果你想知道最常见的类别或数值，就用**众数**。

3.4 中位数

中位数是把数据一分为二的那个值（有一半数据不大于它，另一半数据不小于它）。

步骤1: 把所有数据从小到大排序。
步骤2: 如果n 是奇数，取正中间那一个数据（也就是第$(n+1)/2$ 个）。
步骤3: 如果n 是偶数，取第$n/2$ 个和第$(n/2+1)$ 个数据的中间值。

例如，对于数据集$(1,1,3,4,5)$，中位数是3。对于一个含有偶数个数据点的数据集，例如$(1,1,3,4,5,5)$，中位数是两个中间值之间的中点（在这里是3.5）。

3.5 均值

对于一个样本$x_1, x_2, \ldots, x_n$，样本均值写作$\bar{x}$，定义为

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

均值像一个平衡点：高于均值的数值往上拉，低于均值的数值往下拉，而均值正是使整体达到平衡的位置。它可以被看作数据集的一种“几何中心”。

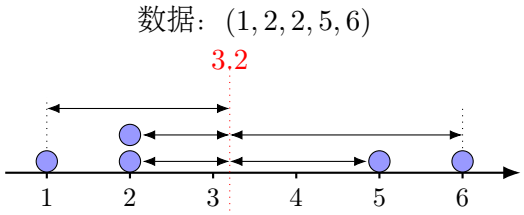
关于均值的一个有用事实

所有数据点到均值的偏差之和总是等于零：

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

因此，总体来看，“高于均值”的数据和“低于均值”的数据会彼此抵消。

为了直观地看到这一点，我们来测量下面图中的距离：



均值左边的三条箭头（均值是3.2）长度之和为$1.2+1.2+2.2=4.6$，而右边两条箭头的长度之和也是$1.8+2.8=4.6$。

3.6 众数

众数是出现频率最高的值。如果存在一个唯一的最高频值，那么这个值就是众数。如果有几个值并列出现次数最多，那么我们说这组数据是多峰的。有些教材会直接说，在这种情况下没有众数。

练习

考虑数据集$(1,1,2,3,4,5)$。

(a) 均值是多少?

(b) 中位数是多少?

(c) 众数是多少?

解答

(a) 总和是$1+1+2+3+4+5=16$，且$n=6$，所以均值是$16/6=8/3\approx 2.67$。

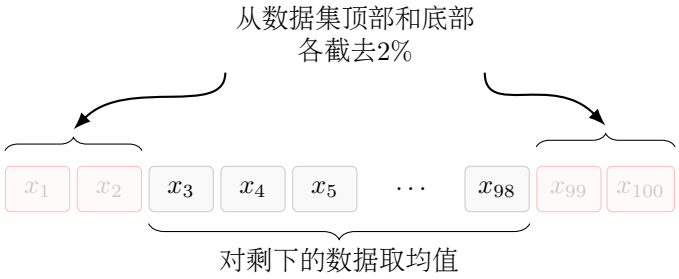
(b) 有序列表已经是$(1,1,2,3,4,5)$。因为$n=6$ 是偶数，所以中位数是第3 个和第4 个数据之间的中点，即$(2+3)/2=2.5$。

(c) 数值1 出现了两次，其余值都只出现一次，所以唯一的众数是1。

3.7 截尾均值

我们可以通过去掉数据集顶部和底部各$x\%$ 的数据，再对剩下的数据取均值，来构造一个“$x\%$ 截尾均值”。这样做的目的是，在保留均值所具有的几何“平衡”含义的同时，也减弱极端离群值的影响。

例如，如果我们有一个已经排好序、大小为100 的数据集，那么一个2% 截尾均值会从数据的两端各去掉两个数，然后对剩下的96 个数求均值。大致如下图所示:



