利用可视化为数据分布拟合钟形曲线

Eric Newburger、Michael Correll 和 Niklas Elmqvist，*IEEE* 高级会员

摘要—正态曲线或其他曲线等理想化概率分布，是验证性统计检验的基础。但人们对这些理想化曲线的理解有多深入？具体而言，人类视觉系统能否让我们将样本数据分布与假设的总体分布进行匹配，而这些样本可能正是从该总体分布中抽取的？不同的可视化技术又会如何影响这种能力？本文介绍了一项众包实验的结果，该实验测试了受访者为四种不同的数据分布可视化拟合正态曲线的能力：条形图直方图、点图直方图、条带图和箱线图。我们发现，众包参与者能够较为成功地估计分布的中心（均值），且偏差较小。我们还发现，人们通常会高估标准差——我们将此称为“雨伞效应”，因为人们往往想用曲线覆盖整个分布，仿佛要为其遮挡上方的天空——而条带图的准确度最高。

索引词——图形推断、视觉统计学、目测统计、分布拟合、众包。

arXiv:2301.04717v1 [cs.HC] 11 Jan 2023

1 引言

许ANY 多重要的可视化任务不仅依赖于对单个数值的评估，还依赖于对整体分布的全面评估：识别离群值、判断变异性、评估各种参数化统计检验是否适用，甚至只是构建一幅关于哪些数值可能出现、哪些不太可能出现的图景，都需要对样本数据进行显式或隐式的分布拟合。以往的研究考察了总体或集成层面的图形感知任务 [1],，但我们认为，对理想化分布曲线进行拟合这一具体任务尚未得到充分研究，而且有望为统计图形的设计提供参考，尤其是面向新手用户的设计。我们选择将正态分布拟合到样本数据上，作为图形感知的“果蝇” [2]：这是一个相对简单且广为理解的问题，但对统计学中的多种任务仍然十分重要。

我们的研究源于可视化领域的一个开放问题：人类能在多大程度上依据统计图形作出视觉判断，并将其作为一种“视觉统计学” [3] 或“图形推断” [4]；也就是说，我们在多大程度上能够依靠读取或汇总图表信息的能力，来评估效应量、为反对零假设提供证据，或判断数据中的趋势。如果这些能力足够稳健，就有可能借助视觉评估增强甚至取代正式统计检验。然而，也有证据表明，在某些方面，对统计属性的视觉估计并不充分、与正式统计检验并不类似，或以其他方式与其结果脱节。偏差 [5], 、满意化策略 [6], 和感知代理 [7], [8]可能导致对数据的统计评估与人类判断或估计不一致。因此，在推广将统计图形用作推断工具或验证工具之前，有必要对人类表现进行实证评估，并测量人类在读取统计图形时的偏差。

本文介绍了一项预注册的众包用户研究结果，该研究考察普通人群在以多种图形形式呈现数据分布时，拟合正态曲线的能力。每次试验中，参与者都会看到数据的可视化，并被要求调整叠加在样本上的高斯曲线的中心点（均值）和宽度（离散程度或标准差），以获得尽可能好的拟合。研究的可视化形式包括条形直方图、Wilkinson 点图直方图、条带图和箱线图（图 1）。

2 相关工作

除了用于交流之外，可视化也常被推崇为一种主要用于探索性数据分析 [9] 的工具，这源于其探究性和数据驱动的特性。然而，可视化也可用于某些形式的验证性数据分析 [10], [11], , 至少可作为一种补充。我们将详细回顾这些主题。

2.1 图形推断

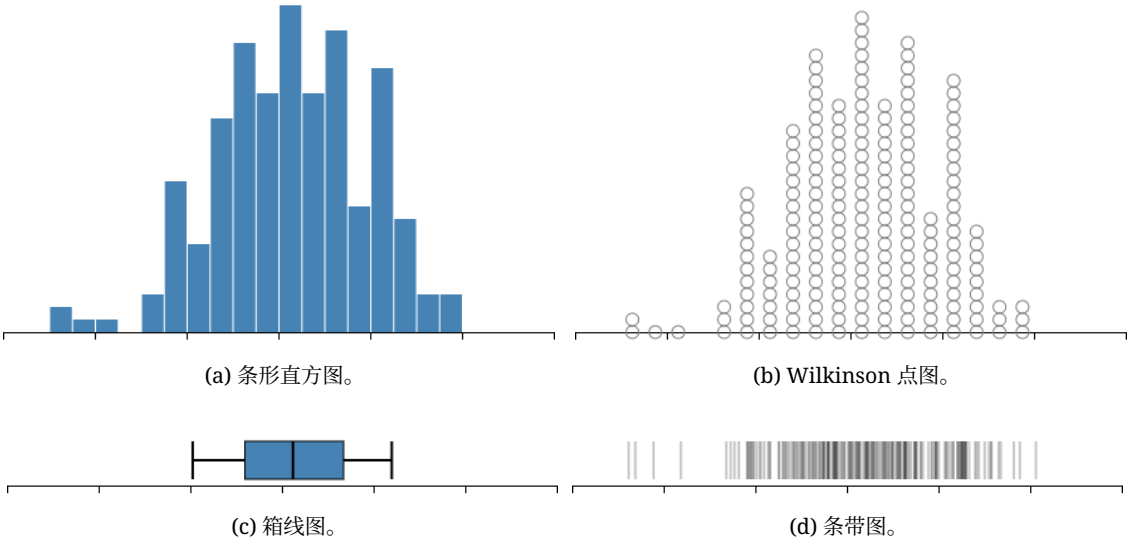
创建数据的图形表示是统计工作流程中常见且自然的一部分 [12], , 甚至是某些工作流程的核心，例如探索性数据分析 [9]。因此，根据这些图形表示进行推断——即图形推断——是对正式统计检验的常见补充。这种做法的早期实例可追溯至 Scott 等人 [13]：他们通过使用模型参数生成模拟星图，并让人们将其与真实图表进行比较，来验证天文模型。

然而，直到最近，学界才开始探讨数据的图形表示如何支持更高阶的任务，而不只是读取数值、趋势和离群值。Buja 等人 [4] 提出了视觉统计学框架，其中可视化表示提供检验统计量，而人类认知则充当统计检验。他们通过对随机数据进行“罗夏墨迹”测试，以及使用由多个小多图组成的排列图（其中只有一个使用真实数据），展示了这种方法。在后续研究中，Wickham 等人 [14] 将这一理念应用于可视化

• *Eric Newburger* 和 *Niklas Elmqvist* 就职于美国马里兰州大学公园市的马里兰大学帕克分校。电子邮箱：*enewburg@terpmail.umd.edu*，*elm@umd.edu*

• *Michael Correll* 任职于美国华盛顿州西雅图的 *Tableau Research*。电子邮箱：*mcorrell@tableau.com*

稿件于 *2021*年 *XXX XX* 收到；修订于 *2* *021.*



(a) 条形直方图。 (b) Wilkinson 点图。

(c) 箱线图。 (d) 条带图。

图 1：**实验设置**。使用我们在实验中研究的四种不同可视化表示形式呈现的一维数据集。在实验中，参与者在这些表示形式上拟合正态曲线。

学界，介绍了如何将这些协议与常见可视化结合使用，以发现新的结论，同时避免假阳性。Beecham 等人 [15] 将这种用于图形推断的“排列图”协议应用于地理聚类可视化。Correll 等人 [16]则利用它探究常见的分布图形是否能有效展示离群值或间隙。

此外，也有研究探讨人们从可视化中提取聚合统计量的能力。Correll 等人 [17]研究了如何设计折线图和其他可视化，以便准确比较时间序列数据中的平均值。Albers 等人 [18] 将这一思路推广到六种聚合任务和八种不同的时间序列可视化。此外，Aigner 等人 [19]为折线图添加颜色，以更好地支持视觉统计学；Fuchs 等人 [20] 则设计了线形字形，以支持更高层次的聚合任务。Correll 和 Heer [21]通过一项众包实验，研究了人们在双变量可视化中拟合趋势线的能力。

从其他可视化表示中推导聚合统计量也与我们的研究相关。Fouriezos 等人 [22] 要求参与者比较两组条形图的平均高度，结果显示准确度较高；条形数量增加会提升准确度，而方差则会降低准确度。Gleicher 等人 [23] 研究了多类别散点图中的均值判断，发现无论点的数量多少，也无论是否存在冲突编码，表现都稳定地处于较高水平。基于众包实验的结果，Correll 和 Gleicher [5] 提出了对条形图中误差条的重新设计，并展示了小提琴图或渐变图如何产生与统计推断更加一致的洞见。最后，Nguyen 等人 [24] 通过一项众包实验，探讨不同的视觉聚合方式会如何影响用户对样本总体摘要陈述的感知。他们发现，视觉聚合策略不会影响参与者的准确度，但指出，看到完整数据的参与者信心更低，也较少陷入二分思维。这表明，分解式可视化能更丰富、更细致地呈现底层数据。

2.2 集合加工

我们研究的一个核心方面，是人们如何通过可视化来直观估计集合特征。视觉研究中的一个常见趋势是使用抽象点簇，因为这样可以精确控制视觉刺激 [25]。Morgan 和 Glennerster [26]研究了人们对这类点簇质心的感知，发现其准确度非常高，但个体之间存在一些差异。Ariely 证实了这些结果，并发现，即使人们很快就会忘记具体细节，他们通常也能相当准确地估计一组点的均值和范围 [27]。这表明，视觉系统提取的是集合的统计特征，而非个体细节。

自然地，人类对几何形状的感知可能与对随机点簇的感知不同。Melcher 和 Kowler [28]研究了人们在由这些点构成的轮廓形状上进行质心估计时的眼球运动（扫视）。他们发现，即使面对偏斜分布的点以及干扰点簇，人们通常也很擅长确定目标形状区域的中心点。这表明，引导感知的是形状本身——也就是它的轮廓——而非单个点基元。

当我们的知觉系统只能感知少数几个实体时，我们又是如何从许多个别的点感知到完整形状的呢？答案或许在于集合加工 [29], , 它探讨视觉系统如何计算多种视觉特征的平均值，以处理复杂的形状和构型。对于现实世界中的真实物体，这种加工更容易，因为它们往往形状规则；而人工点簇则不然。这种加工也称为知觉平均。Chong 和 Treisman [30] 展示了一项实验的结果，表明参与者甚至能对由 12 个形状组成的点簇进行大小平均。Albrecht 和 Scholl [31] 将这些结果扩展到了动态变化的显示，得出了类似的结果。

在可视化领域，已有一些研究探讨这些现象。与我们的工作最相关的是，Szafir 等人 [1]展示了集合加工如何支持统计图形中的更多相关任务，包括对数值进行概括 以及对结构进行估计 。然而，随着视觉复杂度的提高，精确度往往会下降。Yuan 等人 [8] 发现，在多值比较过程中估计可视化中形状的图形属性，往往会归结为简单的知觉线索，其准确度低于感知单个数值时的准确度。这些知觉线索，即代理线索 [7],

