arXiv:2603.13204v1 [eess.AS] 13 Mar 2026

主观与客观测量一致性的界限

Jaden Pieper 和 Stephen D. Voran，*IEEE* 高级会员

摘要—**多媒体质量的客观估计器通常通过将估计值与主观“真实数据”进行比较来评判，最常采用皮尔逊相关系数（PCC）或均方误差（MSE）。然而，主观测试结果包含噪声，因此追求 PCC 为 1.0 或 MSE 为 0.0 既不现实，也无法重复实现。人们已作出许多努力，试图在客观与主观比较中承认并妥善处理主观测试噪声，通常由此形成新的分析框架和评价指标。我们采用一种不同的方法。只作最基本的假设，我们便能够推导出主观测试中 PCC 和 MSE 可能达到的界限。**

与直觉一致，这些界限是主观投票方差的函数。当主观测试包含投票方差信息时，界限的计算十分直接；在这种情况下，我们称所得界限为“完全由数据驱动”。对于无法获得投票方差信息的情况，我们还提供了两种计算界限的方案。一种方案是使用其他包含此类信息的主观测试中的投票方差信息；另一种方案是使用主观投票模型。

因此，我们提出了一种用于主观投票的基于二项分布的模型（BinoVotes），它自然引出了一种平均意见分（MOS）模型，名为 BinoMOS，并具有多项独特的优良特性。BinoMOS 再现了 MOS 值的离散特性，以及这些值对每个文件投票数的依赖关系。该建模提供了推导皮尔逊相关系数（PCC）界和均方误差（MSE）界所需的投票方差信息，并且我们将这一建模与来自18项主观测试的数据进行了比较。该建模得出的 PCC 界和 MSE 界与直接从数据中得到的结果非常吻合。这些结果使人们能够预估任何主观测试可能达到的 PCC 和 MSE，即使该测试不提供投票方差信息。

索引词—**二项分布、相关性、均方误差、客观估计器、评分模型、主观测试、投票模型**

I. 引言

ATTRIBUTES 多媒体信号的质量、噪声程度等特性通常通过主观测试来测量。最常见的做法是，受试者观看一组视频文件或收听一组音频文件，并通过为每个文件评定 1（差）、2（较差）、3（一般）、4（良好）或 5（优秀）来评价其总体质量。对每个文件，将多位受试者的评分取平均，即得到平均意见分（MOS）。这些测试基于这样一个基本假设：每个文件都有真实质量，可通过开展设计良好的主观测试来估算。测试试验可以使用单个刺激，也可以进行比较；既可以在实验室进行，也可以由众包工作者完成。国际电信联盟为语音 [1]–[4], 音频[5], 视频 [6]–[8], 多媒体 [9], [10], 和游戏 [11]领域制定了具体的测试协议。

收稿日期：2025年11月5日；修订日期：2026年3月3日；目前正在审稿。作者供职于美国国家电信和信息管理局电信科学研究所，位于美国科罗拉多州博尔德。

无论主观测试设计得多么完善，其结果都像所有测量结果一样并不完美。尽管如此，这些测试仍是衡量音频、视频或多媒体信号质量及其他属性的黄金标准。它们成本高且耗时，这促使人们不断开发能够在不询问受试者的情况下近似估计 MOS 值的客观估计器。此类估计器的开发与评估依赖于包含媒体文件及其相关主观测试结果的数据库。客观估计器的性能通常通过将其估计值与相应的主观测试结果进行比较来衡量，所使用的指标包括均方误差（MSE）和皮尔逊相关系数（PCC）。均方根误差（RMSE）是均方误差的一种常见替代指标；为方便起见，本文将在二者之间交替使用。

主观测试的结果通常被视为“真实数据”，但事实上，它们是具有固有不确定性的噪声测量值。MOS 值的一个噪声来源是主观评分量表的离散性。另一个来源是为每个信号评分的受试者数量有限。此外，个体偏差以及一般测试设计方面的考量也会带来额外噪声，例如测试中所包含的质量分布，以及测试环境的控制程度。在开发和评估客观估计器的性能时，必须承认这些固有噪声。通过考虑 MOS 值中的固有不确定性，我们可以更准确地理解客观估计器在多大程度上能够匹配这一“真实数据”。

为满足这一需求，我们推导了 MSE 和 PCC 的数学界限，这些界限完全基于通过 MOS 估计质量所受到的限制。这些界限纳入了每个文件的投票数以及测试中的质量分布。只要 MOS 值趋近于真实质量，这些界限便适用，无论投票乃至 MOS 是如何产生的。这些界限有助于判断客观估计器是否还有改进空间，或者考虑到所用主观测试的局限性，它是否已达到预期所能达到的最佳水平。

正如人们所料，这些界限需要了解构成平均意见分（MOS）的个体投票的方差。我们将其称为投票方差，它决定了掩盖“信号”的“噪声”。有投票方差信息时，这些界限能让数据“自行说明”，而无需任何额外假设。但投票方差信息并非总会共享（一些常用于训练客观语音质量估计器的大型众包主观测试便是如此

）。在这种情况下，投票模型可以提供计算界限所需的方差信息。因此，我们还提出了一个简单且易于理解的主观投票模型。该模型基于二项分布，因此我们将其命名为 BinoVotes。正如对真实投票取平均值可得到平均意见分（MOS），对 BinoVotes 取平均值则可得到 BinoMOS。我们展示了 BinoVotes 和 BinoMOS 具有多项符合直觉的理想属性，并且 BinoVotes 与真实投票相符，因为两者的均方误差（MSE）界和皮尔逊相关系数（PCC）界十分接近。

主观测试已得到深入研究，其中广受欢迎的五级评分量表及其产生的 MOS 值尤其受到关注。 [12] 详细讨论了它们的适用范围和局限性，也为读者提供了大量相关研究的参考文献。为了更接近潜在的真实值，研究人员提出了多种替代方案，而非简单地对个体投票取平均值。一种方案是建立最大似然估计问题，然后通过置信传播 [13]或交替投影 [14]求解。其他方案包括采用广义线性模型 [15], 、迭代参数估计算法 [16], 、基于 Z分数的方法 [17],、基于熵的主观质量恢复算法 [18],以及正则化最大似然估计算法[19]。由于实践中永远无法得知潜在的真实值，因此，对这些取代均值的方案的支持或反对，必然基于各种模拟，以及它们对投票所受特定类别受控扰动的稳健性。

此前已有研究使用 α稳定分布 [20],高斯分布[21], 以及高斯混合分布 [21]对投票分布进行建模。 [22] 和 [23] 对投票和 MOS 值中固有的噪声进行了广泛研究和建模。此外，研究人员还提出了多种新统计量，旨在评估客观估计器时恰当地考虑这种噪声。 [24]-[26],提出了基于分类错误的方法； [27],引入了分辨能力； [28]则描述了 ε-不敏感 RMSE。 [29] 中的方法高度依赖数据。该方法利用分类错误率，将客观估计器描述为等同于由 1、2、3、6 或 9 名受试者参与的主观测试。

这些研究路径往往有效且有用，但在此我们提出一种独特的替代方案。我们认同这样一条简单且很有说服力的公理：随着投票数增多，投票均值会收敛于文件内容的真实质量。我们不再为客观估计器提出额外的专门评价指标，而是为两种现有且常用、具有充分理论依据的性能度量推导界限：皮尔逊相关系数和均方误差。

我们工作的另一个特点是，无需建立 MOS 不确定性模型。这一点非常可取，因为此类模型可能存在问题——它们可能违背 MOS 值的离散性或有限取值范围，而截断等临时补救措施又可能带来新的复杂问题，包括引入新的偏差。我们的工作不存在这些问题，因为它以 MOS 真正的数学属性为坚实基础，包括其离散性，以及允许的离散 MOS 值集合取决于计算某个 MOS 值所用的投票数这一事实。基于此，我们推导出 PCC 和均方误差的期望值。据我们所知，尚无以投票和 MOS 的内在属性为唯一依据来界定 PCC 或均方误差的研究。

下一节将介绍评分量表的性质，以及平均意见分（MOS）从这些量表继承的性质。在第三节中，我们利用这些性质推导皮尔逊相关系数（PCC）和均方误差（MSE）的界限。第四节提出一种基于新二项分布的投票模型 BinoVotes，并在第五节介绍由此自然产生的 BinoMOS。第六节推导 BinoVotes 投票模型下 PCC 和 MSE 的界限。第七节介绍三种根据实际主观测试结果估计界限的方法。最后，第八节考察了 22 项不同主观测试的结果。其中 18 项包含投票方差信息，因此我们可以据此构建完全由数据驱动的 PCC 和 MSE 界限，并将其与假设采用 BinoVotes 投票模型以及使用基于经验的投票方差假设所得的界限进行比较。其余四项测试结果不包含投票方差信息，我们采用两种不同方法计算 PCC 和 MSE 的界限。

II. 评分量表和平均意见分的数学性质

现在，我们来考察对任意媒体信号进行评分的情况。可以评分的属性有很多种，但最常见的是“总体质量”，第八节中的所有主观测试结果都涉及“总体质量”。为简洁起见，以下使用“质量”一词，但所有数学结果均与所评属性无关。未来的研究也可以将第八节中的比较扩展到对其他属性的评分。