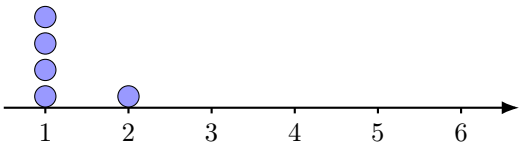
一般来说，当数据带有明显人的因素时，截尾均值很有用。比如，在体育评分中，评委团通常会使用数值评分系统来比较运动员。通常，运动员的分数会通过某种截尾均值来计算，也就是先去掉最高分和最低分，再对其余分数求均值。这样做可以让结果更加客观，也能减少某些容易出错的评委个人偏好所带来的影响。

4 变异

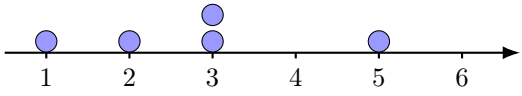
集中趋势大致告诉我们数据“位于哪里”，但它并没有告诉我们数据有多分散。很容易构造出两个拥有相同中心位置、但视觉外观却极不相同的数据集。这就引出了下一个概念：变异（也叫离散程度）。

例如，比较下面两个数据集：

A: $(1, 1, 1, 1, 2)$



B: $(1, 2, 3, 3, 5)$



直观上看，数据集B 更分散，因为它的数值离中间位置更远。

4.1 极差

极差是最大值与最小值之间的距离：

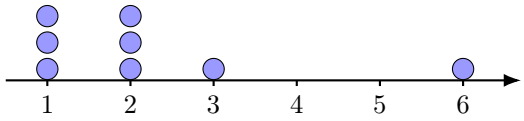
$$\text{极差} = \max(\text{数据}) - \min(\text{数据}).$$

它是对离散程度的一个快速初步描述。极差小通常意味着分散程度较小，极差大通常意味着分散程度较大。

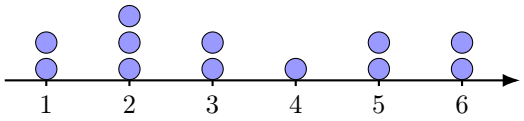
4.1.1 极差的问题

极差只用到了两个数字，也就是最小值和最大值，因此从某种意义上说，这个公式对数据集中绝大多数数据都是“视而不见”的。极差对离群值也非常敏感，因为一个极端数据点就可能让极差变得很大。考虑下面两个数据集：

A: $(1,1,1,2,2,2,3,6)$



B: $(1,1,2,2,2,3,3,4,5,5,6,6)$



注意，数据集A 中的值6 是一个离群值，而B 中则不是。仅从极差来看，这两个数据集的离散程度是一样的。但是，从图上我们可以清楚地看到，A 中绝大部分数据点是聚在一起的，而B 中并不是这样。所以，至少在某种意义上，B 应当比A 更分散。

为了更仔细地刻画数据的分散程度，我们需要一种在公式中使用所有数据点的统计量，而不是只看最小值和最大值。这就自然引向了方差。

4.2 方差

方差是一种比极差更全面、更有信息量的离散程度度量。它通过考察各个数据点到均值的平方距离的平均值来衡量分散程度。如果这些平方距离整体较大，那么数据就更分散。

一种很有用的理解方式是：

> 如果数据点往往靠近均值，那么方差就小。如果数据点往往远离均值，那么方差就大。

4.2.1 核心几何思想

对于每一个数据点x，考察它与均值$\bar{x}$ 的距离，也就是差值$(x-\bar{x})$。如果我们直接对这些偏差$(x-\bar{x})$取平均，那么正值和负值会互相抵消，因此最后得到的结果会等于0。为了避免这种不想要的抵消，我们把这些偏差平方。

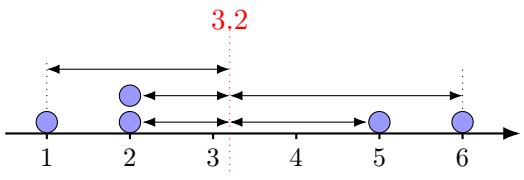
所以，这个基本步骤是：

步骤1: 先求均值$\bar{x}$。
步骤2: 计算每个偏差$(x_i - \bar{x})$。
步骤3: 把每个偏差平方，得到$(x_i - \bar{x})^2$（这样所有值都是正的，而且较大的偏差会被放大）。
步骤4: 对这些平方偏差取平均。