[8], 在视觉系统需要并行比较多个形状时提供了捷径。相比之下，本文的工作聚焦于一项更全面的任务：在单个数据分布可视化中拟合理想化钟形曲线的均值和离散程度。

2.3 分布可视化

单变量分布最常见的可视化方式之一是直方图，它将数据出现次数汇总到离散范围（“分箱”）中，并用条形图展示汇总后的数量。然而，直方图存在一些缺陷，其中大多数与分箱大小和分箱数量有关 [16]。其他表示方式包括条带图、密度图、小提琴图和渐变图 [32], ，但这些方式不像直方图那样容易让人熟悉。即便不考虑分箱方面的问题，直方图汇总数据的特性既是优点也是缺点：优点在于，这种表示方式能够稳健地可视化大型数据集；缺点在于，条形图传达的是各分箱中案例的相对数量，而非绝对数量。想知道绝对人数的用户必须查看坐标轴——这需要进行读取，而不只是观察。因此，Wilkinson 点直方图 [33], 可能优于传统的条形图设计 [16], [34]；在这种图中，分箱内的每个离散项都以圆点的形式堆叠呈现。

不确定性可视化是一个密切相关的主题，因为我们通常会将不确定性计算为点估计周围潜在方差的理想化分布，而近年来，在设计让不确定性成为表示本身内在属性的可视化技术方面，已取得显著进展。例如，这种方法在使用所谓的分位数点图 [32]估计结果时，显著提升了置信度。假设结果图（HOP） [35] 通过动画抽样来展示不确定性，其表现已证明优于小提琴图和误差条。

最后，本文的方法受到近期不确定性可视化研究的启发。这类研究通过图形引导 [36]，让参与者在评估过程中绘制数据的可视化表示。首先，这样做可以提高回忆和理解；Kim 等人 [37] 发现，以图形方式引导参与者呈现其先验知识和观察到的数据，有助于他们记忆并据此进行推理。在后续研究中，Hullman 等人 [34] 让参与者在看到实际分布之前，分别使用连续和离散表示形式绘制对不确定性分布的预测。研究结果表明，这种图形引导也能改善参与者的预测。最后，Kim 等人 [38] 通过引导参与者表达先验和后验信念，构建贝叶斯认知模型，以解释人们如何解读数据可视化。我们的研究采用了类似的引导方法，让参与者为可视化拟合连续曲线的图形表示；但由于我们的分布为高斯分布，因此只要求他们估计均值和离散程度。

3 研究：拟合钟形曲线

我们的研究旨在了解未经训练的人能否（以及能在多大程度上）将正态曲线拟合到从正态分布中抽取的数据样本。为初步探究这一问题，我们在亚马逊土耳其机器人上开展了一项众包用户研究，请参与者调整高斯曲线的位置（均值，即“中心点”）和宽度（标准差，即“离散程度”），使其拟合数据样本。屏幕上以四种不同的可视化形式呈现样本：条形直方图、点状直方图、条带图和箱线图。这些形式的多样性降低了任何一种分布图类型的特征对实验结果产生偏差的可能性，也使我们能够据此提出创建有效数据分布图的设计建议。我们的设计改变了随机样本的大小（$n = 50$ 或 200）、数据中的噪声（变异系数 = 0.2 或 0.4）以及数据样本的视觉表示。为尽量减少“从左到右偏差”的可能性， [39] 即受访者养成总是在屏幕上从左向右移动曲线的习惯，我们调整了部分数据样本的均值，使其中心点在屏幕上向左移动（数值低于零）。这些移动不会影响数据的形状，变异系数（CV）的计算则基于数据集的原始位置。

本研究的原始版本仅包含 Wilkinson 点直方图，并已在 OSF 上预注册。[1] 自最初的研究以来，我们扩展了实验，加入了三种额外的可视化表示：传统条形直方图、条带图和箱线图。新的预注册可在 OSF 上查阅。[2] 以下讨论仅涉及新实验；本文不包含 200 名参与者的原始数据，因此也不会报告这些数据。下文将先回顾我们的方法，随后介绍结果。

3.1 参与者

由于本研究关注的是无需特定训练或数据可视化专业知识的低阶感知任务，我们使用亚马逊土耳其机器人（MTurk）开展研究。使用 MTurk 意味着我们无法对参与者的人口统计学特征、专业知识以及计算机硬件进行多少控制，但先前的研究表明，简单的视觉任务尤其适合采用这种众包研究方式 [40]。

在我们的实验中，所有实验因素均为被试内因素。我们计划招募共100名参与者。由于机构审查委员会施加的税务和补偿限制，我们将参与范围限定在美国。我们对参与者进行了筛选，以确保他们至少具备英语的实用能力；这是理解测试平台上的说明和任务描述所必需的。我们禁止参与者重复参加实验。我们按照至少相当于每小时10美元时薪的标准向所有参与者支付了符合伦理规范的报酬（2020年美国联邦最低工资为每小时7.25美元）。具体而言，每次实验支付2.50美元；按典型的14分54秒完成时间计算，时薪约为每小时10.00美元。

3.2 实验设备

由于采用众包形式，我们无法控制参与者使用的具体计算机设备。研究通过用户的网页浏览器进行。我们要求所有设备均为个人电脑（笔记本电脑或台式电脑）或触控平板电脑；由于智能手机的屏幕空间有限，因此不允许使用智能手机。我们还要求参与者使用窗口尺寸至少为 1280×800 像素的浏览器窗口。

3.3 任务

这些任务包括：使用范围滑块 [41]，通过区间的宽度控制曲线的离散程度（即标准差），将正态（高斯）曲线拟合到数据可视化上

1. https://osf.io/behwz
2. https://osf.io/a9b48

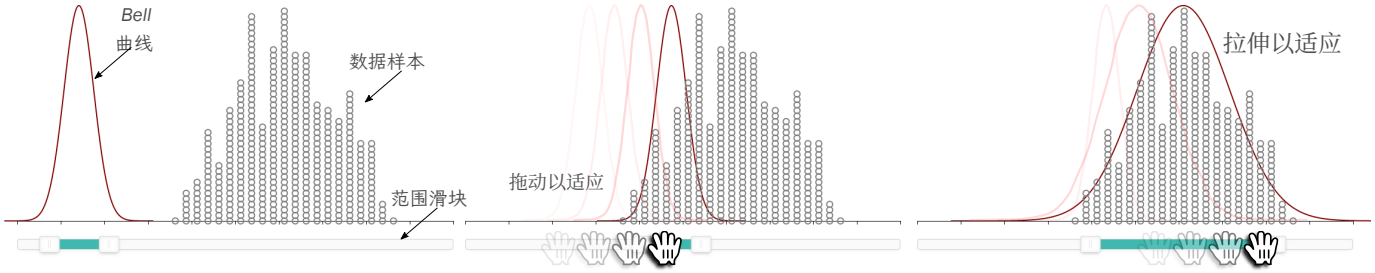


图 2：**曲线拟合任务**。众包曲线拟合实验中的典型流程。参与者使用范围滑块控制正态曲线，将这条连续曲线拟合到数据样本上（此处以 Wilkinson 点图直方图表示）。我们的评估改变了可视化方式，以及样本的数据点数量和变异系数。

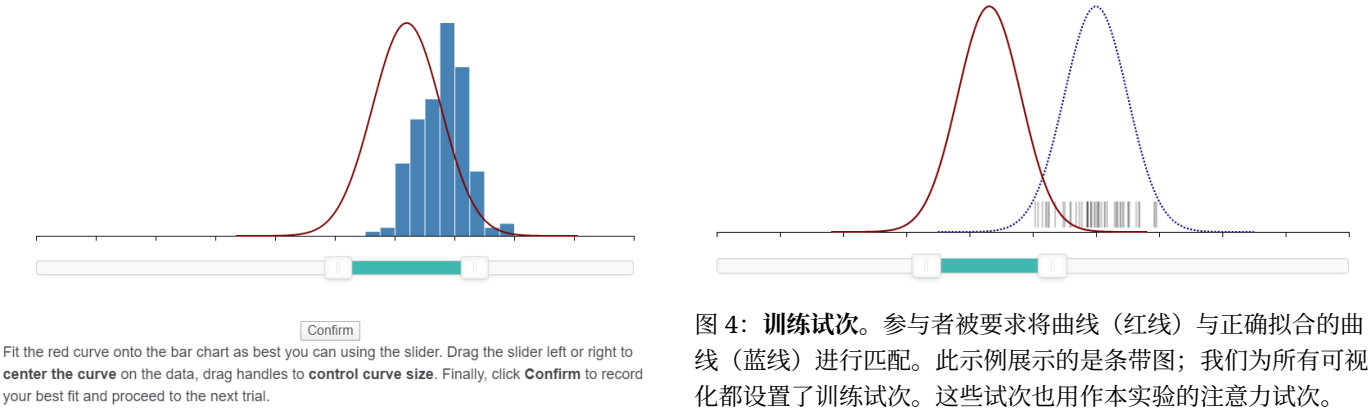


图 3：**曲线拟合任务（条形直方图）**。用户控制红色高斯曲线。条形直方图表示待拟合的样本。坐标轴下方的范围滑块通过范围宽度控制离散程度，并通过滑块上的位置控制中心点。参与者对拟合结果满意后，点击“确认”按钮完成操作，然后继续下一次试验。

并通过移动滑块上区间的位置来调整其中心点（即均值）。图3展示了一个典型任务的截图。参与者被要求在曲线与可视化数据之间找到“最佳拟合”。在每个可视化区块的训练试次中（我们也将其用作注意力试次；详情见下文），参与者会看到一条使用对比色曲线显示的完美拟合，并被要求将自己的曲线调整为正确答案（图4）。

测试平台使用 JavaScript 和 D3 [42] 实现，并嵌入 Qualtrics 问卷中，参与者通过网页浏览器访问该问卷。我们使用 noUiSlider[3] 实现范围滑块。参与者可以移动范围滑块的一个端点来移动滑块中心，也可以拖动范围的中心区域，同时移动两个端点。换言之，参与者能够直接控制曲线的离散程度和均值。不过，参与者无法直接拖动曲线的视觉表示。

3.4 数据集生成

所有试次的数据都经过控制，确保所有参与者在一次会话中看到相同的数据集。所有数据集均随机抽取

3. https://refreshless.com/nouislider/

图 4：**训练试次**。参与者被要求将曲线（红线）与正确拟合的曲线（蓝线）进行匹配。此示例展示的是条带图；我们为所有可视化都设置了训练试次。这些试次也用作本实验的注意力试次。

自正态分布，其离散程度各不相同，然后反复添加抖动，直到标准差与离散程度的预期值相差不超过 1%S （见下文）。生成的数据集均落在 $[-3,7]$ 范围内，水平轴固定为 $[-10,10]$。不过，为尽量减少数字偏差，坐标轴上没有显示数值标签。直方图的水平轴上划分了 50 个分箱，但实际试次通常总共使用 4 到 8 个分箱（具体取决于离散程度）。

我们没有引入任何取自正态分布以外分布的数据集，因为本实验的目的是拟合正态曲线，而不是让参与者判断应使用哪种分布。尽管如此，通过改变实验因素（如下），我们仍能得到外观上足够嘈杂的分布。不过，这意味着我们的数据集都或多或少呈对称分布。