我们将媒体文件内容的真实质量记为 $Y.$ 。我们定义某个连续质量量表，使得 $Y \in [s_L, s_H]$。为了在主观测试中评估 Y，我们将评分量表定义为

$$S = \left\{ \left(\frac{s_H - s_L}{n_s - 1} \right) k + s_L \right\}_{k=0}^{n_s - 1}. \tag{1}$$

评分量表是一组 n_s 个离散值，覆盖连续质量量表。在主观测试中，每位受试者根据自己对媒体文件的感知和看法，通过选择或“投票选择”其中一个 n_s 值，对给定媒体文件的质量进行评分。因此， n_s 决定了单次投票的分辨率。请注意，特殊情况 $s_L = 1$、 $s_H = 5$ 和 $n_s = 5$ 对应非常流行的五点评分主观测试量表，有时称为“MOS 量表”。质量存在于连续量表上，而主观测试会产生一组投票，这些投票对应于该量表上的 n_s 个离散点。

在主观测试中， n_v 名独立投票者或受试者对媒体文件的质量进行评估。我们将这些投票或评分记为 $\{R_j\}_{j=1}^{n_v}$，其中 $R_j \in S$。媒体文件的 MOS 定义为

$$X = \frac{1}{n_v} \sum_{j=1}^{n_v} R_j. \tag{2}$$

由（1）和（2）可知，MOS 值的定义域为

$$M = \left\{ \left(\frac{s_H - s_L}{n_v(n_s - 1)} \right) k + s_L \right\}_{k=0}^{n_v(n_s-1)}, \tag{3}$$

因此 $X \in M$。这一结果清楚表明，在固定评分量表 S 的情况下，随着每个媒体文件的投票数 n_v 增加，可能的 MOS 值数量及 MOS 值的分辨率都会提高。因此，限制 MOS 与真实质量 Y 达成一致的因素之一，就是每个文件的投票数。例如，在通常使用的 1 到 5 质量量表上，如果某个文件的质量为 3.30 且只有一票，则可能达到的最小误差为 0.30。但有两票时，最小可能误差会降至 0.20；有三票时，该界约为 0.03。

这样，对于固定的评分量表 S而言，每个文件的投票数是造成 MOS 与真实质量之间差异的最明显因素。其他差异则来自受试者的随机行为，以及投票在质量范围内的分布方式。如果真实质量为 3.3，受试者并不会只按照适当的比例投票为质量 3 或 4。相反，通常情况下，他们会使用整个质量量表范围。这会增加投票方差，进而增加平均意见分方差，而真正能够缓解这一问题的方法只能是增加每个文件的投票数。增加每个文件的投票数可以用中心极限定理来理解；从本质上说，这相当于降低 MOS 分布的方差。

我们将受试者组针对单个文件提交的个体投票 R_j 建模为相互独立但具有相同分布（i.i.d.）的随机变量。若按受试者对该投票分布进行条件化，则条件分布可能取决于受试者，这通常称为“个体受试者偏差”。这与将给定文件的所有受试者的所有投票 建模为来自单一分布、且独立于各个受试者的独立同分布随机变量完全一致。

如果给定真实质量，投票的期望值就是该质量，我们称投票的数学模型是“良态的”，

$$\mathbb{E}\left(R_j|Y\right) = Y. \tag{4}$$

在本文中，我们假设投票来自良态模型；事实上，这一假设是所有主观测试的基石。

如果投票是良态的，那么 MOS 也具有良态。事实上，当且仅当投票具有良态时，MOS 才具有良态，这一点将在本节末尾得到证明。投票 R_j 是随机变量，其均值 X 也是一个随机变量。根据（2），我们看到，给定真实质量 Y 时，MOS X 的均值也是真实质量

$$\mathbb{E}\left(X|Y\right) = Y. \tag{5}$$

注意到，全期望定律进一步表明

$$\mathbb{E}\left(R_j\right) = \mathbb{E}\left(\mathbb{E}\left(R_j|Y\right)\right) = \mathbb{E}\left(Y\right), \tag{6}$$

$$\mathbb{E}\left(X\right) = \mathbb{E}\left(\mathbb{E}\left(X|Y\right)\right) = \mathbb{E}\left(Y\right). \tag{7}$$

投票 R_j 也具有一定的方差。评分量表是有限的，而要在量表的任一端满足（5），唯一的方法就是使那里的方差为零。但方差不可能在整个评分量表上都为零。因此，我们将条件投票方差函数 $v_r(Y)$定义为

$$v_r(Y) = \mathrm{Var}\left(R_j|Y\right), \quad \forall j = 1, \cdots, n_v. \tag{8}$$

然后，可以利用全方差定律和（5）求得主观测试中 MOS 的方差，

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left(\mathrm{Var}\left(X|Y\right)\right) + \mathrm{Var}\left(Y\right). \tag{9}$$

利用（2）、（8）以及评分 R_j 相互独立这一事实，可以证明 MOS 的条件方差为

$$\mathrm{Var}\left(X|Y\right) = \frac{1}{n_v^2} \sum_{j=1}^{n_v} \mathrm{Var}\left(R_j|Y\right) = \frac{v_r(Y)}{n_v}. \tag{10}$$

将（10）代入（9），可得 MOS 的方差为

$$\mathrm{Var}\left(X\right) = \frac{\mathbb{E}\left(v_r(Y)\right)}{n_v} + \mathrm{Var}\left(Y\right). \tag{11}$$

该结果表明，MOS 的方差等于真实质量分布的方差 Var(Y) 加上一个反映投票方差的项。随着投票数增加，这一附加方差会相应减小。因此，无论投票引入了多大的方差，只要收集足够多的独立投票，MOS 值就会趋近于真实质量值，MOS 的方差也会趋近于真实质量分布的方差。

如上所述，受试者可能表现出个性化的投票行为或偏差。尽管我们目前的建模并未要求加入任何显式的偏差项，但这与我们对投票和 MOS 的建模完全兼容。本节最后将直接讨论在我们所构建的投票模型中的个体受试者偏差。

进行主观测试的人都认同，MOS 值具有良好的性质——只要收集足够多来自偏差各异的投票者的投票，MOS 就会收敛于真实质量。事实上，当且仅当投票具有良好的性质时，MOS 才具有良好的性质。这一点可通过观察以下内容得出。每位受试者的偏差可描述为受试者 V_j 给出的投票具有偏差期望值，如下所示

$$\mathbb{E}\left(R_j|Y, V_j\right) = Y + \delta_j, \tag{12}$$

其中，δ_j 是一个独立同分布的随机变量。对式 (12) 应用全期望定律，得

$$\mathbb{E}\left(R_j|Y\right) = \mathbb{E}\left(\mathbb{E}\left(R_j|Y, V_j\right)\right) = Y + \mathbb{E}\left(\delta_j\right). \tag{13}$$

全期望定律还表明，由具有独立同分布个体偏差的投票得出的 MOS 的期望值为

$$\mathbb{E}\left(X|Y\right) = Y + \mathbb{E}\left(\delta\right). \tag{14}$$

如果我们假设 MOS 表现良好，那么 (14) 意味着$\mathbb{E}(\delta) = 0$，且 (13) 可简化为 (4)。这意味着，带有个体偏差且能产生表现良好的 MOS 值的投票模型，也会产生表现良好的投票。反之，如果我们假设投票表现良好，那么 (13) 意味着$\mathbb{E}(\delta_j) = 0, \quad \forall j$，并且从 (14) 可以看出，MOS 必须

表现良好。在本研究中，个体偏差唯一相关的影响是它们会增大条件投票方差， $v_r(Y)$。

III. MOS 客观估计器的性能界

为了确定客观估计器的性能界，我们考虑可能达到的最佳估计器。这种最佳情况下的客观估计器将是能够访问真实质量 Y 的预言机。这意味着，真实质量Y 与 MOS X 之间的一致性，为任何客观估计器与 MOS 之间的一致性提供了界限。具体而言， Y 与X 之间的 PCC 为实际客观估计器所能达到的 PCC 提供了上界。而 Y 与 X 之间的 MSE，则为实际客观估计器所能达到的 MSE 提供了下界。 [1]

现在，我们考虑一个主观测试，其中包含一组 n_f 个媒体文件 $\{F_i\}_{i=1}^{n_f}$，其对应的真实质量为 $\{Y_i\}_{i=1}^{n_f}$，且文件 F_i 的质量为 Y_i。设 Y_i 按照某个任意的连续分布 f_Y 独立同分布，该分布定义在 $[s_L, s_H]$ 上。每个文件F_i 均由 n_v 名受试者进行评价，他们给出的投票表现良好，因此 MOS 也表现良好，并且满足 (5)。MOS 值X_i 独立同分布，而 X_i 是我们对 Y_i所能获得的最佳测量结果。在实际主观测试中，无法访问描述真实质量的隐变量 Y_i 。获取有关该变量信息的唯一途径，是通过投票及其产生的 MOS 值 X_i。