再次强调，第三步非常重要：一般来说，数据点与均值之间的偏差会因为数据点位于均值左边还是右边，而呈现正值或负值：

为了直观地看到这一点，我们再次观察下面图中的距离：

数据：$(1,2,2,5,6)$



正如我们之前看到的那样，均值左边那些箭头的总和，会和均值右边那些箭头的总和互相抵消，因为它们数值相同但符号相反。但把这些值平方之后，这种抵消就消失了。

4.2.2 样本方差公式

对于样本$x_1, x_2, \ldots, x_n$，若其样本均值为$\bar{x}$，则样本方差记作s^2，定义为

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

分母是$n-1$（而不是n），因为我们通常是在利用样本去估计总体的离散程度。在本课程中，你主要需要知道这个公式，以及它的含义：s^2 越大，就表示数据围绕均值的分散程度越大。

在实际中，你通常会用计算器或计算机软件来计算方差，但理解这个公式究竟在测量什么仍然非常重要。不过，为了让你感受一下这个计算过程是如何进行的，我们还是简单演示一下。

例子。 考虑数据集$(1, 1, 2, 3)$。为了计算方差，我们先求均值：

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

接下来，我们要计算每个数据点到均值的偏差，并把结果平方。最容易的展示方式是做成表格：

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

最后，为了计算方差，我们把最后一列加起来，再除以$n-1=3$：

$$s^2 = \frac{0.5625 + 0.5625 + 0.0625 + 1.5625}{3} = \frac{2.75}{3} \approx 0.917.$$

4.3 标准差

标准差就是方差的平方根：

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

我们要取平方根，是因为方差的单位是“平方单位”（例如，如果数据的单位是日元，那么方差的单位就是日元^2）。而标准差回到了原始单位，因此更容易被解释为数据点到均值的一种典型距离。这个观察在下一讲中会变得极其重要。

4.4 记号的重要说明（样本与总体）

正如前面提到的，推断统计中的一个核心问题，是如何利用样本统计量来确定总体参数。因此，非常清晰地区分样本统计量和总体参数是十分重要的。良好实践的一部分，就是对样本和总体使

用不同的数学记号。出于技术原因，样本对应的公式和总体对应的公式略有不同，而更重要的是，我们会用不同的符号来谈论看上去非常相似的概念。下面作一个总结。

样本统计量与总体参数

- **样本统计量：** $\bar{x}$（均值），s^2（方差），s（标准差）。
- **总体参数：** μ（均值），σ^2（方差），σ（标准差）。

4.5 变异系数

有时我们希望比较不同尺度的数据集的离散程度。例如，如果均值是4，那么标准差为2 就很大；但如果均值是200，那么标准差为2 就很小。于是，我们可以通过把标准差除以均值，来“标准化”标准差的相对大小。标准差与均值之比是一个无量纲数，它也是离散程度的度量。这个无量纲数叫作变异系数，定义如下。

$$\text{总体：}\ \mathrm{CV} = \frac{\sigma}{\mu} \qquad \text{且} \qquad \text{样本：}\ \mathrm{CV} = \frac{s}{\bar{x}}.$$

CV 越大，就表示“相对于数值的典型大小而言，变动更大”。

4.6 把集中趋势与方差看作一张“地图”