3.5 实验因素

我们选择在实验中对三个因素进行建模：

- **数据量 (D)**：拟合数据集中的样本数量。随着样本数量增加，直方图会逐渐接近理想化分布。我们选择了两个水平：50 个和 200 个样本。虽然在实践中，人们可能希望拟合更小的样本量，但我们选择 50 个项目作为最小水平，因为我们希望通过图形推断支持的学生 t 分布 [43]——只有在样本较大（30 个项目或更多）时才开始符合正态分布。

- **离散程度 (S)**：所拟合高斯分布的标准差。我们选择了该因素的两个水平，

以变异系数 (CV)（或相对标准差）表示，即标准差σ与均值 μ (σ/μ)的比值：20% 和 40%。我们通过先导测试确定了这些数值，以确保样本既有相对较大的噪声（S值较高时），也相对集中（S值较低时）。由于我们的数据生成过程具有随机性，只有在 1% 的容差范围内，才能保证数据集具有特定的离散程度。

- **可视化** (V)：用于表示数据集的可视化类型。根据对相关文献的回顾，我们选择了四种各具特色的可视化技术（图 1）：

  – **条形直方图：** 一种“经典”直方图，使用宽度统一的条形来表示每个分箱中汇总的数据项数量（图 1a）。

  – **Wilkinson点直方图：** Wilkinson 最初提出的一种直方图变体 [44], 。点状直方图是单位可视化 [45], 将代表各个项目的独立点（圆形）排列成条形，每个分箱对应一个条形（图 1b）。

  – **箱线图：** 由 John W. Tukey 开创的一种箱线图 [46], , 其中中央矩形包含数据中间一半的部分（从第 25th 百分位数到第 75th 百分位数），中位数（第 50th 百分位数）用一条线标出，而“须线”则标示更宽百分位范围的边界；在本例中，须线标示数据上端 10% 和下端 10% 的边界（即第 10th 和第 90th 百分位数）（图 1c）。本研究中的箱线图**未**添加点或星号等图标来标示离群值。

  – **条带图：** 一种单元可视化 [45] ，其中每个项目都绘制为一条短竖线，并在水平轴上使用不透明度（即不进行垂直数据编码），从而得到一种类似条形码的表示（图 1d）。

虽然还存在其他可视化表示，但我们认为这几种足以代表数据聚合程度的不同范围：条带图绘制所有数值，条形图和点状直方图将它们分箱，而箱线图则舍弃所有数值，转而呈现样本的描述性统计。

此外，分箱数量是直方图的一个重要参数 [16]。我们选择不直接对这一因素建模，而是将水平轴的范围固定为 $[-10, 10]$，并将分箱数量固定为 50，从而根据离散程度间接控制所使用的分箱数量S。

3.6 实验设计

我们采用被试内设计，每位参与者都会看到所有数据量、离散程度和可视化。条件总数相对较少，因此我们能够将每个实验环节控制在 20 分钟以内，以尽量减少疲劳并提高众包工作者的注意力。每位参与者的试次顺序均随机排列。由此形成了以下设计：

2 **数据量** D(50,200 个样本)× 2 **离散程度**S (0.2,0.4)× 4 **可视化**V (条形图、点、方框、条带)× 3重复次数 48 次试验/每位参与者

我们计划从 100 名参与者那里总共收集 4,800 次试次。[4] 对于每次试验，我们都记录完成时间和准确度。完成时间从试次呈现给参与者时开始计时，直到参与者提交答案为止。准确度基于两项指标：

- **均值误差：** 如果 μ_c 是参与者拟合的正态曲线的均值，而 μ_s 是样本的实际均值，则我们将均值误差计算为 $\mu_c - \mu_s$。在大多数分析中，我们取该指标的绝对值。
- **标准差误差（%）：** 如果 σ_c 是参与者拟合的正态曲线的标准差，σ_s 是样本的实际标准差，则标准差误差的计算方式为 $(\sigma_c - \sigma_s)/\sigma_s$，并以百分比表示（由于该指标与屏幕分辨率无关，因此使用比率能提供更多信息）。在大多数分析中，我们取该指标的绝对值。

3.7 流程

所有招募工作均通过亚马逊土耳其机器人进行。符合资格标准的参与者在单独的浏览器窗口中打开问卷。参与结束时，他们将唯一的完成代码复制回 Mechanical Turk 界面；待其工作经过核查后，便会获得报酬。

每次参与开始时，参与者都需要填写一份免签名的知情同意书。未提供同意即会终止实验。参与者被告知，他们可以随时退出本次参与。遗憾的是，我们无法向只完成部分参与的参与者支付报酬。开始参与时，我们已在知情同意书中告知参与者这一情况。

签署知情同意书后，参与者被要求回答有关年龄、教育程度和统计学概念知识的人口统计学问题。我们还请参与者再次确认，他们是使用平板电脑或计算机参与研究的。随后，参与者会看到每种可视化类型对应的练习试次，并获得有关如何阅读可视化内容和完成任务的说明。在每个练习试次中，都会以对比蓝色显示一条拟合正确的曲线。这些练习试次也作为“注意力试次”。设置注意力试次的目的是剔除未认真关注任务的众包工作者的回答。对于这些注意力试次，只要参与者的回答与实际均值之间的误差超过3个标准差，该参与者的整个会话就会被排除在分析之外。我们在知情同意书中说明了这一点。

每个试次都从显示数据集和曲线（图3）开始，并在参与者点击“确认”按钮时结束。参与者至少操作一次范围滑块后，才能确认试次。完成时间从试次显示时开始计量，直到点击该按钮为止。我们告知参与者，如果需要在试次之间休息，可以利用可视化区块之间的间隔。屏幕顶部的进度条显示研究进度。

典型的实验环节持续 14 到 15 分钟。少数参与者完成实验环节花费的时间要长得多，但我们的日志显示，这些参与者在试次之间休息了很长时间（可能是因为受到干扰）。我们认为，实际用于实验的时间不到 15 分钟。

3.8 假设

我们预注册了关于本次实验的以下假设：

- 均值估计将比离散程度估计更准确。我们基于自身经验以及文献综述形成的直觉（例如，人们能够以较高准确度直观地

4. 请注意，由于研究团队内部沟通失误，我们的预注册计划招募 100 位参与者，但最终招募了 150 位。我们将在关于预注册偏离情况的部分对此进行讨论。

确定平均值） [23] 认为，人们能够以较高准确度确定均值，但在样本上拟合曲线的准确度会较低。

- 随着点的数量增加，参与者对均值和离散程度的估计都会更加准确。对于较大的数据集，抽样误差的影响会更小，分布的整体形状也会更加清晰。这应有助于更准确地通过感知估计均值。
- 不同可视化类型的表现会有所差异。具体而言，我们预计：
  - 参与者使用箱线图估计均值时会更加准确。箱线图在视觉表示中直接编码了分布的中位数，而在我们的大多数刺激材料中，中位数接近或等于均值。– 参与者使用条带图估计均值时会不那么准确。利用不透明度和过度绘制的影响来编码密度，使得精确估计变得困难。
- 除箱线图外，参与者会始终高估分布的离散程度。根据先前关于集合加工和视觉统计任务中使用“感知代理”的研究 [7], ，我们预计参与者会尝试覆盖理想化曲线的密度主体内所有可见标记。这会导致参与者高估所有可视化类型中的离散程度；但箱线图除外，因为覆盖中央矩形——其中仅包含 50% 的数据——会导致低估。

4 结果

我们使用自助法分析了研究结果 [47]（$N = 1{,}000$ 次重复），以计算 95% 置信区间（置信区间） [48]。我们的汇总计算未对因素间交互作用进行调整；关于交互效应的分析，请参见第 4.2 节。我们根据这些区间报告效应量。

我们收集了 146 名参与者的数据，他们完成了 48 次试次，总计 7,008 次单独试次。剔除未通过四次注意力试次的 19 名参与者后，还剩 127 名参与者。检查数据后，我们发现另有 10 名参与者似乎在实验的至少一个完整区块中误解了曲线拟合任务。具体而言，这些参与者移动了曲线的位置以拟合均值，却没有改变曲线的宽度以拟合离散程度。我们推测，出现这一问题是因为训练试次未要求受访者调整离散程度。现在回想起来，我们本应要求参与者先改变曲线的离散程度，才能提交试次（我们当时只要求他们以某种方式与滑块互动）。由于我们无法评估这种误导性训练的影响，因此决定在分析中剔除这 10 名参与者的所有数据。最终，参与者总数为 117 名。

根据我们的预注册，我们剔除了完成时间高于 3σ 的离群试次（而非参与者）；总共剔除了 79 个试次（即所有试次的 1.3%）。我们认为，这些试次反映了参与者在试验过程中途受到干扰的情况；其中大多数持续了数百秒（最长达 698 秒）。我们认为，根据完成时间剔除此类试次是合理的，原因有二：(a) 使用在线众包进行数据收集，远不如在实验室环境中可控，因此需要对参与者注意力不集中、延迟和外部干扰的情况作出调整 [40],；(b) 我们的假设均不基于完成时间。

剔除离群值后，最终数据集包含 5,527 个试次。总体平均完成时间为 10.5 秒（标准差为 7.99）。总体绝对平均均值误差为 11.2%（标准差为 25.2%）。总体绝对平均标准差误差为 28.9%（标准差为 30.4%）。下文将分析参与者表现，并详细探讨不同因素的特征。

4.1 平均值与个体分析

图5展示了均值估计误差和标准差误差在各种离散程度、数据量和可视化类型组合下的数据分布。可以看出一些明显的趋势。多种可视化方式似乎会使参与者倾向于高估标准差。箱线图是个例外，它似乎反而会导致低估。图 6 清晰地展示了这种高估现象，其中显示了两种特定试次的中心化拟合曲线，分别代表“紧凑”（尖峰状）和“松散”（较平坦）的分布。

图7比较了个体的表现。每个点代表一位受访者在该图表类型的所有试次中的平均误差。已从这些数据中剔除大于 3σ 的离群值。参与者在估计标准差方面的表现差异相当显著。图中还显示了不同视觉表示两两之间表现的绝对皮尔逊 r 相关值。对于绝对均值误差，箱线图与条带图的相关性似乎最高（$|r| = 0.31$），其次是点图与条形直方图（$|r| = 0.29$）。另一方面，直方图的表现与条带图或箱线图的表现似乎没有明显相关性（$|r| < 0.13$）。