A. 均方误差下界

X_i 与 Y_i 之间的平方误差是随机变量

$$\epsilon_i^2 = (X_i - Y_i)^2. \tag{15}$$

均方误差也是一个随机变量，是这些平方误差的均值，

$$D^2 = \frac{1}{n_f}\sum_{i=1}^{n_f}\epsilon_i^2. \tag{16}$$

均方误差的期望值为

$$\mathbb{E}\left(D^2\right) = \frac{1}{n_f}\sum_{i=1}^{n_f}\mathbb{E}\left(\epsilon_i^2\right) = \mathbb{E}\left(\epsilon^2\right), \tag{17}$$

其中，最后一个等式利用了 ϵ_i^2 独立同分布这一事实。利用全期望定律，我们可以证明

$$\begin{aligned}\mathbb{E}\left(\epsilon^2\right) &= \mathbb{E}\left((X-Y)^2\right) = \mathbb{E}\left(\mathbb{E}\left((X-Y)^2|Y\right)\right)\\ &= \mathbb{E}\left(\mathbb{E}\left((X-\mathbb{E}(X))^2|Y\right)\right) = \mathbb{E}\left(\mathrm{Var}\left(X|Y\right)\right).\end{aligned} \tag{18}$$

根据(10)、(18)和(17)，我们发现均方误差的期望值为

$$\mathbb{E}\left(D^2\right) = \mathbb{E}\left(\epsilon^2\right) = \frac{\mathbb{E}\left(v_r(Y)\right)}{n_v}. \tag{19}$$

[1]有时会在估计器之后应用非线性映射，以改善与平均意见分的一致性。此类映射与使用我们在此推导的界限并不冲突。推导这些界限时，我们考虑了预言机估计器；事实上，它生成的估计值就是真实质量。预言机估计器无需映射，由此得出的界限与 MOS 值中的噪声直接相关。

这是 MOS 与真实质量之间均方误差的期望值。按照本节开头的说明，这也是 MOS 与任何质量客观估计器之间均方误差期望值的下界。

(11) 的结果为 (19) 中的均方误差结果提供了另一种解释：$\mathbb{E}\left(D^2\right) = \mathrm{Var}\left(X\right) - \mathrm{Var}\left(Y\right)$。这有着十分直观的解释：期望均方误差就是主观测试结果（MOS 值）在真实质量分布上引入的额外方差。

B. 总体的皮尔逊相关系数上界

现在我们推导一个基于总体的皮尔逊相关系数界。当然，每项主观测试的样本数都是有限的，因此对样本相关系数给出界限似乎更合适。然而，样本相关系数的期望值会迅速收敛到总体相关系数，这一点将在第六节-B 中展示。

对于总体，皮尔逊相关系数定义为

$$\rho(X,Y) = \frac{\mathbb{E}\left(XY\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}}. \tag{20}$$

利用全期望定律和 (5)，我们可以得出

$$\mathbb{E}\left(XY\right) = \mathbb{E}\left(\mathbb{E}\left(XY|Y\right)\right) = \mathbb{E}\left(Y\mathbb{E}\left(X|Y\right)\right) = \mathbb{E}\left(Y^2\right). \tag{21}$$

由 (7) 可知 $\mathbb{E}(X) = \mathbb{E}(Y)$。利用这一事实以及 (21)，代入 (20) 可得

$$\rho(X,Y) = \frac{\mathrm{Var}\left(Y\right)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}} = \sqrt{\frac{\mathrm{Var}\left(Y\right)}{\mathrm{Var}\left(X\right)}}. \tag{22}$$

最后，代入式 (11) 可得

$$\rho(X,Y) = \sqrt{\frac{\mathrm{Var}\left(Y\right)}{\mathrm{Var}\left(Y\right) + \frac{1}{n_v}\mathbb{E}\left(v_r(Y)\right)}} \tag{23}$$

$$= \sqrt{\frac{\mathrm{Var}\left(X\right) - \frac{1}{n_v}\mathbb{E}\left(v_r(Y)\right)}{\mathrm{Var}\left(X\right)}}. \tag{24}$$

当基于假设的真实质量分布 Y 进行分析时（如第 IV、V 和 VI 节），式 (23) 很有用。当基于主观测试结果进行分析时（如第 VII 和 VIII 节），式 (24) 很有用；主观测试结果会产生 MOS 分布 X。在实践中， Y 始终是一个隐藏的随机变量；然而，式 (7) 和 (11) 使我们能够根据 MOS 分布和投票方差信息估计其一阶矩和二阶矩。或者，用均方误差表示时，可得

$$\rho(X,Y) = \sqrt{\frac{\mathrm{Var}\left(Y\right)}{\mathrm{Var}\left(Y\right) + \mathbb{E}\left(D^2\right)}} = \sqrt{\frac{\mathrm{Var}\left(X\right) - \mathbb{E}\left(D^2\right)}{\mathrm{Var}\left(X\right)}}, \tag{25}$$

这些表达式强调，随着均方误差趋近于零，皮尔逊相关系数趋近于一。式 (25) 给出了 MOS 与真实质量之间的总体相关系数。按照本节开头的论述，这也是 MOS 与任意质量客观估计器之间相关性的上界。 [2]

[2]斯皮尔曼等级相关系数（SRCC）通常与PCC一并考虑。未来的工作可以将这些界限扩展到包括SRCC等替代性一致性统计量。

现在，我们已经得到了均方误差（19）和PCC（25）的界限。均方误差界只需要个体投票的预期方差和投票数。PCC界需要均方误差界，以及MOS 值的方差X 或底层真实质量分布的方差Y。这些界限提供了有关主观测试固有限制性能的信息。换言之，它们说明了MOS的离散性质如何限制不同主观测试所能达到的均方误差和PCC值。当客观估计器针对某项主观测试生成的结果接近这些界限时，即使均方误差并不接近零，或相关性并不接近一，也很可能已经没有多少真正改进的空间了。

IV. 投票的二项分布模型——BinoVotes

现在，我们介绍一种简单、直观且有效的主观投票模型。该模型遵循评分量表有限且离散的特性，因此得到的 MOS 值也同样保持这些属性。我们认为，这比使用连续随机变量近似投票或 MOS 要优越得多。

我们通过二项分布对个体投票或评分进行建模，这一分布非常适合该问题。它是一种取值范围有限的离散分布，也是正态分布的离散类比。鉴于我们的投票模型源自二项分布，我们将其命名为“BinoVotes”。

二项分布描述了在 n 次独立试验中，以成功参数 p 进行伯努利试验时取得给定成功次数的概率。对于随机变量$B \sim \text{Binomial}(n,p)$，其概率质量函数（PMF）为

$$P(B=k)=\binom{n}{k}p^k(1-p)^{(n-k)},\quad k=0,1,\cdots,n. \tag{26}$$

期望值和方差分别为

$$\mathbb{E}(B)=np, \tag{27}$$

$$\text{Var}(B)=np(1-p). \tag{28}$$

我们通过将试验次数 n设为比评分量表中的离散等级数少 1，即 $n=n_s-1$，将二项分布的定义域调整为与式 (1) 给出的任意评分量表相匹配。单个受试者对应的 BinoVotes 随机变量 j为

$$R_j=\frac{s_H-s_L}{n_s-1}B_j+s_L,\quad j=1,\ldots,n_v, \tag{29}$$

其中， B_j 是二项分布随机变量

$$B_j\sim\text{Binomial}\left((n_s-1),\frac{Y-s_L}{s_H-s_L}\right). \tag{30}$$

这里， $Y\in[s_L,s_H]$ 是表示真实质量的随机变量，而 Y取自任意分布 f_Y。我们称 R_j ～为 BinoVotes(n_s， Y)。请注意， R_j 的定义域为评分量表 S 定义见式 (1)。

BinoVotes 随机变量 R_j 满足式（4）中给出的良态投票模型条件。利用式（27）中的期望值以及式（29）和（30）中的定义即可证明这一点：

$$\mathbb{E}(R_j|Y)=s_L+\frac{(s_H-s_L)(n_s-1)(Y-s_L)}{(n_s-1)(s_H-s_L)}=Y. \tag{31}$$

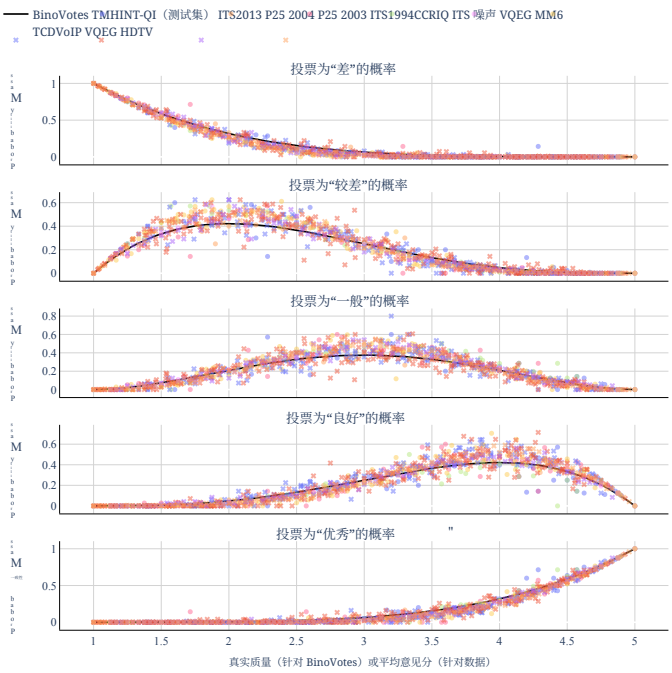


图1. BinoVotes 模型对通用 MOS 量表五个评分的概率（相对于真实质量 Y）以及10次测试的数据（相对于平均意见分）。

利用式（29）和（30）中的定义，将式（28）代入即可得到投票 R_j 的条件方差:

$$\begin{aligned}\text{Var}(R_j|Y)&=\left(\frac{s_H-s_L}{n_s-1}\right)^2\frac{(n_s-1)(Y-s_L)(s_H-Y)}{(s_H-s_L)^2}\\&=\frac{(Y-s_L)(s_H-Y)}{n_s-1}.\end{aligned} \tag{32}$$

