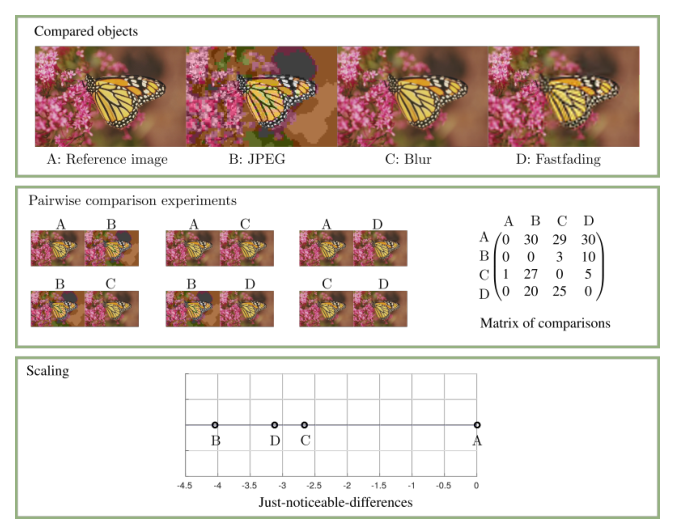
构建一维相对测量量表，表示对一组对象或条件的顺序偏好或判断，直接通过评分，可以是利克特量表或基数量表的形式，也可以是成对或成套的比较判断，因为实验过程简单，这里选择成对比较。

成对比较数据转换成测量尺度的过程：

我们通过引入离群值分析、提供计算置信区间和统计检验的方法，以及引入先验知识来改进现有的标度方法，从而在观察者数量较少时减少估计误差。我们的大多数示例都侧重于图像质量评估。

介绍：

图像质量排序中最简单的方法是两两比较，一次只显示两个条件，要求参与者根据特定标准选择其中一个条件。例如，如果我们想分析三种渲染方法（A、B和C）中哪一种产生最高质量的结果，我们可以将这些方法产生的图像成对呈现（AB、BC、AC），然后询问观察者每对中哪个图像的质量更好。如果收集到足够的数据，我们就可以将算法从最好的到最差的排序，估计这种排序的可信度，并对排序分数进行缩放，这样就可以很容易地根据更好的感知质量的概率来解释它们。



投影在一条线上表示偏好判断，直观看出各组之间的信息，条件之间的距离可以解释为更好感知质量的概率。

通常，数据分析仅限于统计测试：表明观察到的差异不太可能是偶然产生的。但统计意义可能无法转化为实际意义。

**1. 直接评级和成对比较的区别**

直接评级需要仔细训练，为给定的实验建立一个定义好的缩放，即使精细训练的情况下，在不同天数的实验中参与者之间的缩放仍可能会有较大差异。当比较条件差异过大时直接评级实验难以进行，例如，实时图像质量数据集是在7个不同的实验环节中收集的，每个环节只涉及一种类型的失真（例如JPEG压缩、噪声、模糊等）。隔离每种畸变类型简化了实验任务，但它使每一次实验中获得的质量等级彼此不同。为了校准所有的量表，作者必须进行8次重新校准实验，其中7次实验中的一部分图像被再次评估，收集的分数被用来线性地重新校准之前收集的分数。这一相当复杂的过程表明了在评级实验中获得统一质量标准的挑战。

而成对比较有诸多优点：1.这是一个非常简单的实验任务，因此非常适合非专家参与者；2.它避免了基数测量中经常遇到的校准问题；3.与直接评级相比，它通常具有更高的灵敏度和更低的测量误差；4.与直接缩放相比，它可以更快地运行（尤其是因为对参与者进行成对比较更容易、更快，而且使用自适应程序可以减少比较的数量）。

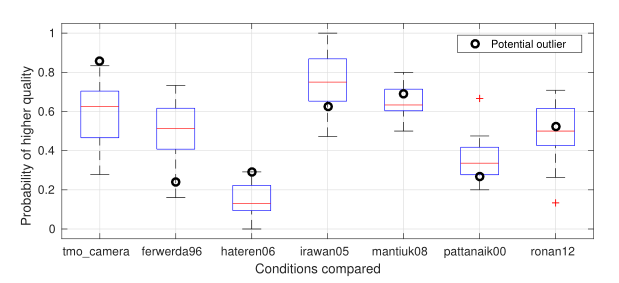
**2. 投票数和缩放的区别**

成对比较最简单的呈现方法是计算投票计数，投票计数是以序数表示结果，会产生正确的条件排名，但不能正确地捕捉条件之间差异的大小，另一方面，成对比较比例将这些条件放在一个连续的区间比例上，该比例可以捕获条件的顺序和差异的大小。有学者证实了当按比例缩放成对比较数据时，质量等级更容易被捕获。此外，当并非所有条件都相互比较（不完全设计）或并非所有观察者都比较相同条件（不平衡设计）时，计票是困难的。定标方法可以很好地处理此类非标准实验设计。