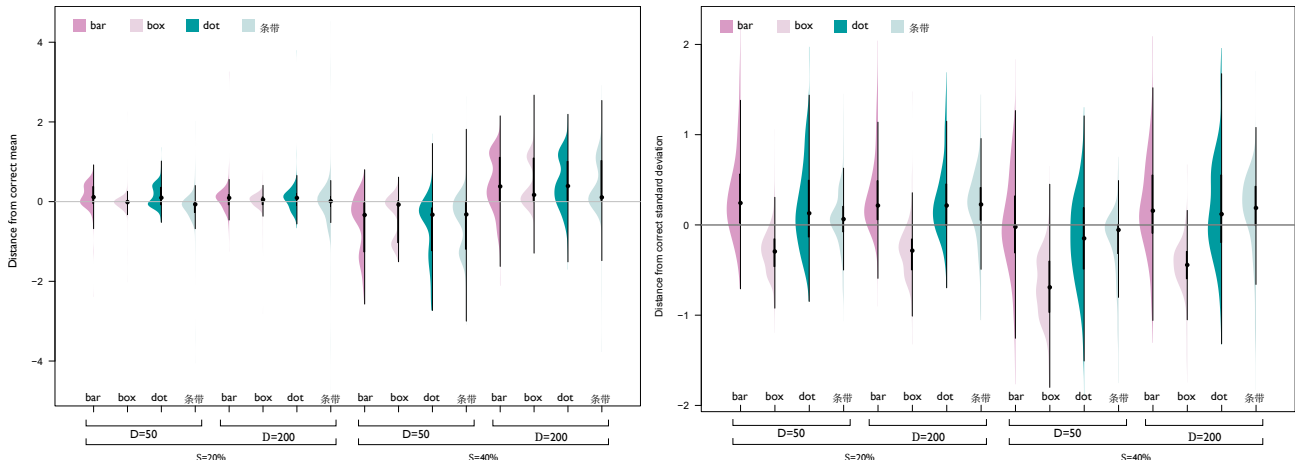
对于绝对标准差误差，绝对相关性要高得多——例如，条形直方图和点图高度相关（$|r| = 0.82$）；点图与条带图、条形直方图与条带图也呈现相关性（二者均为 $|r| = 0.53$）。

表1：**效应量**。可视化类型 V 的绝对均值误差和绝对标准差误差（%）。95% 置信区间通过自助法计算；参见第 4 节。

可视化 V	mean	95% 置信区间	s.d.	Cohen 的 d
绝对均值误差：				
– 条形直方图	0.312	[0.29, 0.34]	0.480	−0.001
– 箱线图	0.242	[0.20, 0.28]	0.713	−0.140
– 点图	0.328	[0.30, 0.36]	0.581	0.031
– 条带图	0.367	[0.33, 0.41]	0.833	0.110
绝对标准差误差（%）：				
– 条形直方图	32.2	[30.2, 34.4]	39.3	0.148
– 箱线图	30.7	[29.9, 31.7]	17.1	0.082
– 点图	30.6	[28.9, 32.5]	34.6	0.078
– 条带图	21.9	[20.6, 23.1]	24.6	−0.308

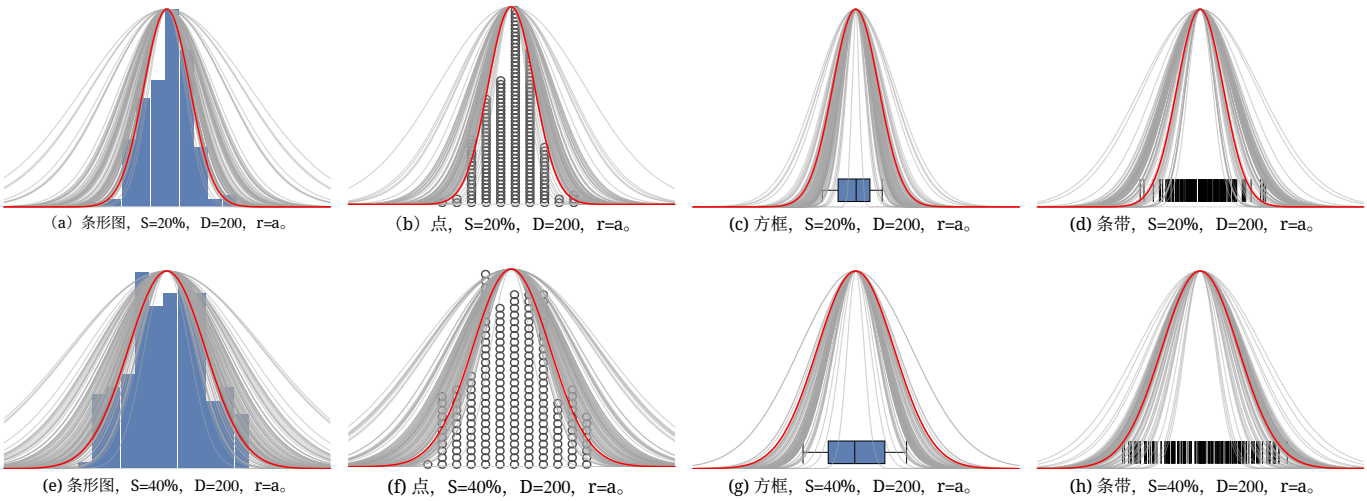
4.2 按因素特征分析

图 8 的前三行总结了基于可视化V、数据量D和离散程度S 的三项测量指标的表现，使用的是 95% 置信区间（按照上文所述通过自助法计算）。此外，表1总结了绝对均值误差和标准差误差（%）的效应量。我们采用了经典的 Cohen$d$ 公式，即不依赖于实验设计，因此便于与其他实验进行比较。对于可视化（第一行），与其他图表相比，使用箱线图估计均值时的误差更低（绝对均值误差=0.24，Cohen’s $d = -0.14$）



（a）均值估计。 （b）标准差估计。

图 5：**各可视化方式的详细性能**。按离散程度 S、数据量 D 和可视化方式 V 分组的所有试次中，均值（左）和标准差（右）误差的分布。



（a）条形图，S=20%，D=200，r=a。 （b）点，S=20%，D=200，r=a。 (c) 方框，S=20%，D=200，r=a。 (d) 条带，S=20%，D=200，r=a。

(e) 条形图，S=40%，D=200，r=a。 (f) 点，S=40%，D=200，r=a。 (g) 方框，S=40%，D=200，r=a。 (h) 条带，S=40%，D=200，r=a。

图 6：**曲线拟合示例**。两个特定试次的所有拟合曲线：S=20%，D=200，rep=a（上排）和S=40%，D=200，rep=a（下排）。曲线已围绕均值居中。红色曲线表示给定数据样本的实际高斯分布。为提供背景信息，相应的可视化已添加到背景中。

。条带图的表现稍差（绝对均值误差=0.37，$d = -0.11$），但与条形图和点图的表现相当。与其他图表相比，参与者使用条带图估计标准差时准确得多（标准差绝对百分比误差=0.21，Cohen's $d = -0.31$）。条形图在估计标准差方面的准确性最低（标准差绝对百分比误差=0.32，Cohen's $d = -0.15$）。

图 8 第二行汇总了数据量的测量结果。尽管较大的数据量与更好的表现相关，但这种影响对于均值误差（$D = 50$为 0.32，$D = 200$为 0.30，Cohen's $d = 0.02$）和标准差误差（$D = 50$为 29.4%，$D = 200$为 28.4%，Cohen's $d = 0.03$）而言都很小。

在图 8 的最后一行中，我们看到相同的数据在离散程度上的结果。这里，虽然离散程度较大时均值误差略高（20% 时为 0.307，40% 时为 0.317，Cohen's $d = 0.0142$），但标准差的相对性能提升幅度更大（20% 时为 34.1%，40% 时为 23.7%，Cohen's $d = 0.341$）。

最后，我们在图 9 中展示了离散程度与数据量和可视化之间的交互作用。以下是一些值得注意的观察：

- 对于数据量与可视化之间的交互作用，数据量越大，绝对均值误差似乎越小。箱线图是个例外，在 50 个和 200 个项目时，其绝对均值误差始终较低。至于绝对标准差误差，也出现了相同趋势——数据量越大，误差似乎越小。与其他技术相比，条带图在两种数据量下的误差始终较低。

- 对于离散程度与可视化之间的交互作用，绝对均值误差似乎会随着离散程度的增加而增大；然而，这一效应并不适用于箱线图，因为在两个离散程度水平下，箱线图显示出相近的绝对均值误差。此外，对于这一相同的交互作用，与较低离散程度（20%）相比，所有可视化在较高离散程度（40%）下的标准差误差都较低，但箱线图再次例外；在这两种条件下，箱线图的标准差误差基本保持不变。

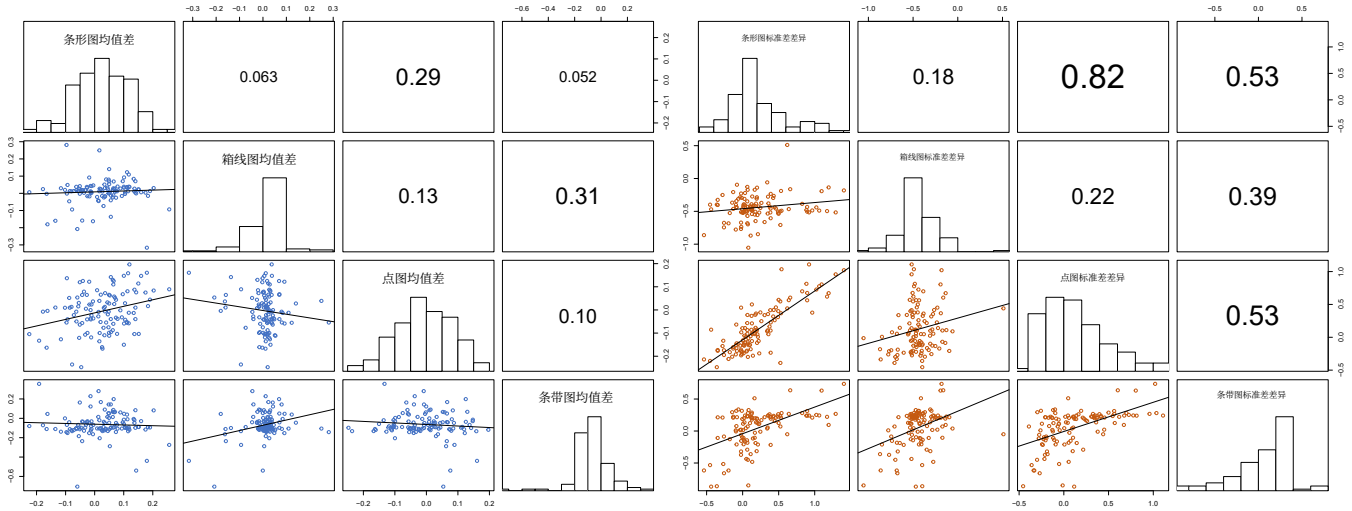


图 7：**性能比较**。比较各图表类型下每位受访者的平均表现。每个点代表一位受访者在所有试次中使用给定图表类型查找数据集均值或标准差时的平均绝对误差。矩阵中的单元格便于比较受访者在每一对可能的图表类型上的表现。查找均值的表现以蓝色显示（左侧），查找范围内标准差的表现则显示在右侧。每对可视化类型的皮尔逊相关系数r的绝对值标于对角线上方。

4.3 人口统计信息与参与者反馈

图 10 总结了人口统计调查的结果。这些结果与此前对亚马逊土耳其机器人用户群体的了解尝试大体一致 [49]。重要的是，我们的受访者中仅有极少数（8 人）具备专业统计学经验，而只有大约七分之一的受访者具备中等程度的统计学经验。