因此，对于 BinoVotes，给出投票方差与真实质量之间函数关系的函数 $v_r(Y)$ 就是

$$v_r(Y)=\frac{(Y-s_L)(s_H-Y)}{n_s-1}. \tag{33}$$

从第三节可知，该函数的期望值将在下文中发挥非常重要的作用，

$$\begin{aligned}\mathbb{E}(v_r(Y))&=\frac{1}{n_s-1}\left((s_H+s_L)\mathbb{E}(Y)-\mathbb{E}\left(Y^2\right)-s_Ls_H\right)\\&=\frac{(\mathbb{E}(Y)-s_L)(s_H-\mathbb{E}(Y))-\text{Var}(Y)}{n_s-1}.\end{aligned} \tag{34}$$

选定评分量表后，BinoVotes 使用单个参数 Y 对投票进行建模。乍看之下，只有一个参数似乎是个缺点。但二项式方法实际上非常自然，因为这个参数以恰当的方式将投票的方差与投票的均值联系起来。具体而言，如式（33）所示，在质量量表的两端，方差必然为零；而在量表的中间位置，感知多样性和投票多样性可能最大，方差也达到最大值。这一点此前已在 $s_L=1$、 $s_H=5$ 的背景下进行过讨论，我们将在第VI-A节中把当前的 BinoVotes 结果与此前研究的这一方面联系起来。

图1展示了 BinoVotes 概率质量函数（以 Y 为自变量），以及投票分布（以平均意见分为自变量）；其中包括10项包含个体投票的主观测试，我们将在第八节中对其作更详细的介绍。对于每项测试，在给定平均意见分的情况下，分别绘出被标记为“差”“较差”“一般”“良好”或“优秀”的投票比例。该图表明，尽管 BinoVotes 并非完全准确，但确实捕捉到了受试者投票方式的正确趋势，而且二项分布是投票模型非常好的基础。

V. 由 BinoVotes 得到平均意见分——BinoMOS

按照惯例，我们对 BinoVotes 求平均值以得出 BinoMOS。也就是说，当 $R_j \sim \text{BinoVotes}(n_s,\ Y)$ 且 X 为

$$X = \frac{1}{n_v}\sum_{j=1}^{n_v} R_j, \tag{35}$$

我们称 $X \sim \text{BinoMOS}(n_v,\ Y)$。该分布还取决于 s_L、s_H 和 n_s，它们在 S 的定义及 Y 的支撑域中被隐式使用。

由于 BinoVotes 满足 (4) 和 (8)，因此 BinoMOS 随机变量满足 $\mathbb{E}(X|Y) = Y$(5)，且 $\mathbb{E}(X) = \mathbb{E}(Y)$ (7)。此外，我们可以利用 (11) 和 (34) 得出 BinoMOS 方差为

$$\text{Var}(X) = \frac{(\mathbb{E}(Y)-s_L)(s_H-\mathbb{E}(Y))+(n_m-1)\text{Var}(Y)}{n_m}, \tag{36}$$

其中 $n_m = n_v(n_s-1)$。请注意，X 可取的不同值总数为 n_m+1。增加每个文件的投票数 n_v 会增大 n_m，并降低 BinoMOS 方差，这正是我们所期望的。

接下来，我们求 BinoMOS 的 PMF。我们将利用这样一个事实：独立二项随机变量之和仍为二项随机变量。具体来说，将 n_v 个相互独立的 $\text{Binomial}(n_s-1, p)$ 随机变量相加，得到的随机变量服从 $\text{Binomial}(n_v(n_s-1),\ p)$，且该随机变量描述了 X，其尺度因子和平移因子与（29）中相同。由于 $X \sim \text{BinoMOS}(Y, n_v)$，$n_m+1$ 的允许取值为 X、$x_k = s_L + k\frac{s_H-s_L}{n_m} k = 0, 1, \cdots, n_m$。因此，这些允许取值各自的条件概率为

$$P(X = x_k|Y = y) = P(B = k), \tag{37}$$

其中 $B \sim$ 服从 $\text{Binomial}(n_m, \frac{y-s_L}{s_H-s_L})$。X 的概率质量函数可简化为

$$\begin{aligned} P(X = x_k) &= \int_{s_L}^{s_H} P(X = x_k|Y = y) f_Y(y)dy \\ &= \binom{n_m}{k} \frac{\int_{s_L}^{s_H}(y-s_L)^k(s_H-y)^{n_m-k} f_Y(y)dy}{(s_H-s_L)^{n_m}}. \end{aligned} \tag{38}$$

请注意，对于 Y 服从缩放和平移的贝塔分布这一特殊情况，BinoMOS 分布可用缩放和平移的贝塔-二项分布来描述。

VI. BinoMOS 客观估计器的性能界

现在，我们可以将第四节和第五节中的 BinoVotes 和 BinoMOS 结果应用于第三节的一般性能界，从而求得主观测试投票符合 BinoVotes 模型时客观估计器的性能界。将期望的 BinoVote 方差（34）代入一般均方误差界（19），即可得到均方误差下界：

$$\mathbb{E}\left(D^2\right) = \frac{(\mathbb{E}(Y)-s_L)(s_H-\mathbb{E}(Y)) - \text{Var}(Y)}{n_m}. \tag{39}$$

同样，将 BinoMOS 的均方误差（39）代入一般的 PCCbound（25），即可得到总体 PCC 上界：

$$\begin{aligned} \rho(X,Y) &= \sqrt{\frac{\text{Var}(Y)}{\text{Var}(Y)+\mathbb{E}(D^2)}} \\ &= \sqrt{\frac{n_m\text{Var}(Y)}{(n_m-1)\text{Var}(Y)+(\mathbb{E}(Y)-s_L)(s_H-\mathbb{E}(Y))}}. \end{aligned} \tag{40}$$

假设投票遵循 BinoVotes 模型后，我们现在可以得到均方误差和皮尔逊相关系数的界限，这些界限完全由底层真实质量分布 Y 的期望值和方差确定。

A. 通用 *MOS* 量表上的 *BinoVotes* 和 *BinoMOS* 示例

现在，我们给出一些适用于非常常见的 1 到 5 MOS 量表的具体 BinoVotes 和 BinoMOS 示例，即 $s_L = 1$、$s_H = 5$ 和 $n_s = 5$。在这种情况下，式 (29) 和 (30) 分别变为

$$R_j = B_j + 1, \tag{41}$$

$$B_j \sim \text{Binomial}\left(4, \frac{Y-1}{4}\right). \tag{42}$$

投票方差函数 (33) 变为

$$v_r(Y) = \frac{(Y-1)(5-Y)}{4}, \tag{43}$$

其期望值 (34) 为

$$\mathbb{E}(v_r(Y)) = \frac{(\mathbb{E}(Y)-1)(5-\mathbb{E}(Y)) - \text{Var}(Y)}{4}. \tag{44}$$

均方误差下界 (39) 变为

$$\mathbb{E}\left(\epsilon^2\right) = \frac{(\mathbb{E}(Y)-1)(5-\mathbb{E}(Y)) - \text{Var}(Y)}{4n_v}, \tag{45}$$

而皮尔逊相关系数上界（40）变为

$$\rho(X,Y) = \sqrt{\frac{4n_v\text{Var}(Y)}{(4n_v-1)\text{Var}(Y)+(\mathbb{E}(Y)-1)(5-\mathbb{E}(Y))}}. \tag{46}$$

我们使用公式 (38) 中的 BinoMOS 概率质量函数结果绘制图 2 中的示例概率质量函数。该图展示了当 f_Y 为贝塔分布（2, 2.5）时，三个 n_v 值对应的概率质量函数。与真实投票一致，BinoMOS 概率质量函数是对真实质量所对应底层连续分布的离散近似。随着每个文件的投票数

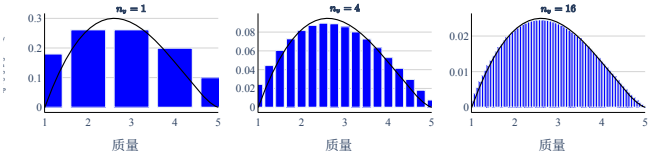


图 2. 真实质量分布为贝塔分布（2, 2.5）时的 BinoMOS 概率质量函数示例（蓝色；黑色曲线为真实质量分布，已缩放以便视觉比较）。每个文件的投票数 n_v 从左至右依次为 1、4 和 16。