**3. 成对比较实验的例子**

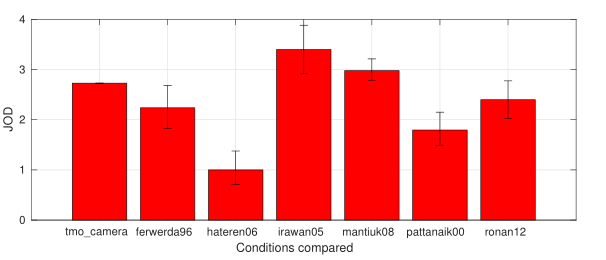
第一步是将表格中的答案转换成一组比较矩阵M，每个观测值对应一个矩阵。在这样的矩阵中，列和行对应于比较的条件，矩阵值cij=n意味着条件Oi是条件Oj的n倍。如果存在一个参考条件，例如非失真图像，则应将其作为第一行和第一列中的第一个条件放入矩阵中。第一个条件将被指定一个固定的质量值0。

第二步是执行异常值分析，以检测与其他人表现非常不同的潜在观测值。执行该分析的函数是[L，L\_dist]=pw\_outlier\_analysis（M），它接收一个矩阵M，其中包含每个观测值的响应，并返回观察每个观测数据的可能性L和四分位间的标准化分数L\_dist，这表示应进一步调查的观测。由于没有客观阈值可以区分高置信度的异常值，我们建议调查L\_dist得分接近或高于1.5的常规阈值的所有观测值。分析数据集中18个观测的结果表明，有一个观测的得分为2.72，这需要进一步关注。为了将指定观测（观测编号n\_obs）的答案与其他观测结果进行比较，我们使用函数compare\_probs\_observer（M，n\_obs），该函数绘制了选择一种条件的概率，如下图所示：

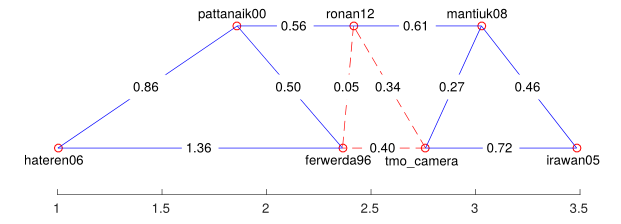


请注意，这种数据表示不涉及缩放，这可能会掩盖特定于异常值的模式。图中的黑色圆圈代表潜在异常值的答案。该图表明，潜在的异常者对操作员Ferwerda96、Hateren06和TMO\_Camera有不同的看法，但其结果与其他观测值没有太大不同。虽然观测值与其他观测值并不完全一致，但我们无法证明从数据集中删除答案是合理的。我们建议对每个观测值进行这样详细的分析，而不是使用任意措施排除观测值。异常值分析的详细信息见第9节。

一旦我们确信数据集中没有异常值，我们就可以使用[jod stats]=pw\_scale\_bootstrp（M）函数缩放结果并计算置信区间。该函数期望每个观测值M的比较矩阵与异常值分析相同，并返回缩放解决方案和一组统计数据。我们的数据集的比例和置信区间如下图所示：



置信区间表示估计质量值的置信度为95%的范围。然而，不应使用置信区间来推断差异的统计显著性。统计测试由函数pw\_plot\_ranking\_triangles（jod，stats）执行，该函数生成下图：

该图中的连续线表示这对条件之间存在统计上的显著差异，虚线表示缺乏统计显著性差异方面的证据。

**4. 实验设计**

成对比较实验的规划需要考虑几个因素，以确保以可能较小的实验工作量收集足够的数据。所需比较的数量取决于比较条件的数量n（例如，不同的算法或失真级别）、不同内容片段的数量k（例如，图像或视频剪辑）以及实验的重复次数t。如果要求每个观察者将每个条件与其他条件进行比较，他们需要执行1/2·n·（n− 1） ·k·t比较。这个数字增长很快，尤其是对于大n。

一个重要的问题是比较条件的选择，因为并非所有的比较都同样有用。产生明显结果的比较，例如比较最高和最低失真水平，对实验结果影响不大，可以避免。这些实验只对选定的一对进行比较，被称为不完全设计，而不是对每一对进行比较的完整设计。如果不是所有观察者都比较同一组条件，而是每个观察者都有不同的实验设计，那么这个实验就被称为不平衡设计。请注意，这种不平衡设计通常是不可取的，至少在不考虑观察者协变量的情况下是如此。如果我们没有关于条件潜在排序的任何先验信息，我们可以使用有效的排序算法（如快速排序）或其他专门设计的技术，如主动抽样。考虑到相同的试验次数，这会导致差异较小。然而，在许多情况下，我们提前知道最可能的条件顺序，例如，在图像压缩中，我们知道较低比特率的图像的质量比较高比特率的图像的质量差。在这种情况下，在失真水平的范围内我们可以限制比较。重要的是要确保比较图像的质量水平相对相似，以便在某些情况下混淆它们。如果所有观察者给出相同的响应，我们将无法可靠地估计他们之间的标度差。第8节和第10.1节对此进行了进一步讨论。

最后，可以在实验中给出第三个结果（即ties）。然而，这通常会使建模更加困难。我们将在第10.3节中更详细地讨论这个问题。我们的一般建议是，在没有ties的情况下进行两个替代性的强制选择实验。

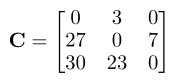
与实验设计相关的其他重要因素的讨论，如控制观看条件、降低学习效果、训练、实验疲劳，不在本报告的讨论范围内。