117 位参与者中有 37 位提供了自由形式的反馈。总体而言，反馈较为积极，几位参与者提到这项任务既有趣又引人入胜。参与者 P21812 写道：“这几乎像是游戏化的内容，就像你会在平板电脑或手机上看到的那种。”同样，“这很有趣！我一直想把所有东西都放进钟形曲线里，虽然说明里并没有这么说。”（P51429）许多其他参与者也提到拟合曲线颇具挑战性：“很多数据都不服从正态分布！”（P15279），“有些真的很难，只能试着凭目测判断。”（P28558），以及“不太确定这些究竟该怎么拟合，有些看起来太奇怪了。”（P49775）

最后，参与者 P21883 提出了一个更笼统的评论：“我上高中时学过统计学，但不记得当时是这样估算曲线的。我只记得做了大量计算。”P43443 说得更深入：“我觉得，要是去证明弦理论，我可能都比做这个练习更有把握，哪怕准确度再低也好。”幸运的是，我们的结果与此相悖。

4.4 对预注册的偏离

在预注册中，我们表示将收集 100 名参与者的数据。在实际研究中，我们总共招募了 150 名参与者（最终纳入 146 名，筛选后则剩下 117 名），这无意中偏离了预注册。造成这一偏离的原因是研究团队内部沟通不畅。虽然此处未展示，但我们对最初的 100 名受访者进行了单独分析，结果显示的效应与我们报告的完整结果相同。

由于训练具有误导性，参与者无需调整离散程度来拟合蓝色曲线，因此我们还移除了

我们认为有 10 名参与者误解了曲线拟合任务，因此移除了他们的数据。如果某位参与者至少有一个完整的 48 次试验区块，其离散程度为 1.0（初始离散程度），我们便将其归类为存在此类误解。我们在 OSF 上提供了未经筛选的数据。

此外 onal deviations include the following:

- 我们曾说，离散程度数值是“通过试测确定的”。这是不正确的——这些数值是根据我们对该实验先前版本的经验确定的。
- 我们的预注册包含第五个假设：“不同参与者的表现会存在不均匀性。”这一假设表述不够具体且含义模糊，因此我们决定在分析中将其舍弃。
- 我们的一些图表和分析——包括对皮尔逊相关和科恩效应量的分析，以及图7中的散点图矩阵——并未纳入预注册。为提供更丰富的分析和结果报告，我们在此将其列出。

5 讨论

我们的研究产生了若干有趣的发现，同时也证实了我们对这项任务的基本直觉。首先，我们结合这些结果回顾各项假设。接着，我们详细解释研究结果，并讨论其普遍适用性。最后，我们将讨论推动这项研究的总体研究愿景。

5.1 回顾假设

我们的结果有力地表明，**均值估计比离散程度估计更准确**。总体而言，图 8 中左侧和中间两列的图表显示，无论可视化技术、数据量或离散程度如何，人们估计数据集中的平均值都比估计其离散程度容易得多。这支持了我们的第一个假设。

我们观察到，对于条形图直方图和点图直方图，随着数据量增加，均值和标准差估计的误差都会降低。然而，对于其他可视化，似乎

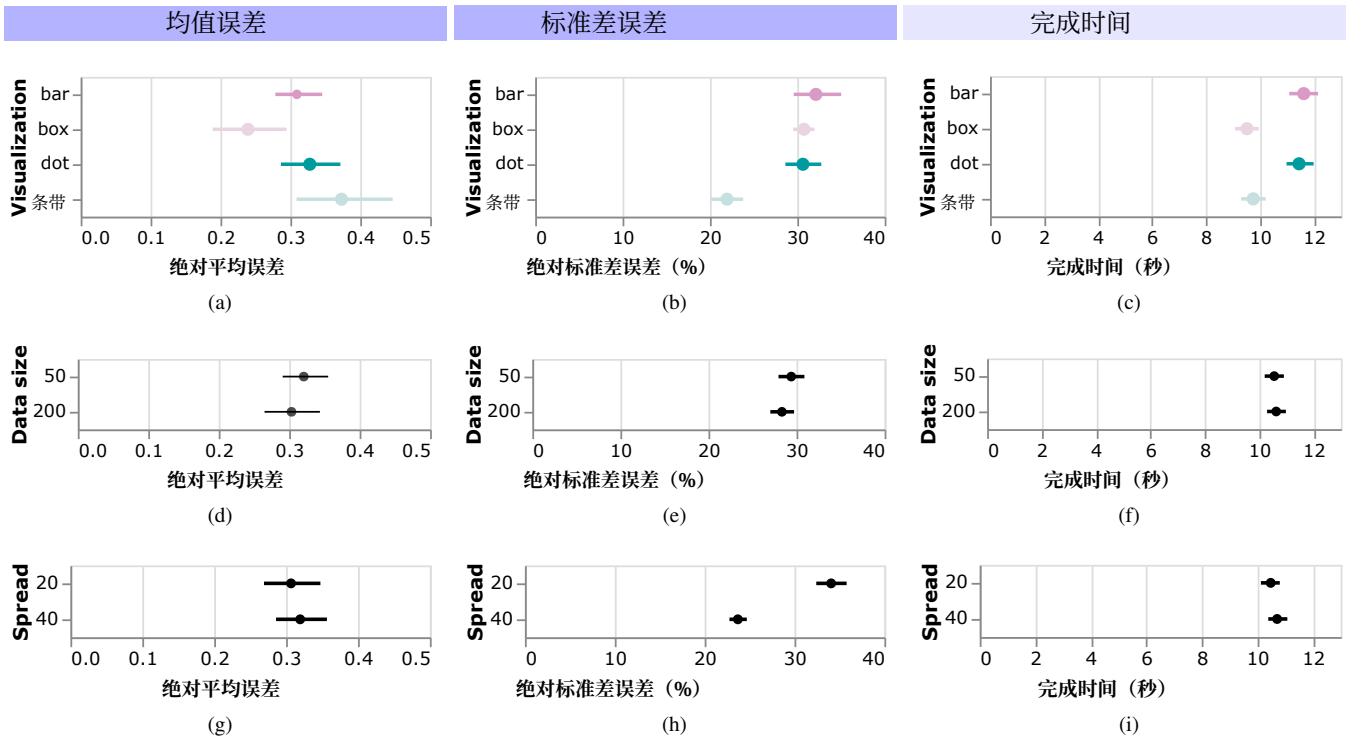


图 8：**整体表现**。按参与者进行自助抽样得到的 95% 置信区间，显示可视化 *V*、数据量 *D*和离散程度 *S*对均值误差、标准差误差和完成时间的影响。（请注意，此处列出完成时间仅为完整起见。）

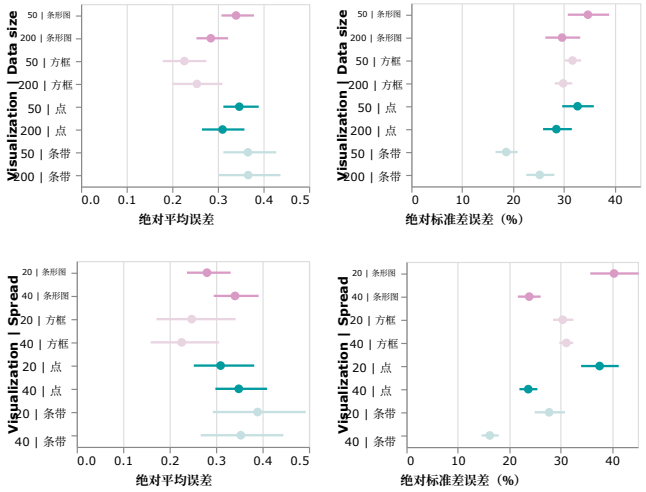


图9：**因素间交互作用**。按参与者进行自助抽样得到的 95% 置信区间，显示可视化*V* 与数据量 *D*和离散程度*S* 对绝对均值误差和标准差误差的影响。

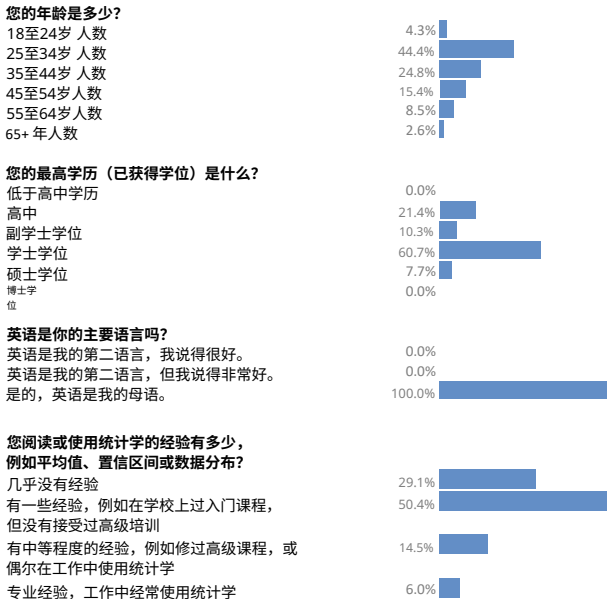


图10：**人口统计学特征**。 参与者人口统计学特征摘要。

并不存在这种效果；对于条带图，误差实际上会随着数据量增加而增大。虽然这可以说在一定程度上支持了我们的第二个假设，但我们认为，这反而说明“数据越多，估计越准确”这一假设过于简单化。例如，箱线图甚至不编码数据量，而随着条带数量增加，条带图似乎也难以良好扩展。

与我们的第三个假设一致，图 8 的顶行显示，不同可视化类型的表现存在一定的不均匀性。正如我们所预期的，**估算均值时，箱线图的误差最低，条带图的误差最高**。

我们的第四个假设涉及这样一种假定：参与者会试图让曲线下方尽可能多地包含样本中可见的数据点。我们称之为“雨伞效应”，仿佛人们在为数据遮雨。如果确实如此，那么对于箱线图以外的可视化，人们会系统性地高估标准差。图5、图6和图7显示，人们经常高估条形图和点直方图的离散程度，但这一现象似乎并未延伸到箱线图和条带图。这表明这些图中存在雨伞效应

，支持了第四个假设。这种效应也可能适用于箱线图，因为低估可能表明参与者是在拟合须线，而须线会排除部分数据；也可能是在拟合箱线图的方框，而方框仅包含50%的数据。

5.2 解释结果

我们注意到，估计均值时的误差不会随着离散程度S的增大而增加；基本上，样本中的噪声越多，离散程度越大，就越难为直方图确定准确的均值。同样，我们的结果表明，随着数据量D 增加，误差并未发生变化。我们原本以为，数据点数量增加会使分布图景更加完整，从而降低寻找中点的感知难度。正如第 4.2 节末尾所述，我们确实观察到一些这方面的证据，但其程度远不如预期。

我们关于标准差误差的结果也值得关注。首先，数据量D 对该误差的影响很小。通常这会令人感兴趣，因为人们可能会认为，增加项目数量会让样本变得更加规整，从而提高准确度。然而，在选择变异系数水平时，我们控制了样本的相对噪声。另一方面，我们确实发现，随着离散程度S 增加，标准差误差会减小。这有悖直觉，但或许可以用前述的“雨伞效应”来解释。图 6 展示了这一现象。我们的结果还表明，大多数试次都高估了标准差，也就是说，估计值大于实际所需值，这进一步支持了这一解释。箱线图的结果则相反：标准差估计误差呈现出明显的负偏差，即反向雨伞效应。我们推测，这可能反映了参与者试图直接将曲线拟合到图形上。使用箱线图估计均值的结果非常出色，这或许只是因为，尽管我们天生具有识别分布均值的能力，但在个体层面上，我们无法与自动计算所提供的精度相匹敌。在来自正态分布的样本中，箱线图中线所表示的预先计算出的中位数与分布均值非常接近。构成箱线图的其他点也表示类似的预先计算值。因此，箱线图呈现出的分布看起来几乎没有随机噪声。