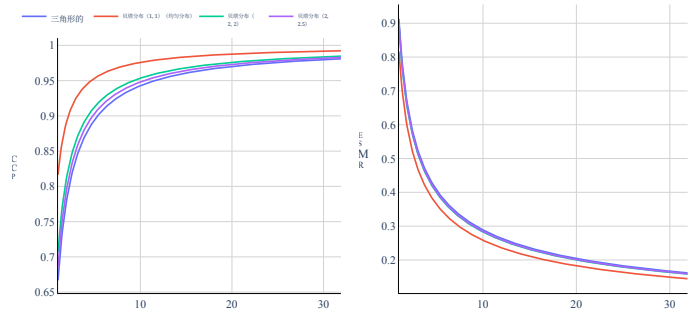


图 3. 四种真实质量分布 $f_Y(y)$ 对应的一致性统计量界限随每个文件的投票数 n_v 的变化。(a) PCC 上界。(b) RMSE 下界。三角形分布、贝塔分布（2, 2）和贝塔分布（2, 2.5）的界限曲线在视觉上难以区分。

增加，BinoMOS 值的分辨率也会提高，近似结果会更加接近。

图 3 展示了各种 f_Y 分布对应的 PCC 和 RMSE 界限随每个文件的投票数变化的情况。从图中可以清楚地看出，每个文件可用的投票越多，投票均值与客观估计器之间就越有可能实现更低的 RMSE 和更高的 PCC。底层质量分布的影响较小；对于三种非均匀分布，PCC 界限彼此接近，而 RMSE 界限几乎完全相同。

（43）中的抛物线（$Y-1$）（$5-Y$）决定了整个质量范围内的投票方差，这是 BinoVotes 模型直接带来的数学结果。在 [30] 中，对整个质量范围内可能出现的最大投票方差进行推导，得到的也是同一条抛物线。随后，我们将该抛物线的缩放版本拟合到 28 组涵盖广泛应用领域的主观测试结果上（见[30]中的表 1）。这 28 个缩放因子的中位数为 0.24，与（43）中的 $\frac{1}{4}=0.25$ 缩放因子十分接近。也就是说，BinoVotes 方差函数处于 [30]所研究的 28 组方差与平均意见分关系的“中间位置”。 [21] 中的图像质量评估工作要求将这条抛物线拟合到六组主观图像质量测试结果上。所得中位数为 0.15，略低于 BinoVotes 投票方差函数中的缩放因子。在 [16] 中，同一条抛物线的缩放版本被引入作为“不一致性因子”，用来“模拟这样一种事实：我们预期量表中间位置的不一致性较大，而越接近两端，不一致性就必须逐渐减小，直至

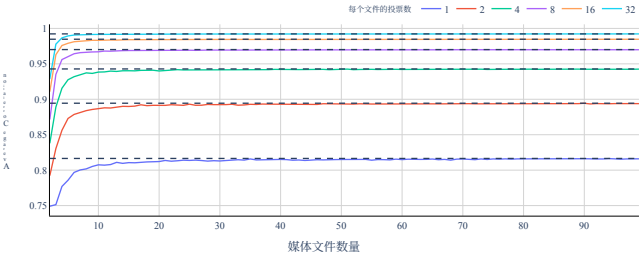


图 4. 根据式（46）得出的 BinoVotes 总体相关性界限（黑色虚线），以及通过模拟得到的 BinoVotes 样本相关系数（实线，颜色各异）与样本中文件数量的关系。样本相关系数取 10,000 次重复实验的平均值，该平均值迅速趋近于总体相关性界限。真实质量分布 f_Y 为均匀分布。

零。”BinoVotes 模型天然实现了这一点。

B. 样本的皮尔逊相关系数

对于包含 n_f 个媒体文件且对应真实质量值为 n_f的给定主观测试，皮尔逊相关系数确实取决于 n_f。样本相关系数是一个随机变量，能够反映这种 n_f 依赖关系，而总体相关系数（40）则不能。但图 4 显示，随着n_f 增加，样本相关系数的期望值会迅速收敛至总体相关系数。每个文件的投票数越多，收敛速度越快。无论每个文件的投票数是多少，对于任何规模合理的主观测试（例如 $50 \leq n_f$），样本相关系数与总体相关系数之间的差异都微不足道。因此，我们认为所推导的基于总体的皮尔逊相关系数界限在所有实际应用中都相当令人满意。

VII. 对主观测试结果的应用

到目前为止，我们主要关注的是在给定真实质量的情况下对主观投票进行建模。仅基于关于投票行为的基本假设，我们推导出了一般界限，用于衡量 MOS 在均方误差和皮尔逊相关系数方面与真实质量的一致程度，分别见式（19）和（25）。随后，我们提出了 BinoVotes 投票模型，并推导出 BinoMOS 结果与真实质量之间关系的界限，见式（39）和（40）。

现在，我们将重点转向数据集，并推导针对特定数据集上任意客观估计器的性能界限，以均方误差和皮尔逊相关系数衡量。[3] 我们再次指出，最佳的客观估计器是能够直接获取真实质量信息的预言机。此前的结果是根据真实质量分布给出的；现在，我们将推导仅使用 MOS 分布样本中可获取信息的结果。

对于任意主观测试结果，真实质量分布都是隐藏的，我们只能获取相应 MOS 分布的一个样本。因此，为了计算界限，我们必须先估计式（19）和（25）所需的临界值： $\mathbb{E}(v_r(Y))$ 和 $\mathrm{Var}(X)$。为方便起见，我们将这些值以及 $\mathbb{E}(X)$ 标记为

$$\mu_X = \mathbb{E}(X), \sigma_X^2 = \mathrm{Var}(X), \sigma_V^2 = \mathbb{E}(v_r(Y)). \quad (47)$$

[3]均方误差界和 BinoVotes 的软件实现可在 https://github.com/NTIA/its-mos-agreement 获取。

我们必须根据数据估计这些值，并将这些估计值标记为 $\hat{\mu}_X, \hat{\sigma}_X^2$ 和 $\hat{\sigma}_V^2$。对于每个文件有 n_v 个投票的数据集，均方误差界的估计值为

$$\hat{d}^2 = \frac{\hat{\sigma}_V^2}{n_v}, \tag{48}$$

皮尔逊相关系数界的估计变为

$$\hat{\rho} = \sqrt{\frac{\hat{\sigma}_X^2 - \frac{1}{n_v}\hat{\sigma}_V^2}{\hat{\sigma}_X^2}}. \tag{49}$$

现在考虑两种情况：对于每个文件，主观测试结果同时包含 MOS 和投票方差，$\{(x_i,\ \sigma_i^2)\}_{i=1}^{n_f}$；以及对于每个文件，主观测试结果仅包含 MOS 值，$\{x_i\}_{i=1}^{n_f}$。在这两种情况下，MOS 分布的均值和方差估计相同：

$$\hat{\mu}_X = \frac{1}{n_f}\sum_{i=1}^{n_f} x_i, \qquad \hat{\sigma}_X^2 = \frac{1}{n_f - 1}\sum_{i=1}^{n_f}(x_i - \hat{\mu}_X)^2. \tag{50}$$

当每个文件都提供投票方差信息时，期望投票方差的估计就是观测投票方差的平均值，我们将其定义为

$$\hat{\sigma}_{DV}^2 = \frac{1}{n_f}\sum_{i=1}^{n_f}\sigma_i^2. \tag{51}$$

当有投票方差信息时，使用式 (50) 和 (51) 给出的估计值来计算式 (48) 和 (49) 中的界。这些结果就是完全数据驱动的性能界。

当没有投票方差信息时，我们必须作出假设，以生成 $\hat{\sigma}_V^2$ 的值。一种方法是采用其他提供投票方差信息的测试所得的经验结果。一个自然的选择是取所有可用 $\hat{\sigma}_{DV}^2$ 值的全局平均值，并将其作为固定值，我们将其记为全局平均观测投票方差 $\hat{\sigma}_{GV}^2$。在这种情况下，$\hat{\sigma}_{GV}^2$ 将在式 (48) 和 (49) 中用作 $\hat{\sigma}_V^2$。

另一种方法是使用 BinoVotes 投票模型和观测到的 MOS 样本来估计平均投票方差。我们将 BinoVotes 模型下的平均投票方差定义为 $\hat{\sigma}_{BV}^2$。由 (34) 和 (7) 可知

$$\hat{\sigma}_{BV}^2 = \frac{(\hat{\mu}_X - s_L)(s_H - \hat{\mu}_X) - \hat{\sigma}_Y^2}{n_s - 1}. \tag{52}$$

通过重排 (11)，并使用 BinoVotes 投票方差估计值，可以得到

$$\hat{\sigma}_Y^2 = \hat{\sigma}_X^2 - \frac{\hat{\sigma}_{BV}^2}{n_v}. \tag{53}$$

将 (53) 代入 (52)，并求解 $\hat{\sigma}_{BV}^2$，可得 MOS 分布的给定样本在 BinoVotes 模型下的期望投票方差为

$$\hat{\sigma}_{BV}^2 = \frac{n_v}{n_m - 1}\left((\hat{\mu}_X - s_L)(s_H - \hat{\mu}_X) - \hat{\sigma}_X^2\right). \tag{54}$$