**5. 问题描述**

假设我们的目标是比较n个条件（例如，n幅图像，通常具有相同的内容，每幅图像都用不同的算法处理），基本真实质量分数q=（q1，…，qn），qi∈ R.该分析的目的是估计近似真实质量分数q的分数ˆq=（ˆq1，…，ˆqn）。这可以从t试验中从m个观测值收集的成对比较中获得（可能还有k份内容，每个内容单独处理）。因为成对比较是相对的，所以我们也假设q1=ˆq1=0。

**5.1 比较矩阵**

成对比较实验通常用计数矩阵C表示，其中每个元素cij测量条件Oi被选为优于条件Oj的个数（考虑到m个观测值和t个试验）。例如，在一个有三种条件的实验中，得到的矩阵可以如下所示：



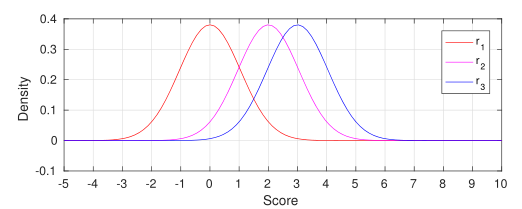
也就是说，c12=3告诉我们条件O1比条件O2好三倍，c21=27告诉我们条件O2比条件O1好27倍。可以使用矩阵C中的经验信息来估计一种条件被选择为优于另一种条件的概率（表示为Oi和Oj的pij）：



例如，O2被选为优于O1的概率可以估计为c21/(c21+c12)=27/(27+3)=0.9

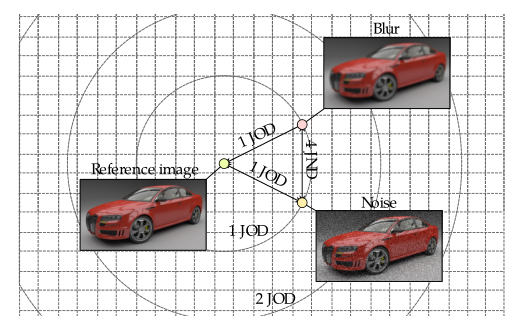
**5.2 观测模型**

在本文中，我们使用了Thurstone提出的模型。该模型假设观测值通过每个条件分配一个质量值来做出质量判断并且条件质量都是一个随机变量，从而解释这些实验的主观性。也就是说，条件Oi的感知质量被建模为一个随机变量：ri∼ N（qi，σ）（即，假设分布的平均值为真实质量分数qi）如下图所示：

观测值对质量的看法各不相同（观测值之间的差异），当他们重复相同的实验（观测值内部的差异）时，他们的观点也可能发生变化。Thurstone案例V模型假设观测值间和观测值内的方差都可以用正态分布来解释，并且该分布的方差在每种情况下都是相同的（噪声参数σ对于所有项目都是相同的，并解释了比较中的不确定性）。成对比较实验的目标是找到每种情况下分数分布的期望值ˆq。实际上，由于分数是相对的，我们对恢复分数之间的距离感兴趣。

**5.3 JNDs和JODs**

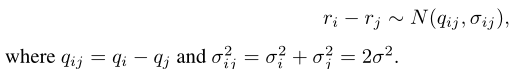
配对比较的结果通常以显著差异（JND）为单位。如果75%的观测值能表示两种刺激之间的差异，那么两种刺激之间的距离为1jnd。然而，我们认为，将测量的差异视为“明显”会导致对实验结果的错误解释。以下图所示的两个扭曲图像为例：

一个图像因噪声而扭曲，另一个因模糊而扭曲。他们绝对是明显不同的，直觉上他们应该相距超过1 JND。然而，在图像质量实验中，我们问的问题不是它们是否不同，而是哪一个更接近完美质量参考。请注意，回答这个问题不需要显示参考图像，因为我们通常对高质量图像的外观有一个概念。因此，我们收集的数据与图像之间的视觉差异无关，而是与完美质量参考的图像质量差异有关。因此，我们将这种质量度量描述为令人反感的差异（JOD），而不是JND。请注意，JOD的测量更类似于视觉等效性，或者更类似于以差异平均意见分数表示的质量，而不是JND。

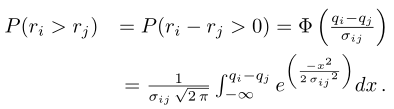
**6. 缩放方法**

成对比较可以看作是两种情况下潜在质量差异的嘈杂样本。缩放的目标是基于成对比较的噪声数据估计这些潜在差异。给定观测模型，我们可以使用以下方法之一将收集到的概率ˆpij转换为标度质量分数ˆq。

**6.1 从概率到距离**

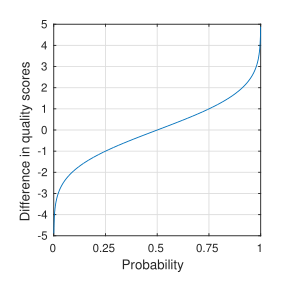
在缩放数据时，我们最感兴趣的是在基本质量分数qi和qj之间的qi-qj（因为分数是相对的）。该距离与条件Oi的质量高于条件Oj的概率有关。请注意，两个高斯数ri和rj的差也是一个高斯随机变量：

选择Oi而不是Oj的概率可以使用差值ri上的累积正态分布Φ来计算ri− rj:



从概率到分数差的映射由Φ的倒数给出（称为probit，如下图所示）：





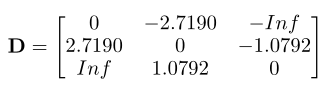
Thurstone模型假设噪声参数σ已知且在所有条件下都是常数，因此σij=σ.然而，我们不知道它的价值。一种常见的方法是选择σij，以便在随机猜测和完全确定之间的中间，将0.75的概率映射到1个JOD单位的得分距离。2个JODs的差值对应于0.91的概率，依此类推。当标准偏差σij为1.4826时，反向累积分布穿过pij=0.75的1值。

**6.2 最小二乘距离解（极大似然估计方法更好）**

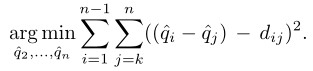
一旦我们建立了概率和得分差异之间的关系，我们就可以用公式中的估计值ˆpij代替P（ri>rj），以获得距离的估计值：



当这些概率转化为分数差时，我们得到以下距离矩阵：



我们的目标是找到一个估计ˆq，使得不同分数之间的距离与矩阵D中的距离非常相似。这种质量分数通常通过解决形式的优化问题来找到（Engeldrum，2000）：

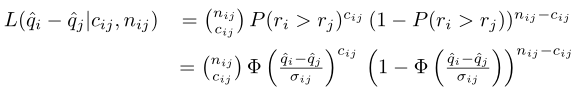


这个公式类似于当我们缩放到一个维度时的多维缩放问题，只是我们的距离是有符号的。由于不可能只考虑到它们之间的距离来优化绝对分数值，其中一个分数通常是固定的（最常见的是ˆq1=0）。

不幸的是，由于D中的无穷大值，上面式子的解在我们的例子中不可行。这两个无穷大值对应于所有观察者给出相同（一致）答案且概率等于0或1的情况。当逆累积正态分布在0和1处达到其一个渐近线时，分数中相应的距离是无限的。正无穷远或负无穷远的距离绝对是一个错误的估计，但考虑到数据，也不可能确切地说出真正的距离应该是多少。在实验中得到一致的答案是很常见的，因此设计一种处理这些情况的方法是非常重要的。有时，一致的答案会被忽略，但这会从数据中删除有效的观察结果。在其他情况下，距离范围受到限制，例如在-3和3之间，但这会在估计中引入偏差。在下一节中，我们将介绍一种更适合这些情况的优化方法。

**6.3 极大似然估计MLE（作者推荐）**

极大似然估计（MLE）为缩放提供了一种更优雅、更稳健的解决方案。MLE寻找质量分数的差异，以最大化观测数据C的概率。为此，我们需要将质量差异与比较矩阵C中收集的数据联系起来。如果我们知道选择Oi的真实概率优于Oj（P（ri>rj）），从nij=nji=cij+cji试验总数中，在cij试验中选择Oi而非Oj的概率由二项分布给出：



注意，概率P（ri>rj）取决于质量分数的差异，由累积正态分布Φ给出。

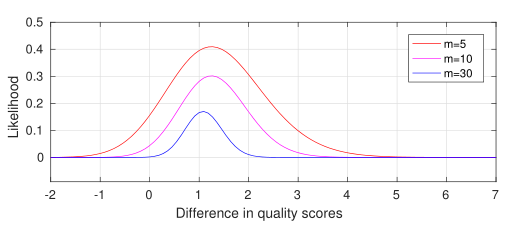
为了衡量所有比较条件，我们将所有条件对的可能性乘积最大化：



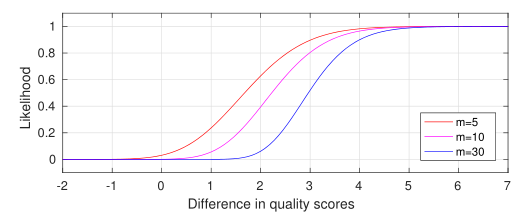
这里Ω是至少进行了一次比较的所有对的集合：cij+cji>0。注意，在实践中，极大化似然函数的对数更方便。

最小二乘距离解与MLE对比，MLE有许多优点：

1. MLE解释了比较的数量，从而衡量了我们对数据的信心。下图显示了pij=0.75的三种样本大小的可能性。样本越大，分数之间可能存在的差异范围就越窄。当实验设计不平衡时，MLE解的这一性质特别有用。



1. MLE解决方案（几乎）以一致的答案优雅地处理案例。下图将概率L绘制为质量分数差异的函数，当pij等于1且观察者人数m为5、10和30时。在每种情况下，最有可能的距离都大于5，但也有可能较小的距离，尤其是当m很小时。



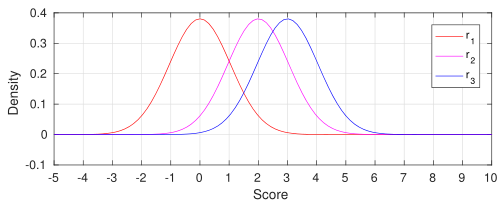
3.MLE允许我们处理不完整的实验设计，当只比较一部分对时。

**7. 统计分析**

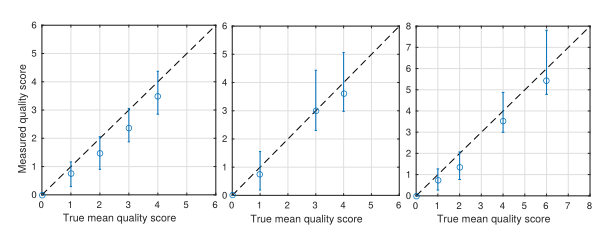
由于任何实验只给出真实质量值的估计值，因此分析和报告数据中的不确定性水平非常重要。在本节中，我们将展示如何计算置信区间并测试统计差异。