集中趋势告诉你数据在哪里，而方差（或标准差）告诉你数据分散得有多宽。因此，均值和标准差可以看作描绘数据的两种方式：

$$\text{均值} = \text{中心在哪里}, \qquad \text{标准差} = \text{数据通常离中心有多远}.$$

在下一讲（正态分布）中，这种“地图”式的理解会变得更加具体，因为钟形曲线能让我们把“离均值多少个标准差”直接翻译成近似百分比。

5 一个完整例子

一个好的统计研究，会把现实背景与数学技术结合起来，从而发现关于某个系统的新信息。现在，我们把第1 节中的工作流程应用到课堂上收集到的数据中。

5.1 Haribo 数据（原始样本）

课堂上，我们从19 袋中等大小的Haribo 糖果袋中收集了数据。对于每一袋，我们统计了各种类型糖果各有多少个。下面就是原始数据：

样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
环形糖	3	7	4	2	6	6	7	3	7	4	2	6	6	4	7	5	5	2	6
心形糖	5	4	6	4	5	2	5	7	1	3	6	3	4	6	4	3	2	4	5
瓶子糖	4	1	4	2	4	2	3	3	4	6	4	0	2	4	4	4	8	1	5
蛋形糖	4	5	3	6	0	6	1	2	6	2	4	4	1	2	1	4	3	7	1
小熊软糖	7	8	6	12	11	10	10	9	7	11	6	14	15	8	10	9	6	12	7
每袋总数	23	25	23	26	26	26	26	24	25	26	22	27	28	24	26	25	24	26	24

5.2 总数与平均个数

在这19 袋中一共统计到476 颗糖。每袋糖果总数的平均值是

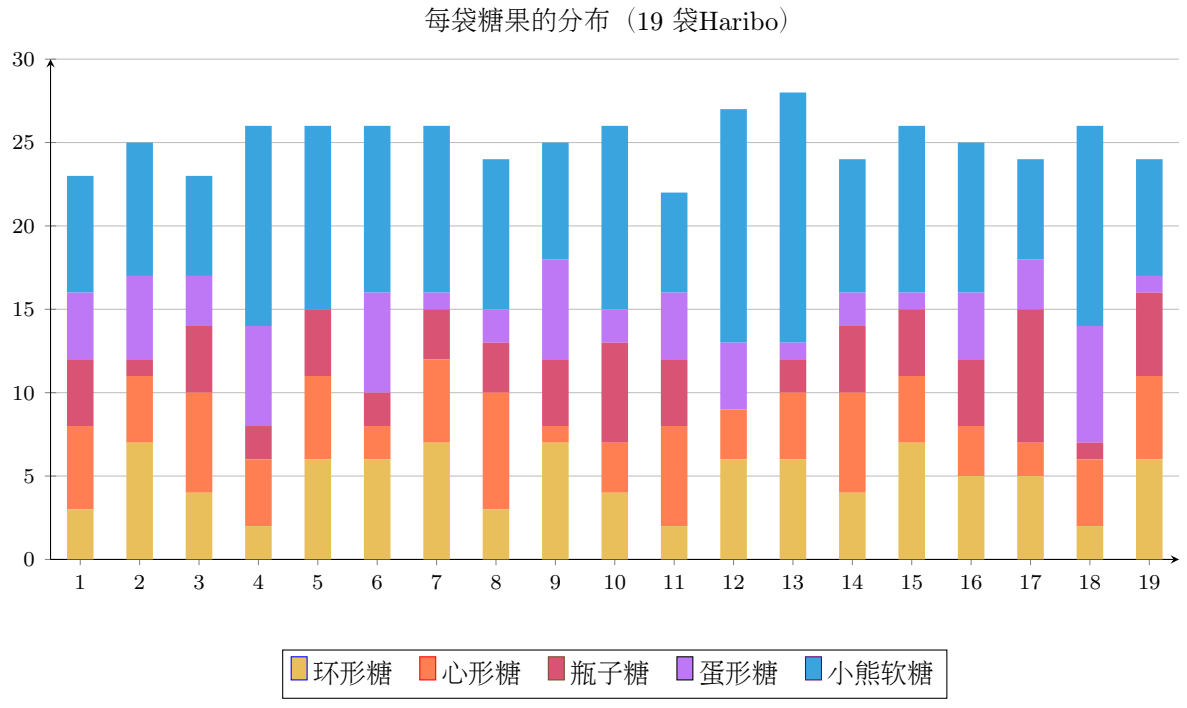
$$\frac{476}{19} \approx 25.05.$$

按种类分别统计总数和每袋平均值，也可以用类似的方法得到：

种类	总数	每袋平均值	平均值（近似）
环形糖	92	92/19	4.84
心形糖	79	79/19	4.16
瓶子糖	65	65/19	3.42
蛋形糖	62	62/19	3.26
小熊软糖	178	178/19	9.36