在这种情况下，$\hat{\sigma}_{BV}^2$ 将用作（48）和（49）中的$\hat{\sigma}_V^2$。

表 I 按每个文件的平均投票数排列的、包含方差信息的主观测试。†表示图像质量测试；††表示视频质量测试；其余均为语音质量测试。*表示对 [31] 中陆地移动无线电系统进行的测试。

Name	n_v	$\hat{\mu}_X$	$\hat{\sigma}_X^2$	$\hat{\sigma}_V^2$
TMHINT-QI（测试集） [32]	3.52	3.11	0.99	0.93
PSTN [33]	4.57	3.12	0.76	0.80
P25 2005*	5.20	3.05	1.51	0.48
NISQA 仿真平均意见分 [34]	5.24	2.99	1.20	0.54
ITS2013 [35]	5.64	3.41	0.60	0.83
ITS1997（ [36]中的测试16和17）	6.40	2.85	0.89	0.58
P25 2004*	7.47	2.95	0.87	0.73
P25 2003*	7.91	2.92	0.65	0.74
ITS1994（测试 4， [37]）	8.00	3.12	0.88	0.63
CCRIQ [38] †	9.85	2.89	1.10	0.63
ITS 噪声 [39] †	20.84	3.18	0.63	0.69
VQEG MM6 [40]††	21.30	3.06	1.15	0.66
NISQA LiveTalk MOS [34]	24.00	2.76	0.89	0.66
TCDVoIP [41]	24.00	3.22	0.99	0.62
VQEG HDTV [42]††	24.00	3.10	1.33	0.51
ITU-T P.Sup23 实验 2，意大利语 [43]	24.00	2.95	0.80	0.57
NISQA NSC MOS [34]	27.17	2.55	0.91	0.48
NISQA P501 MOS [34]	28.33	2.60	1.04	0.42

VIII. 18 项主观测试结果分析

现在，我们使用来自 2 项图像主观测试、2 项视频主观测试和 14 项语音主观测试的主观测试结果，计算均方根误差和皮尔逊相关系数的界限。我们选取这些测试作为示例，因为它们包含投票方差信息以及 MOS 值。投票方差信息可以原始投票的形式提供，也可以通过随每个 MOS 值附带的方差或标准差值提供。我们注意到，许多主观测试结果在共享时并未包含这些信息；若能将其纳入，将会显著提升结果的价值。

投票方差信息使我们能够计算完全数据驱动的性能界。我们将这些界限与第七节中介绍的两种替代方法进行比较。第一种方法是使用观测投票方差的全局平均值 $\hat{\sigma}_{GV}^2$ 来计算界限。第二种方法是使用 BinoVotes 模型和观测到的 MOS 样本，通过 (54) 得到投票方差 $\hat{\sigma}_{BV}^2$，并据此计算界限。这些比较展示了当我们不得不弥补投票方差信息缺失时，界限会如何变化。

表 I 和图 5 总结了 18项主观测试。每项测试的文件数（n_f）范围为 200 至 58,448，每个文件的平均投票数（n_v）范围为 3.52 至 28.33。MOS 平均值范围为 2.55 至 3.41，MOS 方差范围为 0.63 至 1.51。平均投票方差范围为 0.42 至 0.93。图 5 展示了 MOS 分布。我们仅使用覆盖完整 MOS 范围（1 至 5）的测试。更具体地说，我们从 1 到 5 创建了 8 个宽度为 0.5 的均匀分箱，并仅纳入所有 8 个分箱均非空的主观测试。

A. 一致性统计量界限比较

式（48）和（49）中的性能界取决于 MOS分布方差估计 $\hat{\sigma}_X^2$、预期投票方差估计 $\hat{\sigma}_V^2$以及每个文件的投票数 n_v。

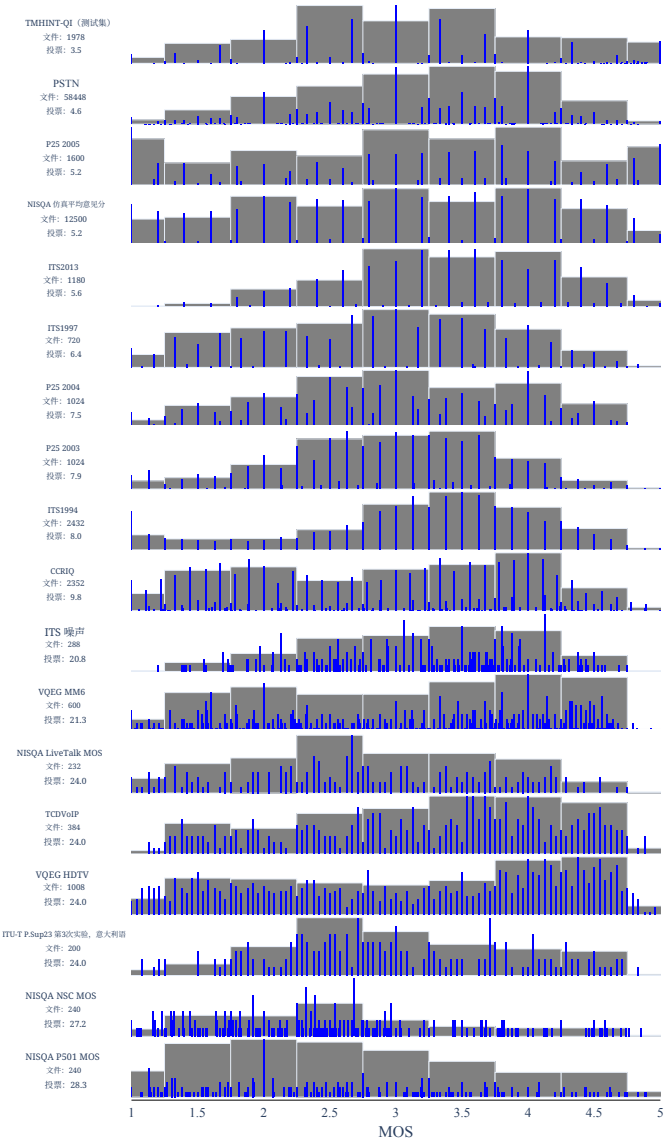


图5. 18项主观测试的MOS分布。低分辨率灰色直方图突出显示分布形状，高分辨率蓝色直方图突出显示分布分辨率，分辨率由每个文件的投票数决定。

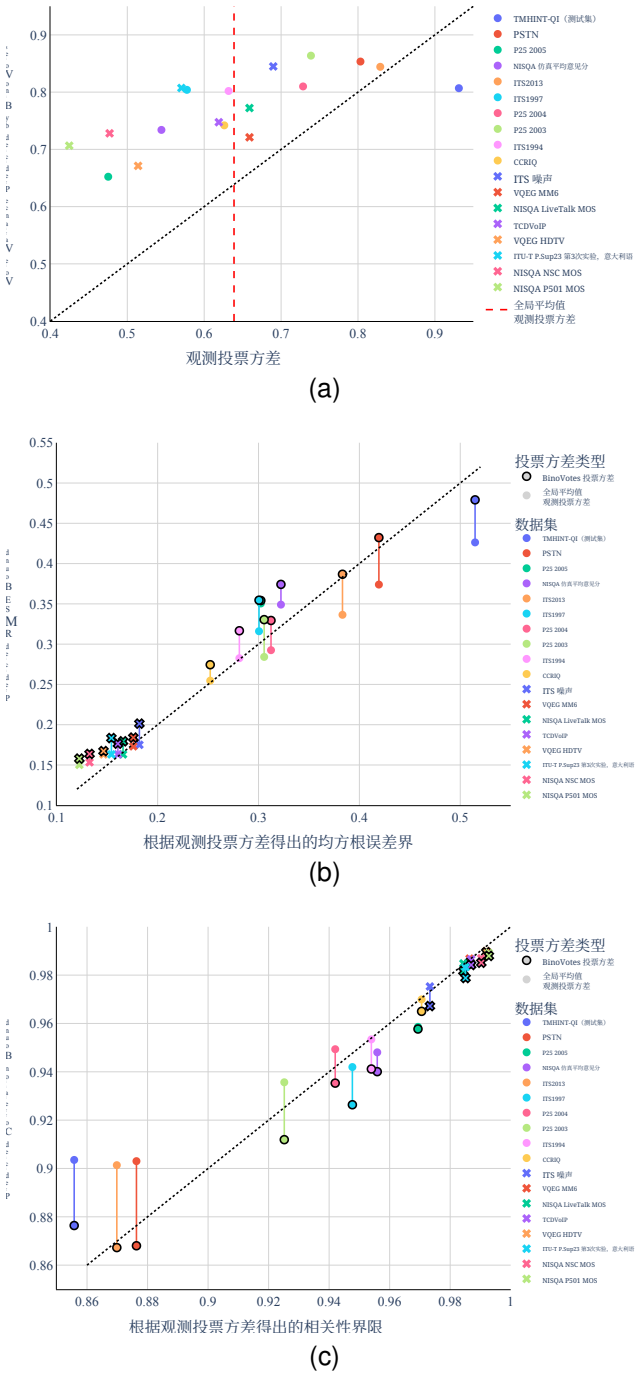


图 6. BinoMOS 结果与 18项主观测试结果的比较。（a）BinoVotes 下的期望投票方差与每项测试的平均观测投票方差的比较。（b）BinoVotes 下以及全局平均观测投票方差下的 RMSE 下界与完全数据驱动界限的比较。（c）BinoVotes 下以及全局平均观测投票方差下的 PCC 上界与完全数据驱动界限的比较。