**7.1 置信区间**

由于多种条件相互影响，使用分析方法计算标度质量分数的置信区间非常困难。Thurstone案例V的原始公式不允许计算置信区间。解决此问题的代价是将模型简化，也可以使用数值方法计算置信区间，例如重采样。重采样通常用作估计采样分布的统计方法。当参数假设存在疑问时，它代表了基于参数假设的推理的稳健替代方案。一个常见的例子是使用自举技术。这种方法总是从样本中重新采样，因此依赖于从采集的样本中生成伪样本。给定一个测量样本（成对比较实验的结果），我们通过随机复制一些参与者的数据并删除其他参与者的数据来生成相同大小的新样本。这个过程被称为随机抽样和替换。为了计算置信区间，生成大量伪样本（通常超过500个），然后使用MLE方法对每个样本进行缩放，最后计算所有样本中每个条件下JOD值的2.5和97.5个百分位。这给出了平均JOD分数的95%置信区间。

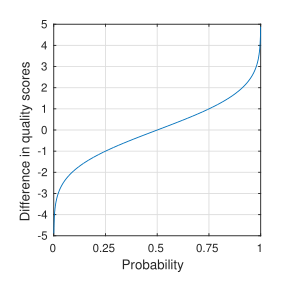


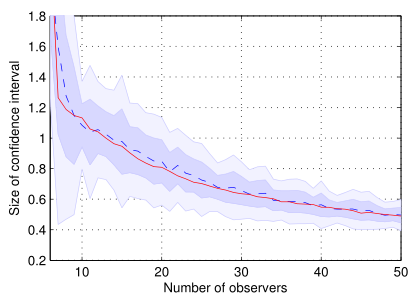
注意上下两图，上图为对三个条件进行评分的概率，下图显示了为模拟实验计算的三个置信区间示例。假设一组固定的真实分数，我们可以通过从分布中绘制模拟答案来模拟观测值判断的随机性。在我们的例子中，十名虚拟观察者（n=10）进行了三次重复（t=3）的实验，并对所有对进行了比较。从图中可以得出一些结论：



1.距离参考点0较远的质量分数的置信区间较大。由于绝对分数是根据对之间的距离估计的，因此第一个和第二个条件之间的估计误差会传播到第三个条件，依此类推。

2.随着条件之间的距离增加，置信区间变大。直观地看，probit图，较大的距离投射到较小的概率差异上。因此，当JOD距离较大时，概率估计中的小误差可能会导致估计距离中的大误差。

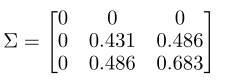


为了分析bootstrapping在我们的问题中估计置信区间的准确程度，我们在模拟中分析了它的性能，在模拟中可以高精度地估计真实的置信区间。我们假设我们知道真正的质量分数是q=（0,1,2,3,4）。然后，我们通过将一定数量的观察者的答案随机化（将随机高斯噪声N（0，1.4826）添加到q），生成相应的比较矩阵并运行我们的缩放方法，模拟了10000次实验。我们计算置信区间的平均大小（97.5%和平均值之间的距离的平均值；以及平均值和2.5%和平均值之间的距离），并将其绘制在下图（红色连续线）中不同观测数量的实验中。

然后，我们使用相同的程序模拟50个实验（对于每个观测数量），我们对这些实验运行自举，并以相同的方式计算置信区间的平均大小。引导结果的分布如上图中蓝色阴影区域和蓝色虚线所示。可以看出，平均而言，自举给了我们一个正确的估计。然而，我们需要记住，自举只是一种估计，计算的区间很容易被低估或高估，尤其是在观察者人数较少的情况下。因此，即使在置信区间内，我们也需要有有限的置信度。

**7.2 两种情况之间的统计差异**

成对比较数据的置信区间分析比典型的直接评分实验更复杂，因为计算出的JOD值不是独立的。由于所有条件都通过成对比较相互“链接”，因此更改一个条件的值将“推送”所有直接或间接链接条件的值。条件之间的这种相关性可以在协方差矩阵∑中捕捉，如下式所示：



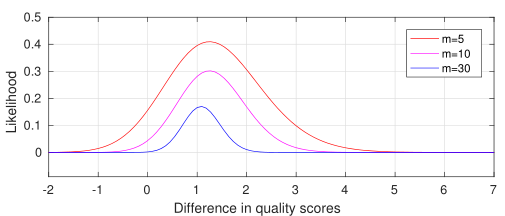
第一行和第一列都有0，因为O1总是固定为0，不能变化。值∑22=0.431和∑33=0.683表示O2和O3的方差。值∑23=∑32=0.486表示一对条件之间的差异。如果我们想要拒绝两种情况之间的JOD得分差异为0的H0，我们需要计算该差异的方差，如下所示：