正态曲线更进一步，是对数据样本完全无噪声、理想化的呈现。仔细想来，参与者更容易将正态曲线应用于我们测试的四种可视化中噪声最少的一种，这似乎合情合理。作为数据可视化研究者，我们往往认为，数据偏离某种理想化模型的部分包含有用信息；我们会从细节中寻找意义。然而，同样的信息——离群值、意料之外的相关性以及数据中的间隙——显然也可能分散人们凭直觉「估读」传统汇总指标的注意力。传统统计学所关注的理想化形式，有时可能与我们对更杂乱的现实世界数据的直觉不太相符。每当我们计算均值和标准差时，都会将正态曲线这一结构套用到较小的样本上，而它并非天然就是合适的拟合。这或许凸显了在抽象或汇总数据之前，先通过观察数据来了解它的重要性。但这也说明，能够精确识别（往往至关重要的）集中趋势指标的汇总计算同样重要，因为数据中的噪声可能会干扰我们的视觉判断能力。

5.3 结果的推广

我们的参与者中年轻成年人占比较高，拥有大学学历的比例也高于美国普通人群。然而，参与者自述的统计学知识水平相对基础。统计学训练水平较低，使我们有信心认为这些发现可能具有更广泛的普遍性。

我们认为，我们试验数据集的样本量（50 个和 200 个数据项）代表了许多研究领域中人们通常采用的样本量。较小的样本量确实引发了一个问题：这些结果是否适用于未来关于视觉推断的研究，因为当案例数少于 100 时，学生 t 分布并不完全等同于正态分布 [43]。不过，在 $n = 50$ 时，这两条曲线相似，因此我们预计，小样本的噪声会使其更难拟合，这一挑战将是更为重要的影响因素。

我们测试的可视化在聚合程度上各不相同，涵盖了从零聚合（条带图）、部分空间聚合（点状直方图）、中等聚合（条形直方图），到高程度聚合（箱线图）的完整范围。我们认为，用于可视化分布的其他技术，例如小提琴图或密度图，也可能会在这一维度上获得类似的评价。这样一来，未来的研究人员或许可以利用本研究的结果，为这些图形或其他图形的性能预测提供参考。不过，其他技术可能表现出完全不同的行为。例如，假设结果图（HOPs） [35], 使用动画来传达不确定性，因此，尽管它们没有经过聚合，却依赖于空间记忆和其他我们在本研究中未调查的视觉机制。

正如我们在交互研究（第 4.2 节）中所示，聚合是一把双刃剑。其他可视化类型在数据量和离散程度变化时，估计标准差的表现也会随之变化；而箱线图在这些属性上的表现始终如一。然而，这种始终如一的 误差相对较高：无论分布如何，箱线图的外观都相对稳定，但这种简洁性可能会掩盖分布更细致的属性，或使观众产生误解。

5.4 局限性

尽管此处呈现的结果是我们这项研究的第二个版本，但我们的实验仍有一些不足。首先，我们对预注册（第 4.4 节）做了几处偏离，而且有些分析也未在预注册中明确说明。在未来的工作中，我们将努力在规划阶段就做到更加精确和规范。

我们的训练试次通过在每种可视化类型的数据分布上方显示一条预先拟合的理想化曲线，让参与者为回答问题做好准备。然而，可以想见，这种训练教给人们的可能更多是如何针对特定可视化拟合曲线，而非如何根据数据拟合曲线。这一观点完全合理，本质上也正是我们实验的目的：了解不同可视化如何传达正态数据样本的均值和标准差。尽管如此，我们仍试图避免训练参与者按照预设模式机械地拟合曲线，因此每种可视化类型只为参与者提供一次训练试次。

此外，我们的研究仅涉及正态分布数据。这是本次实验的明显局限，未来还需要开展更多工作来探索这一领域。此外，为了使实验设计的规模保持可控，我们决定不将分箱大小的数量和配置纳入本次实验；已有研究表明，它们是直方图设计中的重要因素 [16]。我们改为固定分箱数量（50 个）和每个分箱的大小（$20.0/50 = 0.40$）。这是我们工作的另一项局限，因为我们的两种技术——条形图和点状直方图——显然会受到这一选择的影响，而另外两种——条带图和

箱线图——则不会。虽然我们认为分箱大小总体上适合该数据集的属性，但这仍是一个重要因素，我们希望未来能对此展开研究。

最后，我们的研究聚焦于拟合曲线，以便将其用于经典参数统计检验。有人可能会认为，鉴于更现代的检验可能更适用于图形推断，以这类经典检验为基础开展这项研究适得其反。虽然我们不一定不同意这种看法，但我们指出：(1) 对离散样本拟合正态分布是许多基本统计运算中的常见任务——其中包括对分布进行简单描述的能力；(2) t 检验及其他参数统计检验仍是基本的统计工具。

5.5 未来工作

本文研究的总体愿景是降低使用真正统计方法的门槛，使数学或统计学背景有限的人也能使用这些方法。这一愿景受到感知与可视化研究发现的启发（例如， [5], [18], [23]），这些发现表明，恰当的视觉表征能够让人们掌握复杂的统计学知识。

本文的研究是朝着这一愿景迈出的一步。我们的发现表明，受过极少训练的人也能从视觉呈现中获取有用的统计信息。更重要的是，我们的发现显示，人们的表现一致；根据所使用的可视化方式，他们往往会高估或低估标准差。这使我们有理由相信，可以在下一步设计视觉干预来克服这些偏差；同时也为未来的实证研究指明了有益方向，例如探索更复杂的情形，如比较多个分布或构建区间。

展望未来，我们认为，对单个分布中的统计矩进行简单估计，是推进“视觉统计学” [3]并厘清“图形推断” [4]的局限与潜在陷阱这一工作的一个重要且必要的起点。例如，我们的实验中不同可视化类型的表现差异，表明存在改进受众对分布特性理解的潜在干预空间，例如结合多种单变量可视化的“混合型”可视化 [16]。

6 结论

我们介绍了一项预注册众包用户研究的结果，该研究考察了参与者如何为通过不同形式呈现的不同含噪程度的数据样本拟合正态曲线。我们发现，参与者能够较准确地确定数据样本的均值（误差约为 12%–13%），并且能够确定样本的标准差（误差为 28%–29%）。我们认为，这些结果对依赖受众充分理解均值和离散程度的可视化设计师而言令人鼓舞，尤其是在设计师依赖隐式估计或满意化策略 [6]的情况下。

参考文献

[1] D. A. Szafir、S. Haroz、M. Gleicher 和 S. Franconeri，《数据可视化中的四种集合编码类型》，《视觉杂志》，第 16 卷，第 5 期，第 11 11, 03 2016 [ ] 10 1167 16 5 11 – 页。在线。获取地址：https://doi.org/ 。/ 。. R. A. Rensink，《可视化科学的前景》。纽约，纽约州：Springer New York，2014年，第 147–175 页。 [在线]。获取地址：https://doi.org/ 10 1007 978 1 4614 7485 2 6 。/ - - - -M. Correll，《改进视觉统计学》，博士学位论文，威斯康星大学麦迪逊分校，2015年。

[4] A. Buja、D. Cook、H. Hofmann、M. Lawrence、E.-K. Lee、D. F. Swayne 和 H. Wickham，《用于探索性数据分析和模型诊断的统计推断》，英国皇家学会哲学汇刊，第 367 卷，第 1906 期，第 4361–4383 页，2009年。 [在线]。可获取：https://doi.org/ 10 1098 2009 0120 . /rsta. . M. Correll 和 M. Gleicher，《误差条有害：探索均值与误差的替代编码》，IEEE《可视化与计算机图形学汇刊》，第 20, 12, 2142 2151, 卷，第 期，第 – 页。在线。可获取：https://doi.org/ . /TVCG. . 10 1109 2014 2346298

[6] A. Kale、M. Kay 和 J. Hullman，《效应量判断与决策的视觉推理策略》，IEEE《可视化与计算机图形学汇刊》，第 27 卷，第 2 期，第 272–282 页，2021年。 [在线]。可获取：https://doi.org/ 10 1109 2020 3030335 . /TVCG. . B. D. Ondov、F. Yang、M. Kay、N. Elmqvist 和 S. Franconeri，《利用对抗样本揭示感知代理线索》，IEEE《可视化与计算机图形学汇刊》，第 27, 2, 1073 1083, 卷，第 期，第 – 页。在线。可获取：https://doi.org/ . /TVCG. . 10 1109 2020 3030429

[8] L. Yuan、S. Haroz 和 S. Franconeri，《从数据可视化中提取均值的感知代理线索》，《心理学通报与评论》，第 26 卷，第 2 期，第 669–676 页，2019年。 [在线]。可获取：https://doi.org/ 10 3758 13423 018 1525 7 . /s - - - J. W. Tukey，《探索性数据分析》。美国马萨诸塞州雷丁：Addison-Wesley，1977年。[10] E. L. Lehmann 和 J. P. Romano，《统计假设检验》，第 3 版。美国纽约州纽约市：Springer，2005年。[11] M. J. Schervish，《统计学理论》。美国纽约州纽约市：施普林格出版社，1995年。[12] W. S. Cleveland，《数据可视化》。美国新泽西州萨米特：Hobart出版社，1993年。[13] E. L. Scott、C. P. D. Shane 和 M. D. Swanson，《照相底片上星系合成分布与实际分布的比较》，《天体物理学》，第 119 卷，第 91–112 页，1954年。 [在线]。可获取：https://doi.org/ 10 1086 145799 . / H. Wickham、D. Cook、H. Hofmann 和 A. Buja，《面向信息可视化的图形推断》，IEEE《可视化与计算机图形学汇刊》，第 16 卷，第 6 期，第 973–979 页，2010年。 [在线]。可获取：https://doi.org/ 10 1109 2010 161 . /TVCG. . R. Beecham、J. Dykes、W. Meulemans、A. Slingsby、C. Turkay 和 J. Wood，《地图阵列：空间结构对图形推断的影响》，IEEE《可视化与计算机图形学汇刊》，第 23 卷，第 1 期，第 391–400 页，2017年。 [在线]。可获取：https://doi.org/ 10 1109 2016 2598862 . /TVCG. . M. Correll、M. Li、G. L. Kindlmann 和 C. Scheidegger，《我觉得看起来不错：将可视化用作合理性检查》，IEEE《可视化与计算机图形学汇刊》，第 25 卷，第 1 期，第 830–839 页，2019 年。 [在线]。可获取：https://doi.org/ 10 1109 2018 2864907 . /TVCG. . M. Correll、D. Albers、S. Franconeri 和 M. Gleicher，《时间序列数据中的均值比较》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2012年，第 1095–1104 页。 [在线]。可获取：https://doi.org/ 10 1145 2207676 2208556 . / . D. Albers、M. Correll 和 M. Gleicher，《时间序列可视化中聚合的任务驱动评估》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2014年，第 551–560 页。 [在线]。可获取：https://doi.org/ 10 1145 2556288 2557200 . / . W. Aigner、A. Rind 和 S. Hoffmann，《结合定量数据与定性抽象的交互式时间序列可视化的比较评估》，《计算机图形论坛》，第 31 卷，第 3 期，第 995–1004 页，2012年。 [在线]。可获取：https: //doi.org/10.1111/j.1467-8659.2012.03092.x [20] J. Fuchs、F. Fischer、F. Mansmann、E. Bertini 和 P. Isenberg，《小多图情境下时间序列数据替代字形设计的评估》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2013年，第 3237–3246 页。 [ ] 10 1145 2470654 2466443 在线。可获取：https://doi.org/ . / . M. Correll 和 J. Heer，《目测回归：估计双变量可视化中的趋势》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2017年，第 1387 1396 [ ] 10 1145 3025453 3025922 – 页。在线。可获取：https://doi.org/ . / . G. Fouriezos、S. Rubenfeld 和 G. Capstick，《视觉统计决策》，《知觉与心理物理学》，第 70 卷，第 3 期，第 456–464 页，2008年。 [在线]。可获取：https://doi.org/ 10 3758 70 3 456 . /PP. . . M. Gleicher、M. Correll、C. Nothelfer 和 S. Franconeri，《多类别散点图中平均值的感知》，IEEE《可视化与计算机图形学汇刊》，第 19, 12, 2316 2325, 卷，第 期，第 – 页。在线。可获取：https://doi.org/ . /TVCG. . 10 1109 2013 183