对于第七节给出的三种计算界限的方法，估计值 $\hat{\sigma}_X^2$ 和数值 n_v 均相同。唯一的差异在于期望投票方差 $\hat{\sigma}_V^2$，因此我们先比较这些数值。

图6a显示了观测平均投票方差 $\hat{\sigma}_{DV}^2$以及 BinoVotes 模型给出的投票方差 $\hat{\sigma}_{BV}^2$。图中还显示了全局平均观测投票方差，这一数值在这18项测试中为 $\hat{\sigma}_{GV}^2 = 0.64$。各项测试中观测投票方差的标准差为0.13。观测平均投票方差的最小值为0.42，来自 NISQA P.501MOS；最大值为0.93，来自 TMHINT-QI（测试）。图中显示，在18项测试中的17项（除 TMHINT-QI 外），BinoVotes 得出的投票方差都高于数据中观测到的方差。对此的一种可能解释是，在测试过程中，受试者可能会把评分量表的两端留给那些媒体文件

对于属性最极端的情况，BinoVotes 则不会如此。换言之，大多数受试者都知道，中等质量的事物既不差也不优秀，因此投出这类票的可能性低于使用 BinoVotes 时。绝对差值最大的是 NISQA P501 MOS，为 0.28；此时 BinoVotes 预测的期望投票方差为 0.70。BinoVotes 预测的平均投票值与实际值之间的平均差异

方差与观测方差为 0.13。

尽管投票方差存在这些差异，BinoVotes 模型仍能提供非常有用的均方根误差和皮尔逊相关系数界限。图 6b 展示了使用 σ^2_{GV} 或σ^2_{BV}（y 轴）时均方根误差的下界，并与完全数据驱动的 σ^2_{DV}（x 轴）进行比较。基于BinoVotes的界始终大于基于全局平均值的界，而且通常也大于完全数据驱动界限；TMHINT-QI（测试集）再次是唯一的例外。基于全局平均值的界与完全数据驱动界限之间观测到的最大差值为 0.09，出现在 TMHINT-QI（测试集）上。这直接归因于该主观测试的观测投票方差最大，且与全局平均值的偏差也最大。基于全局平均值的界与完全数据驱动界限之间的平均差值为 -0.004，这表明尽管预测的界限并不完美，但仍具有参考价值。基于BinoVotes的界与完全数据驱动界限之间观测到的最大差值为 0.05，出现在 ITS1997 上，平均观测差值为 0.02。本文中的均方根误差界值范围为 0.12 至 0.51；与这些数值及其覆盖范围相比，此处提到的所有差值都较小。

图 6c 展示了与图 6b 相同的比较，但针对的是 PCC 上界，并呈现出类似的模式。基于BinoVotes的界始终低于基于全局平均值的界；BinoVotes 往往会低估相关性界限，而基于全局平均值的界有时会高估、有时会低估（这符合预期，因为其利用了全局平均观测投票方差的性质）。基于全局平均值的界所观测到的最大差异为 0.05，同样出现在 TMHINT-QI（测试集）中，而平均差异为 0.005。基于 BinoVotes 的界所观测到的最大差异为 0.021，同样出现在 ITS1997 中，而平均差异为 -0.006。作为背景参考，PCC 值的范围为 0.86 至 0.99。因此，尽管估计界限并不完美，但在我们没有投票方差信息的情况下，它们显然能够提供有价值的信息。

B. 不含方差信息的主观测试 *ion*

主观测试结果有时不会附带任何方差信息——其中仅包含 MOS 值和每个文件的投票数。表 II 列出了四项常用于推进语音质量估计研究的主观测试。对于这些测试，我们可以使用全局平均观测投票方差和 BinoVotes 模型，分别求得皮尔逊相关系数（46）和均方误差（45）的界限。需要注意的是，BinoVotes 适用于不同的评分量表，而全局平均观测投票方差 σ^2_{GV}必须基于采用单一评分量表的数据计算，且只能应用于采用相同量表的数据。由于 IU 数据不采用标准的 1 到 5 MOS 量表，因此我们未在表 II 中给出使用 $\hat{}\sigma^2_{GV}$ 得到的 IU 界限估计。

IX. 结论

我们推导出了主观测试结果集与这些结果的客观估计器之间一致性的界限。具体而言，我们推导出了皮尔逊相关系数的上界和均方误差的下界。唯一的假设是

表 II 不包含投票方差信息的四项主观测试。仍可使用 BINOVOTES 或固定的$\hat{\sigma}^2_{GV} = 0.64$ 值计算界限。

数据集	$\hat{\mu}_X$	$\hat{\sigma}^2_X$	n_v	RMSE 界		皮尔逊相关系数界	
				BinoVotes	固定	BinoVotes	固定
VCC18	2.92	0.79	4	0.46	0.40	0.85	0.89
IU	5.25	4.56	5	0.64	NA	0.95	NA
VMC22	2.93	0.85	8	0.32	0.28	0.94	0.95
腾讯	2.85	1.38	20	0.18	0.18	0.99	0.99

对于任何经过测试的媒体文件，其主观测试响应均来自一个分布，且该分布的期望值就是真实的媒体文件值。

这些结果具有很强的实用价值。对于任意给定的主观测试，都可以利用 (48) 和一组简单的计算，求出任何客观估计器在该测试中能够达到的均方误差下界。同样，也可以利用 (49) 和一组简单的计算，求出任何客观估计器在该测试中能够达到的皮尔逊相关系数上界。这些界限具有重要价值，因为对于每个数据集，它们都为设计客观估计器的研究人员提供了适当的目标。研究人员自然会努力追求均方误差为零、皮尔逊相关系数为一的目标，但这无法实现。本文给出的界限为每个数据集提供了切实可行且恰当的目标。

对于包含投票方差信息的数据集，这些界限完全由数据驱动，仅取决于观测平均投票方差和平均意见分分布的方差。请注意，观测平均投票方差包含个体受试者偏差的影响——偏差越大，观测平均投票方差越大，进而使皮尔逊相关系数上界降低、均方误差下界升高。

对于不提供投票方差信息的数据集，我们可以假设投票来自受二项分布启发的BinoVotes过程，然后估计界限。我们考察了18项提供投票方差信息的主观语音和视频测试的结果，涵盖了总计86,450个文件和超过493,000次投票。我们发现，基于BinoVotes的界限通常准确且信息丰富，为不提供投票方差信息的数据集估计性能界限提供了依据。尽管BinoVotes产生的投票方差往往略高于数据中观测到的方差，情况依然如此。向BinoVotes 模型添加受试者偏差会增大其产生的投票方差，从而进一步拉大模型输出与观测数据之间的差距。由于界限取决于投票方差，这些界限也会偏离最准确的结果。此外，在任何BinoVotes高估投票方差可能造成问题的情况下，我们都可以改用此处提供的全局平均观测投票方差$\hat{}\sigma^2_{GV}$。

请注意，这些界限不适用于基于机器学习的估计器在训练过程中遇到的训练数据——估计器当然可以学习训练数据中的所有噪声。但在未见的非训练数据上使用估计器时，这些界限则适用。而且这些界限实际上是期望值，因此客观估计器偶尔超过它们是有可能的。但它不应能够持续

超过它们。

我们已将此处得出的结果应用于 22 项不同的主观测试，这一数量颇为可观，但并不全面。未来的潜在工作可以纳入更多数据集，并重点关注视频、图像和多媒体领域，而这些领域在本文中确实涉及不足。未来工作的另一个候选方向是，将此处 प्रस्तुत的结果扩展到研究界关注的其他一致性统计量。该工作的首选对象将是斯皮尔曼秩相关系数。

参考文献

[1] ITU-T P.800 建议书《传输质量主观测定方法》，国际电信联盟：日内瓦，1996年。[2] ITU-T P.805 建议书《会话质量的主观评估》，国际电信联盟：日内瓦，2007年。[3] ITU-T P.806 建议书《采用多重评分量表的主观质量测试方法》，国际电信联盟：日内瓦，2014年。[4] ITU-T P.808 建议书《采用众包方法评估语音质量》，国际电信联盟：日内瓦，2021年。[5] 国际电信联盟无线电通信部门 BS.1116-3 建议书《音频系统轻微劣化的主观评估方法》，国际电信联盟：日内瓦，2015年。[6] 国际电信联盟无线电通信部门 BT.500-15 建议书《电视图像质量主观评估方法》，国际电信联盟：日内瓦，2023年。[7] ITU-T P.918 建议书《基于维度的视频内容主观质量评估》，国际电信联盟：日内瓦，2020年。[8] ITU-T P.915 建议书《3D视频质量主观评估方法》，国际电信联盟：日内瓦，2016年。[9] ITU-T P.910 建议书《多媒体应用的主观视频质量评估方法》，国际电信联盟：日内瓦，2025年。[10] ITU-T P.920 建议书《视听通信的交互式测试方法》，国际电信联盟：日内瓦，2000年。[11] ITU-T P.809 建议书《游戏质量的主观评估方法》，国际电信联盟：日内瓦，2018年。[12] R. C. Streijl、S. Winkler 和 D. S. Hands，“重新审视平均意见分（MOS）：方法与应用、局限性和替代方案”，《多媒体系统》，第22卷，第2期，第213–227页，2016年3月，DOI: 10.1007/s00530-014-0446-1。 [在线]。网址：http://link.springer.com/10 1007 00530 014 0446 1 。/s - - - Z. Li 和 C.