5.3 糖果分布

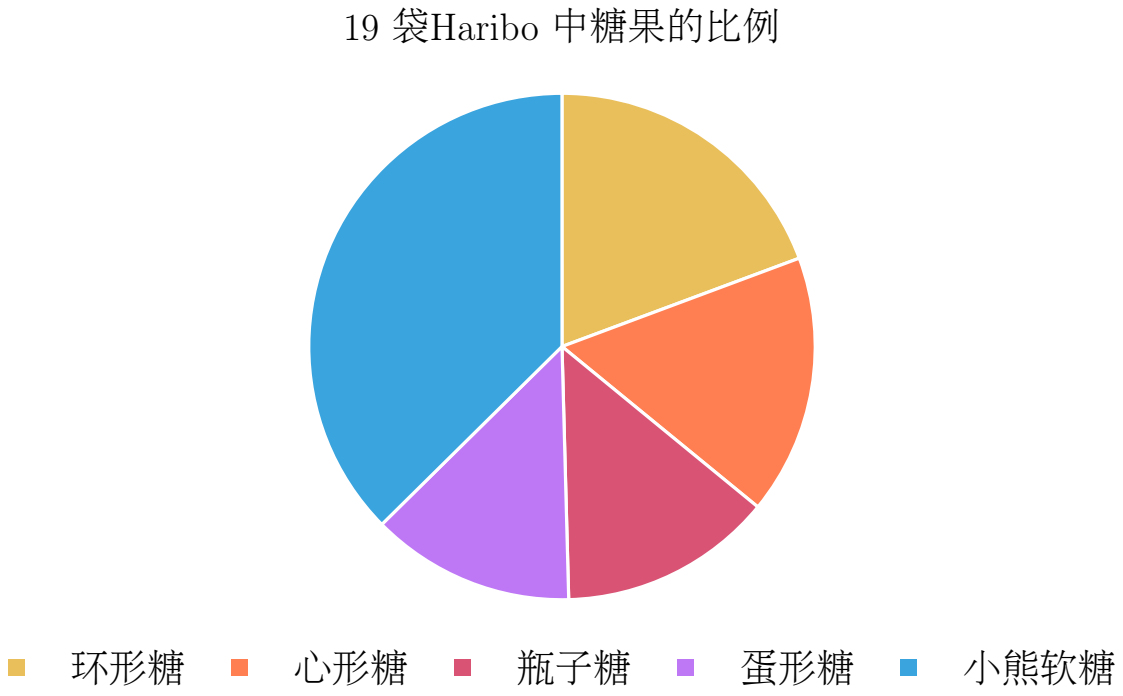
利用原始数据，我们可以把每袋中的糖果分布画成下面这个大图：

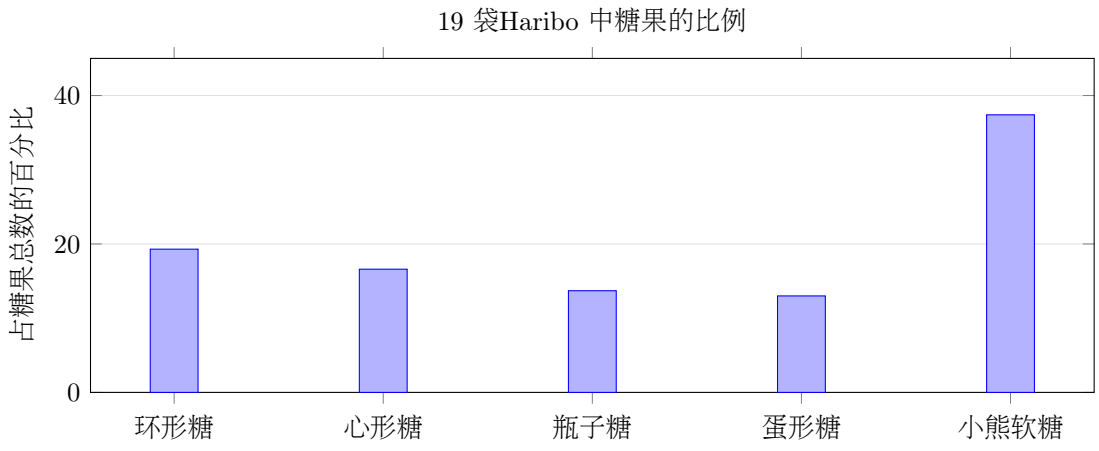


看起来，每袋糖的总数都相当稳定，大约在25 左右。然而，每袋内部糖果种类的分布似乎变化很大。

5.4 糖果比例（相对频率）

种类	比例	百分比（近似）
环形糖	92/476	19.3%
心形糖	79/476	16.6%
瓶子糖	65/476	13.7%
蛋形糖	62/476	13.0%
小熊软糖	178/476	37.4%





5.5 Haribo 计数中的离散程度与变异

类别	均值	方差	标准差	CV
心形糖	4.16	2.47	1.57	0.38
蛋形糖	3.26	4.32	2.08	0.64
瓶子糖	3.42	3.48	1.87	0.55
小熊软糖	9.36	7.13	2.67	0.29
环形糖	4.84	3.25	1.80	0.37
每袋糖果总数	25.05	2.27	1.51	0.06

每袋糖果总数的变异很小，但与之相比，各类糖果的变异系数则相当高。这支持了我们之前的观察：袋子大小似乎相当统一，但内部组成变化明显。

5.6 如何在Excel 中快速计算这些量

概念	Excel 命令
单元格求和	`=SUM(...)`
样本均值	`=AVERAGE(...)`
样本标准差	`=STDEV.S(...)`
中位数	`=MEDIAN(...)`
众数	`=MODE.SNGL(...)`