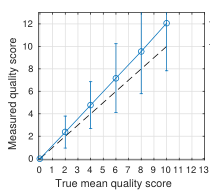
利用方差和JOD分数的差异，可以使用双尾检验来测试给定置信水平的H0。

**8. 有限距离先验**

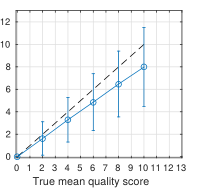
对于缩放方法来说，一致的答案是有问题的，因为它们对两个条件之间的距离没有上界，因此在估计中引入了偏差。当样本量（观察者的数量）很小时，问题最为明显。这是因为i）观察者较少时，获得一致答案的概率增加；ii）样本越小，分数之间可能存在的差异范围越大,像下图这样。

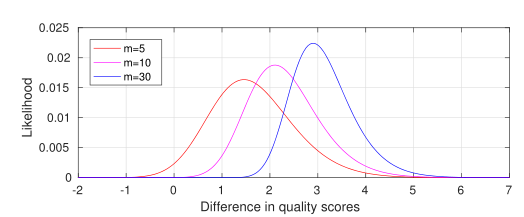


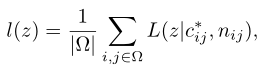
然而，通过在似然函数之前添加一个简单的距离，可以使缩放更加稳健。



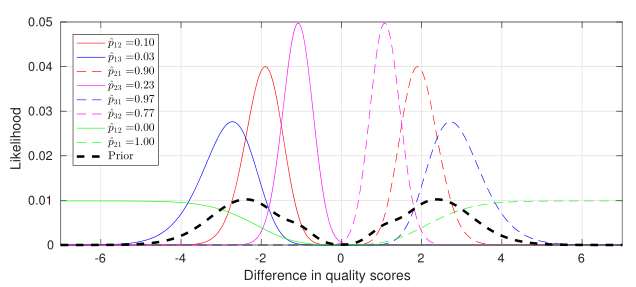
在上图中，这个问题更容易看到，其中模拟了1000次实验，以获得q=（0,2,4,6,8,10）的真实分数。如果该方法没有偏差，平均而言，实验应该给出正确的答案，正好在上图左侧的黑色虚线上。然而，由于存在偏差，测量的分数大于真实分数。原因是一致答案的情况下，条件之间的距离没有上限。



上图中所示，可以看出这会导致低估JOD值。尽管图中的似然函数允许条件之间的距离为无穷大，但我们知道，在实践中，所有距离都是有限的，通常是中等数值。这种关于有限距离的知识将是我们的首要任务。我们可以将随机选择的两种情况下，在质量分数中观察到特定距离的可能性定义为先验知识。给定条件对的这种可能性用极大似然估计表示。根据一开始的比较矩阵，我们在下图绘制了所有条件对的可能性。观察到任何差异的概率是所有绘制概率的归一化总和。然而，问题是，一致答案的可能性（图中的绿色线为ˆp12和ˆp21）有无限的支持度，因此无法标准化。为了避免这个问题，我们将这个答案转换为最接近的非一致回答。在这一步之后，我们可以计算观察任意两个随机条件之间距离z的概率，如下所示：

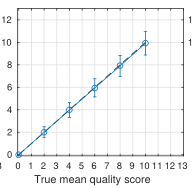


这里的c\*ij是cij最接近的非匿名版本。求和中的主项由极大似然估计给出，我们的先验值取决于优化方法当前迭代中的估计距离和迭代方式的变化。该概率如下图中的黑色虚线所示。结果表明，两个随机选择的条件之间的最大概率差约为2.5个JODS，并且该概率函数的支持度是有限的。我们可以在似然函数之前加上距离：

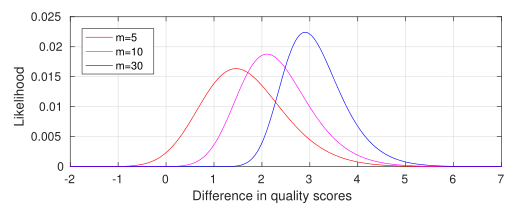
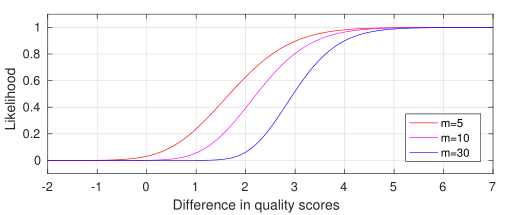




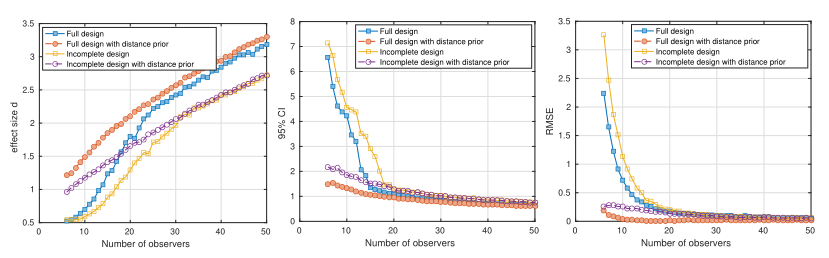
请注意，这只是先前的调制距离，而不是约束。为了允许选择其他距离，我们在先验距离的基础上加上一个小的偏移量γ=0.1。



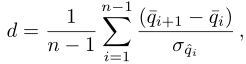
上图显示了在似然函数中包含先验信息时，如何减少偏差。下左图显示了下右图中的似然函数，当它与先验知识相乘时。可能性不再平稳，只有一个最大值，这也提高了优化的稳定性。