[24] F. Nguyen、X. Qiao、J. Heer 和 J. Hullman，《聚合选择对未受训练的可视化用户从数据中进行概括的影响探索》，《计算机图形论坛》，第 39 卷，第 6 期，第 33–48 页，2020年。 [ ] 10 1111 13902 在线。可获取：https://doi.org/ . /cgf. [25] D. Whitaker 和 H. Walker，《随机点簇游标对准中的质心评估》，《视觉研究》，第 28， 7， 777 784, 卷，1988 [ ] 10 1016 0042 6989 88 90024 第 期，第 – 页。在线。可获取：https://doi.org/ . / - ( ) - 7 [26] M. J. Morgan 和 A. Glennerster，《人类观察者定位点簇中心的效率》，《视觉研究》，第 31 卷，第 12 期，第 2075–2083 页，1991年。 [在线]。可获取：https://doi.org/ 10 1016 0042 6989 91 90165 2 . / - ( ) - D. Ariely，《看见集合：以统计属 [27] 性表征》，《心理科学》，第 12 卷，第 2 期，第 157–162 页，2001年。 [在线]。可获取：https://doi.org/ 10 1111 1467 9280 00327 . / - . D. Melcher 和 E. Kowler，《形状、表面与扫视》，《视觉研究》，第 39 卷，第 17 期，第 2929–2946 [28] 页，1999年。 [在线]。可获取：https://doi.org/ 10 1016 0042 6989 99 00029 2 . /S - ( ) - G. A. Alvarez，《将多个对象表 [29] 征为一个整体可增强视觉认知》，认知科学趋势，第 15 卷，第 3 期，第 122–131, 2011 [ ] 10 1016 2011 01 003 页。在线。可获取：https://doi.org/ . / j.tics. . . S. C. Chong 和 A. Treisman，《统计属性的表征》，《视觉研 [30] 究》，第 43 卷，第 4 期，第 393–404 页，2003年。 [在线]。可获取：https://doi.org/ 10 1016 0042 6989 02 00596 5 . /s - ( ) - A. R. Albrecht 和 B. J. Scholl，《连续视觉世界中的知觉平均：随时间提取统计摘要表征》，《心理科学》，第 21 卷，第 [31] 4 期，第 560–567 页，201年。 [在线]。可获取：https://doi.org/ 10 1177 0956797610363543 . / M. Kay、T. Kola、J. R. Hullman 和 S. A. Munson，《我的公交车什么时候（左右）到？：日常移动预测系统中以 [32] 用户为中心的不确定性可视化》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2016年，第 5092 5103 [ ] 10 1145 2858036 2858558 – 页。在线。可获 [33] 取：https://doi.org/ . / . K.-W. Moon，《Wilkinson 点图》，载于《使用 Shiny 应用学习 ggplot》。美国纽约州纽约市：Springer，Cham，2016年，第 103–109 页。 [在线]。可获取：https://doi.org/ 10 1007 978 3 319 53019 2 12 . / - - - - J. Hullman、M. Kay、Y. Kim 和 S. Shrestha，《想象重复实验：图形预测与离散可视化可 [34] 改善对效应不确定性的回忆与估计》，IEEE《可视化与计算机图形学汇刊》，第 24 卷，第 1 期，第 446–456 页，2018 年。 [在线]。可获取：https://doi.org/ 10 1109 2017 2743898 . /TVCG. . J. Hullman、P. Resnick 和 E. Adar，《假设结果图在推断变量排序可靠性方面优于误差条和小提琴图》，PLOS ONE，第 10 卷，第 11 期，第 1–25 页，2015年11月。 [35] [ ] 10 1371 0142444 在线。可获取：https://doi.org/ . /journal.pone. N. Crilly、A. F. Blackwell 和 P. J. Clarkson，《图形引 [36] 导：将研究图示用作访谈刺激材料》，《定性研究》，第 6 卷，第 3 期，第 341–366 页，2006年。 [在线]。可获取：https://doi.org/ 10 1177 1468794106065007 . / Y. Kim 、 K. Reinecke 和 J. Hullman，《解释差距：可视化个人预测可提高对数据的回忆 [37] 与理解》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2017年，第 1375–1386 页。 [在线]。可获取：https://doi.org/ 10 1145 3025453 3025592 . / . Y. Kim、L. A. Walls、P. M. Krafft 和 J. Hullman，《改进数据可视化的贝叶斯认知方法》，载于《ACM计算机系统中 [38] 的人为因素会议论文集》。美国纽约州纽约市：ACM，2019年，第 682:1–682:14 页。 [在线]。可获取：https://doi.org/ 10 1145 3290605 3300912 . / . T. M. Spalek 和 S. Hammad，《返回抑制中的从左到右偏差源于阅读方向》，《心理科学》，第 16 卷，第 1 期，第 15–18 页，2005年。 [在线]。可获取：https://doi.org/10.1111/j.0956-7976.2005.00774.x [39] [40] J. Heer 和 M. Bostock，《众包图形感知：使用 Mechanical Turk 评估可视化设计》，载于《ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，2010年，第 203–212 页。 [在线]。可获取：https://doi.org/ 10 1145 1753326 1753357 . / . C. Ahlberg、C. Williamson 和 B. Shneiderman，《用于信息探索的动态查询：实现与评估》，载于《 [41] ACM计算机系统中的人为因素会议论文集》。美国纽约州纽约市：ACM，1992年，第 619–626 页。 [在线]。可获取：https://doi.org/ 10 1145 142750 143054 . / . M. Bostock、V. Ogievetsky 和 J. Heer，《D³：数据驱动文档》，IEEE《可视化与计算机图形学汇刊》，第 17 卷，第 12 期，第 2301–2309 页，2011年12月。 [在线]。可获取：https://doi.org/10.1109/TVCG.2011.185 [42] [43] Student，《均值的可能误差》，*Biometrika*，第 6 卷，第 1 期，页码 1 25, 1908 [ ] 10 1093 6 1 1 —。在线阅读。链接：https://doi.org/。/biomet/。。 [44] 53, 3, L. Wilkinson，《点图》，*TheAmerican Statistician*，第 卷，第 页，276–281，1999年。 [在线]。链接：https://doi.org/10.1080/00031305. 1999.10474474 [45] D. Park、S. M. Drucker、R. Fernandez 和 N. Elmqvist，《Atom：单位可视化语法》，*IEEE*《可视化与计算机图形学汇刊》，第 24 卷，第 12 期，3032–3043 页，2018年。 [在线]。链接：https://doi.org/ 10 1109 2017 2785807 。/TVCG.。 H. Wickham 和 L. Stryjewski，《箱线图的 年》，2011年，未发表。[在线]。链接：[46] 40 http://vita.had.co.nz/papers/boxplots.pdf [47] B. Efron，《自助法：再看删刀法》，载于统计学突破。美国纽约州纽约市：施普林格出版社，1992年，第 569–593 页。 [在线]。链接：https://doi.org/10.1007/ 978-1-4612-4380-9 41 [48] P. Dragicevic，《人机交互中的公平统计沟通》，载于人机交互现代统计方法，J. Robertson 和 M. Kaptein 编。美国纽约州纽约市：施普林格出版社，2016年，第 291–330 页。 [在线]。链接：https://doi.org/ 10 1007 978 3 319 26633 6 13 。/ - - - - J. Ross、L. Irani、M. S. Silberman、A. Zaldivar 和 B. Tomlinson，《众包工作者是 [49] 谁？：Mechanical Turk 中人口统计学特征的变化》，载于*ACM*计算机系统中的人为因素会议扩展摘要。美国纽约州纽约市：ACM，2010年，第 2863–2872 页。[在线]。链接：https://doi.org/10.1145/1753846.1753873



EricNewburger 于1995年获得美国威斯康星州麦迪逊市威斯康星大学麦迪逊分校应用经济学硕士学位。他是美国马里兰州大学公园市马里兰大学帕克分校信息研究学院的博士候选人和兼职讲师。他也是 UMD 人机交互实验室（HCIL）的学生成员。



Michael Correll 于2015年获得美国威斯康星州麦迪逊市威斯康星大学麦迪逊分校博士学位。他是美国华盛顿州西雅图市 Tableau Software 旗下 Tableau Research 的高级研究科学家。他的研究兴趣包括数据伦理、向大众传播统计学，以及研究有偏差或具有误导性的数据可视化。



Niklas Elmqvist 于2006年在瑞典哥德堡的查尔姆斯理工大学获得博士学位。他现任美国马里兰州大学公园市马里兰大学信息研究学院教授，同时是高级计算机研究所（UMIACS）的成员，曾任 UMD 人机交互实验室（HCIL）主任。他还是 IEEE 和 IEEE计算机学会的高级会员。