变异系数（CV）	`=STDEV.S(...)/AVERAGE(...)`

5.7 结果解释（三个结论）

我们对19 袋Haribo 样本的研究表明，有三个主要发现：

1. 在这个样本中，各类糖果所占比例并不接近20%。特别是，小熊软糖似乎比其他种类更常见。
2. 每袋糖果的大小通常比较一致，大约在每袋25 颗左右。
3. 不同袋子中糖果的混合组成变化相当明显。

5.8 可能的抽样偏差

任何一个好的统计研究，至少都应当考虑研究过程中可能出现的偏差。对我们来说，一个合理的总体，是在我们抽样期间、当地（例如日本）出售的所有中等大小Haribo 袋装糖，而我们从中抽取了19 袋作为样本。下面列出一些我们应当意识到的可能偏差。

- 我们抽到的19 袋糖，可能来自同一个生产批次。
- 我们无法确定每个人记录数据时是否都完全正确。
- 在把数据输入Excel 时，可能会出现录入错误。
- 我们不知道生产这些糖果的本地工厂，是否与其他工厂相同。

- 我们也不知道这一特定批次的Haribo 是否存在本地生产误差。
- 我们不知道Haribo 的生产标准是否会因国家而异，而且即使这种差异存在，我们也没有记录这些样本究竟是哪个国家生产的。

5.9 用工作流程语言总结这项研究

- **收集**：学生们把数据写在纸上。
- **整理**：Excel。
- **分析**：平均个数、变异程度和图表。
- **解释**：这个样本对典型袋装大小和糖果组成的启示是什么。
- **呈现**：图表和总结性陈述。

5.10 接下来要做什么

接下来，我们将发展一些技术，用它们来检验这些结论，并进一步推断所有Haribo 袋装糖总体的信息。