为了评估先验估计带来的改进，我们分析了估计的精度如何随观察者的数量而变化。我们对真实质量分数q=（0,1,2,3,4）进行了10000次蒙特卡罗模拟，并采用了与第7.1节中估计置信区间相同的假设。我们对完整设计（比较所有条件）和不完整设计（只比较最近邻）进行了模拟。对于每个模拟，我们获得一组10000个估计质量分数ˆq，我们的目标是将其与q中的真实质量分数进行比较。我们将qi估计的平均值定义为“q-i”。下图显示了三种不同测量的结果：



1. 效应大小d：估计质量分数之差除以估计误差标准差的比率：



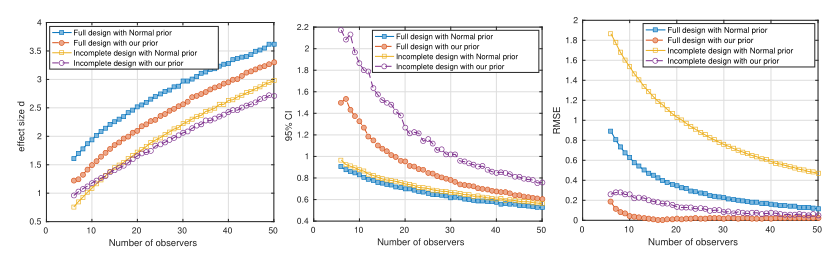
式中，∑ˆqi是每个个体估计结果与分布q-平均值的标准偏差。效应大小是实验方法灵敏度的一个有用度量，用于计算它是否能够检测到一对条件之间的差异，并证明其统计意义。

1. 95%置信区间的平均大小，按第7.1节计算。
2. 均方根误差（RMSE），用于测量与地面真相的偏差，如下所示：



图中还显示了随着观测数量的增加，这些措施是如何改进的。它还表明，如果观测的数量少于20，RMSE和置信区间都可能非常大。所提出的距离先验显著提高了估计的准确性和鲁棒性，特别是对于小样本。

为了证明选择正确的先验知识的挑战，我们使用我们的距离来比较标度，然后再使用中提出的先验知识。作者引入了一个先验知识，假设质量分数来自正态分布。如下图所示，尽管他们的先验极大地降低了置信区间（就像大多数先验一样），但它也在估计中引入了很大的误差（较大的RMSE）。



**9. 异常检测**

在实际中，一些观测可能无法完全理解或遵循实验的指示，尤其是在控制较少的众包实验中。重要的是要检测出这些不属于整体模式的观测值，因为他们的答案可能会将缩放推向错误的解决方案。本节介绍了一种新的方法来检测这些异常观测值。请注意，这种方法仅用于支持实验者，实验者最终决定是否应将观测值视为异常值并从数据集中删除。

为了表明某个特定的观测值是否可以被视为异常值，我们将其答案与样本的其他部分进行比较。首先，我们从数据集中排除一个给定的观测值（一个接一个），并使用最大似然估计（MLE）方法来寻找标度距离，从而得到概率P（ri>rj）。鉴于其余样本的这些概率，我们使用似然度的乘积来计算观察所考虑的观测值结果的概率。如果所考虑的观测值与人群中的其他人一致，则相应的概率将很高。在实践中，我们使用对数概率之和，因为它不仅简化了后续分析，而且在数值上也有帮助，因为大量小概率的乘积很容易低于浮点数的数值精度。

可以使用不同的规则来检测异常值，其中大多数规则都考虑了到分布中心度量的距离和数据范围。在我们的例子中，我们考虑J. Tukey关于四分位数的规则。我们将对数似然度转换为分数，分数以四分位区间的倍数表示到分布中心范围的距离。四分位范围是75%和25%之间的距离。我们只考虑分布左侧的异常值，即显示属于样本的可能性很低的情况，计算到第一个四分位数的距离。

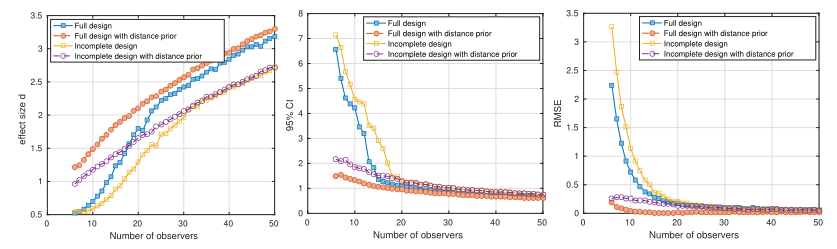
我们进行了一系列蒙特卡罗模拟，以确定异常值的存在如何影响缩放结果，以及我们的方法是否可以用于自动确定异常值。正如预期的那样，当有效观测值（非离群值）的数量很小时，离群值可能会引入最大的误差。但更有趣的是，我们观察到，当重复次数t较小时，异常值更难检测。因此，我们建议每个观测值至少重复相同的比较3次。在调查之前论文中的实际（非模拟）数据集时，我们发现，鉴于实验的主观性，这种检测异常值的自动标准可能并不总是准确的。因此，我们建议将此决定留给实验者，实验者应调查异常值分数较高的标记观测值的结果（如第3节所述）。