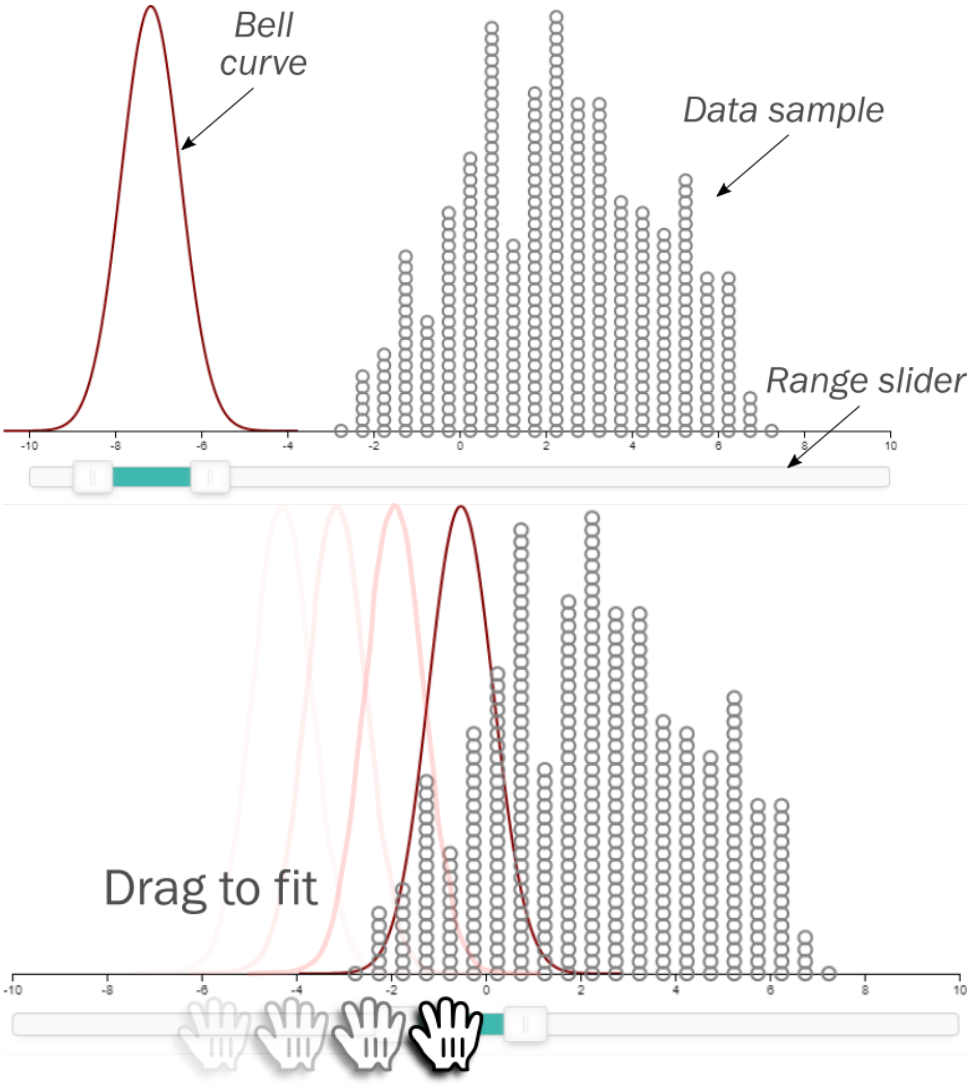
使用可视化将钟形曲线拟合到数据分布补充材料

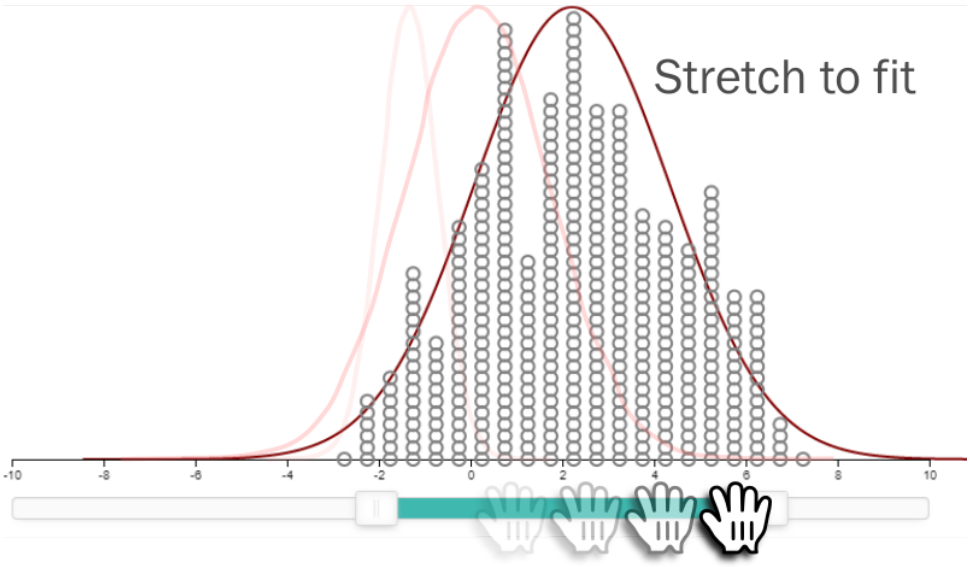
附录 A 补充材料

所有用户研究材料均已发布在我们的 OSF 网站上：https://osf.io/r3jpn/

附录 B 研究说明

我们请你为一组带有噪声的数据样本拟合一条平滑的钟形曲线。该样本是从一个本身呈钟形曲线分布的总体中随机抽取的；你可以想象，这些数据代表一大群人的身高，有些人高，有些人矮，但大多数人身高处于中间水平。然而，由于样本规模较小且是随机抽取的，其形状可能会有些不规则。你的任务是调整曲线的中心和离散程度，将钟形曲线拟合到数据样本上，以此尽可能估计源总体的大致形状。





这里的灰色点代表每个样本点。如图所示，目标是将红色曲线“拟合”到灰色点上。这意味着两点：(1) 曲线应**以灰色点为中心**。曲线应**拉伸**，使其近似呈现这些点的形状。

拟合时，目标**不是** 让所有点都落在曲线内，而是要找到点与曲线之间的最佳近似拟合。满意后，点击“**提交**”。

附录 C
补充可视化

图 1 和图 2　　分别展示个体参与者在均值误差和标准差误差方面的表现。

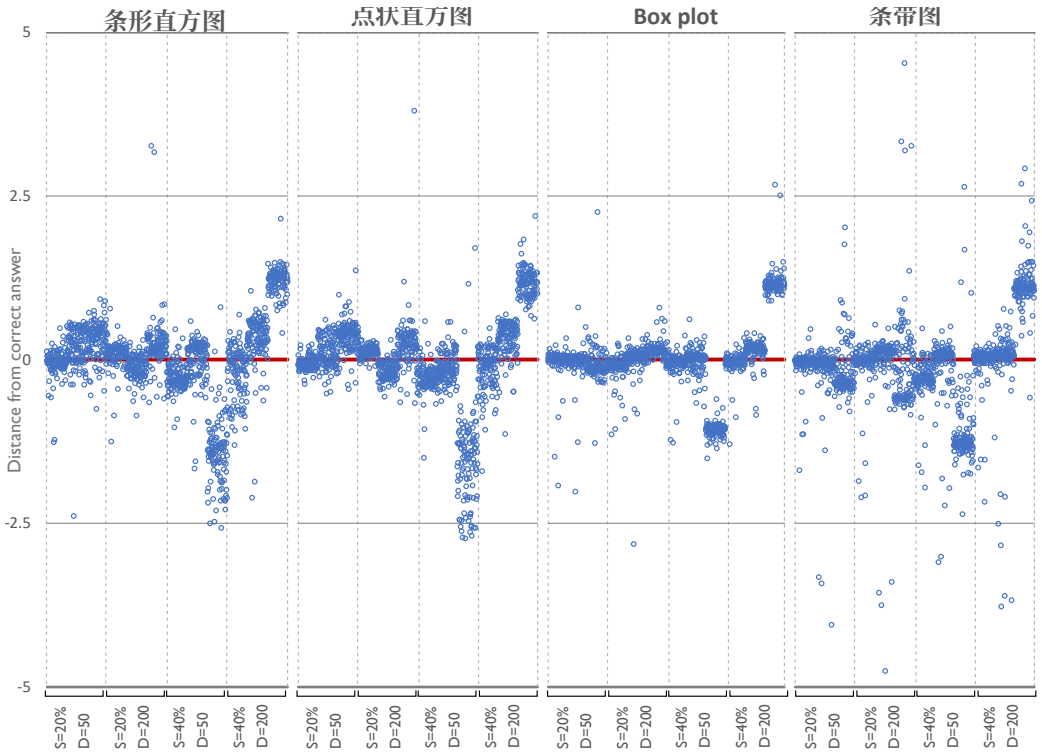


图 1：**均值误差表现**。真实值与个体参与者对数据集均值的估计值之间的距离。红线表示零距离（答案正确）。每个特定离散度*S* 和数据集大小*D* 对应的区域代表一组试次，每组使用相同的数据集）。

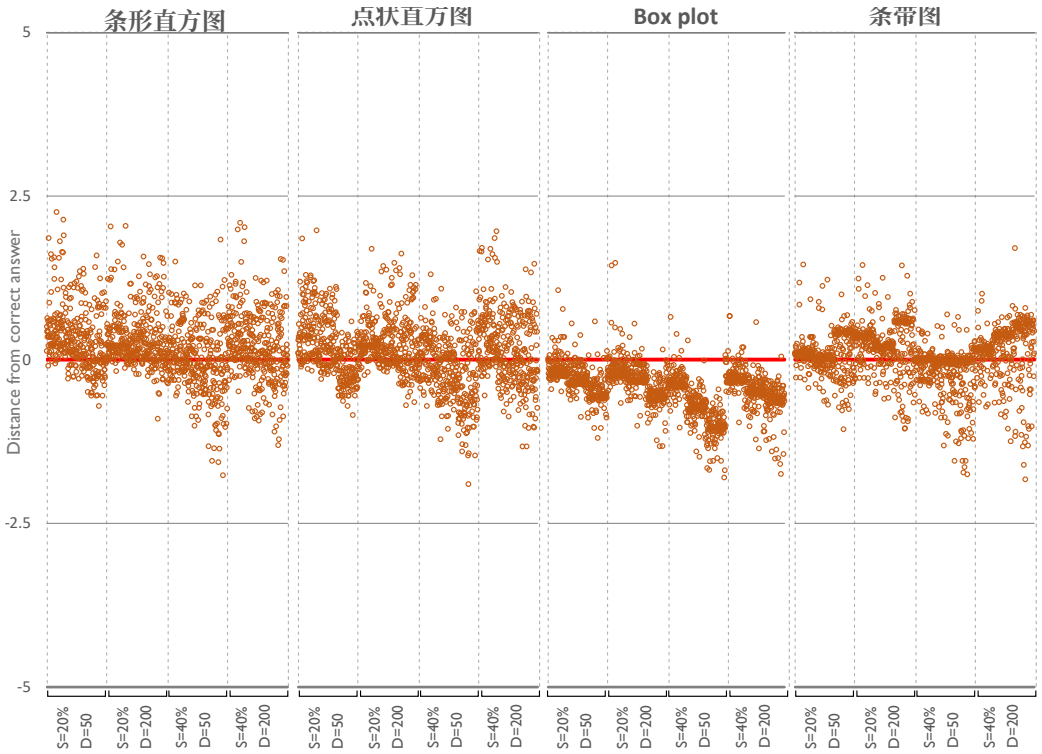


图 2：**标准差误差表现**。真实值与各参与者对数据集标准差的估计值之间的距离。红线表示零距离（正确答案）。特定离散程度S 和数据集大小D 对应的每个区域均代表一组试次，每组使用相同的数据集）。