[13] G. Bampis，“从噪声测量值中恢复主观质量分数”，载于2017年数据压缩会议（DCC），2017年4月，第52–61页，ISSN：2375-0359。DOI: 10.1109/DCC.2017.26。

[ ] 7921900 在线。网

[14] 址：https://ieeexplore.ieee.org/document/ Z. Li、C. G. Bampis、L. Janowski 和 I. Katsavounidis，“主观实验中受试者行为的简易模型”，《电子成像》，第32卷，第11期，第131–1–131–1页，2020年，DOI: 10.2352/ISSN.2470-1173.2020.11.HVEI-131。 [在线]。网址：https://library.imaging.org/ei/articles/32 11 00010 / /art S. Pezzulli、

[15] M. G. Martini 和 N. Barman，“从主观测试中估计质量分数——超越受试者的 MOS”，《IEEE 多媒体汇刊》，第23卷，第2505–2519页，2021年，DOI: 10.1109/TMM.2020.3013349。 [在线]。网址：https://ieeexplore.ieee.org/document/9154548 [16] L. F. Tiotsop、A. Servetti 和 E. Masala，“在主观实验中考虑受试者特征变异性的评分模型”，载于2023年第15届多媒体体验质量国际会议（QoMEX），2023年6月，第43–48页，ISSN：2472-7814。DOI: 10.1109/QoMEX58391.2023.10178476。 [在线]。网址：https://ieeexplore.ieee.org/document/10178476

[17] J. Zhu、A. Ak、P. Le Callet、S. Sethuraman 和 K. Rahul，“ZREC：稳健恢复平均及百分位意见分”，载于2023年 IEEE 国际图像处理会议（ICIP），2023年10月，第2630–2634页，DOI: 10.1109/ICIP49359.2023.10222033。 [在线]。网址：https://ieeexplore.ieee.org/document/10222033

[18] A. Altieri、L. F. Tiotsop 和 G. Valenzise，“从含噪原始意见评分中恢复主观媒体质量：一种非参数视角”，《IEEE 多媒体汇刊》，第26卷，第9342–9357页，2024年，DOI: 10.1109/TMM.2024.3390113。 [在线]。网址：https://ieeexplore.ieee.org/document/10504622

[19] L. F. Tiotsop、A. Servetti、M. Barkowsky 和 E. Masala，“基于离散质量量表的主观实验受试者评分行为建模”，《IEEE 多媒体汇刊》，第26卷，第8742–8757页，2024年，DOI: 10.1109/TMM.2024.3382483。 [在线]。网址：https://ieeexplore.ieee.org/document/10480466

[20] Y. Gao、X. Min、W. Zhu、X.-P. Zhang 和 G. Zhai，“基于α稳定模型的图像质量评分分布预测”，《IEEE 视频技术电路与系统汇刊》，第33卷，第6期，第2656–2671页，2023年，DOI: 10.1109/TCSVT.2022.3229839。[21] Y. Gao、X. Min、Y. Cao、W. Lin、B. S. Lee 和 G. Zhai，“基于高斯混合分布的盲图像质量评估”，《IEEE 图像处理汇刊》，第34卷，第4660–4675页，2025年，DOI: 10.1109/TIP.2025.3586512。[22] L. Janowski 和 M. Pinson，“质量实验中受试者的准确度：理论受试者模型”，《IEEE 多媒体汇刊》，第17卷，第12期，第2210–2224页，2015年，DOI: 10.1109/TMM.2015.2484963。[23] M. Pinson，“评估图像、视频、语音或视听质量的主观测试和客观指标的置信区间”，NTIA，技术报告 TR-21-550，2020年，DOI: 10.70220/kk0d38dw。[24] S. Voran，“比较客观与主观语音质量测试的技术”，载于《语音质量评估研讨会论文集》，鲁尔大学，德国波鸿，1994年。[在线]。网址：https://its.ntia.gov/publications/details?pub=2852

[25] L. Krasula、K. Fliegel、P. Le Callet 和 M. Kl´ıma，“关于客观图像与视频质量模型的准确度：一种新的性能评估方法”，载于2016年第八届多媒体体验质量国际会议（QoMEX），2016年，DOI: 10.1109/QoMEX.2016.7498936。[26] A. Ragano、H. B. Martinez 和 A. Hines，“超越相关性：使用受限一致性指数评估多媒体质量模型”，《IEEE 多媒体汇刊》，第27卷，第5604–5616页，2025年，DOI: 10.1109/TMM.2025.3542991。[27] M. Brill、J. Lubin、P. Costa 和 J. Pearson，“视频质量指标的准确度与交叉校准：来自 ATIS/T1A1 的新方法”，载于《国际图像处理会议论文集》，第3卷，2002年，DOI: 10.1109/ICIP.2002.1038897。[28] ITU-T 建议书 P.1401《客观质量预测模型的统计评估、资格认定与比较方法、指标和程序》，国际电信联盟：日内瓦，2020年。[29] M. H. Pinson，“媒体质量比较的精度与可重复性：测量和新的统计方法”，《IEEE 广播汇刊》，第69卷，第2期，第378–395页，2023年，DOI: 10.1109/TBC.2023.3236528。[30] T. Hoßfeld、R. Schatz 和 S. Egger，“SOS：仅有 MOS 还不够！”，载于《第三届多媒体体验质量国际研讨会论文集》，2011年，第131–136页，DOI: 10.1109/QoMEX.2011.6065690。[31] “公共安全通信官员协会项目25数字陆地移动无线电”，https://www.apcointl.org/technology/interoperability/project-25/，[访问日期：10-23-2025]。[32] Y.-W. Chen 和 Y. Tsao，“InQSS：一种采用多任务学习网络的语音可懂度与质量评估模型”，载于《Interspeech论文集》，2022年10月，DOI: 10.21437/Interspeech.2022-10153。[33] G. Mittag、R. Cutler、Y. Hosseinkashi、M. Revow、S. Srinivasan、N. Chande 和 R. Aichner，“基于深度神经网络的无参考 PSTN 语音质量预测”，载于《Interspeech论文集》，2020年10月，DOI: 10.21437/interspeech.2020-2760。[34] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，“NISQA：一种用于基于众包数据集进行多维语音质量预测的深度 CNN 自注意力模型”，载于《Interspeech论文集》，2021年，第2127–2131页，DOI: 10.21437/Interspeech.2021-299。[35] S. D. Voran 和 A. A. Catellier，“话段内何时应允许语音编码质量提升？”，载于2013年 IEEE 国际声学、语音与信号处理会议，2013年5月，第8149–8153页，DOI: 10.1109/ICASSP.2013.6639253。[36] S. Voran，“感知语音质量的客观估计进展”，载于《1999年 IEEE 语音编码研讨会论文集：模型、编码器与误差准则》（目录号 99EX351），1999年6月，第138–140页，DOI: 10.1109/SCFT.1999.781510。[37] S. Voran，“感知语音质量的客观估计——第二部分：测量归一化块技术的评估”

,” *IEEE*

语音与音频处理汇刊，第7卷，第4期，第383–390页，1999年7月，DOI：10.1109/89.771260。[38] M. Saad、M. Pinson、D. Nicholas、N. Van Kets、G. Van Wallendael、R. Jaladi 和 P. Corriveau，“图像质量体验：一项针对消费者体验的主观测试”，电子成像，第2016卷，2016年2月，10 2352 2470 1173 2016 16 124 DOI：. /ISSN. - . . .HVEI- . [39] J. Dumke 和 M. Pinson，“ITSnoise：含传感器噪声的图像质量数据集”，NTIA，技术报告 TR-22-563，2022年，DOI：10.70220/gms9bj0x。[40] M. H. Pinson、L. Janowski、R. Pepion、Q. Huynh-Thu、C. Schmid- mer、P. Corriveau、A. Younkin、P. Le Callet、M. Barkowsky 和 W. Ingram，“受试者和环境对视听主观测试的影响：一项国际研究”，*IEEE* 信号处理精选主题期刊，第6卷，第6期，第640–651页，2012年，DOI：10.1109/JSTSP.2012.2215306。[41] N. Harte、E. Gillen 和 A. Hines，“TCD-VoIP：用于评估 VoIP 应用中质量的劣化语音研究数据库”，*2015*年第7届多媒体体验质量国际研讨会（*QoMEX*），2015年，DOI：10.1109/QoMEX.2015.7148100。[42] A. Webster 和 F. Speranza，“高清晰度视频内容视频质量模型验证报告”，视频质量专家组，技术报告，2010年。 [在线]。网址：https://vqeg.org/projects/hdtv [43] *ITU-T P*系列补充文件*23*语音数据库，国际电信联盟：日内瓦，1998年。