**10. 实际问题**

在这一部分中，我们探讨了关于成对比较实验的三个相关问题：完整和不完整设计之间的比较、质量分数之间的距离以及实验中的关系容差。

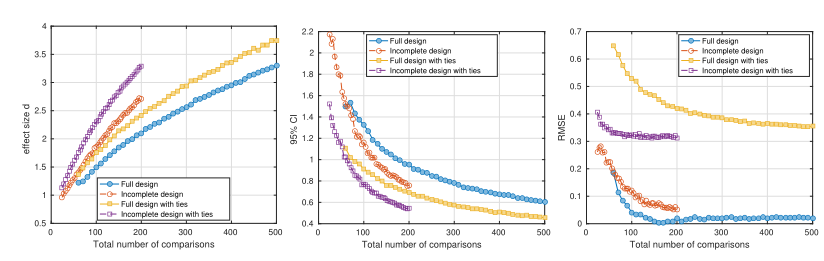
**10.1 完整或不完整的设计**

在设计实验时，我们可以选择比较所有可能的条件对（完全或完全设计），或者只选择最相似的条件对（不完全设计）。我们感兴趣的是知道哪种方法更有效，并导致更准确的结果。



在讨论先验知识对于小样本量的重要性时，我们已经在上面三个结果图中展示了完整和不完整结果的一些结果。在该图中，不完整设计的精度相似，但总体稳定性较低。然而，这些图并没有考虑到这样一个事实：在完整的设计中，每个参与者都需要进行更多的比较。在我们的模拟中，假设n=5个比较条件，完整设计需要比较（5·4）/2=10对，但在不完整设计的情况下，我们只比较4对：q1⇔q2，q2⇔q3，q3⇔q4和q4⇔q5。

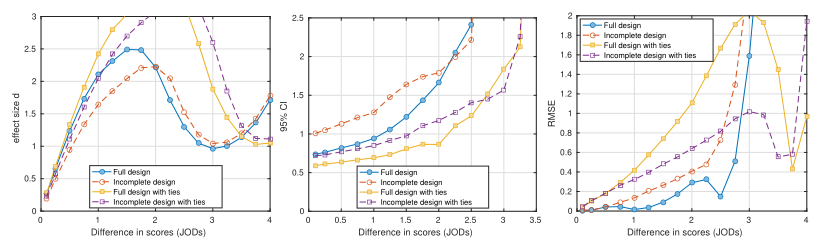
在下图中，我们将数据替换为比较次数的函数，而不是观测的函数。请暂时忽略“ties”的曲线，关注完整和不完整设计的蓝色和红色线条。图中显示，在相同的实验环境下不完整的设计会导致更稳定和更准确的估计。增益将取决于要比较的条件的数量。例如，如果我们有10个条件，完整设计将需要比较45对，但在不完整设计中只需要比较9对，从而获得更大的增益。。



第一组结果图

**10.2 质量分数之间的距离**

衡量方法的准确性取决于质量分数之间的距离。如果质量分数之间的距离大于2 JODs（即pij>0.91），则缩放变得尤其不可靠。当我们怀疑感知属性将在大于2个jod的范围内进行缩放时，差异缩放方法可能更合适。为了测试这种效果，我们对真实JOD分数之间的不同假设距离（均相等）进行了蒙特卡罗模拟，并总结在了下图。让我们专注于完整的设计（蓝色连续线），暂时忽略所有其他曲线。如中所示，即使条件之间的距离变化平稳，RMSE和置信区间也会迅速增加。很明显，由于真值范围沿x轴增加，RMSE等测量值将增加，然而，增加的幅度比预期的线性增加更突然。



第二组结果图

**10.3 ties实验**

在成对比较实验中，允许观察者在看不到差异时选择第三个“无偏好”选项是一个有争议的问题。

在统计分析中引入联系有不同的方式。在下一个实验中，我们选择了等分法：如果观察者选择“无偏好”，我们把选票一分为二，每项条件加上一票半。这可能会导致非整数投票数，我们将其四舍五入到最接近的较小或较大整数（随机选择，并考虑到需要保持一致的比较次数）。我们模拟了当两个条件之间的差异小于某个阈值时，做出“无偏好”选择的观察者。当两个条件相同时，不同的观察者不可能有相同且一致的观点，因此我们的“无偏好”阈值是JOD单位空间中的随机变量N（0.7,0.3）。使用tie模拟10000次实验运行的结果与上面两组结果图中没有tie选项的相同实验进行比较。

我们的模拟表明，提供“无偏好”选项可以减少置信区间的大小，并改善效应大小。但这是以更大的错误为代价的。仔细观察结果，我们发现解决方案总是被低估。这一结果有一个直观的解释：提供“无偏好”选项会导致更多的“无差异”反应，而差异实际上存在，给出较小的JOD距离和负偏差（低于预测）。偏差大到足以抵消减少的置信区间内的任何增益。这种偏差可能会被消除，但它需要对“无参考”选择进行建模，并找到该模型的参数：在实际没有差异的情况下，观察者选择“无偏好”的可能性有多大。这反过来又需要收集额外的数据：观察者对两种相同条件的反应。当前版本的pwcmp软件不支持缩放时的建模关系，因此，当此软件用于缩放时，我们不建议提供“无偏好”选项。