评判的一致性与正确率

摘要

本文通过运用统计方法，对评价者们的评价的正确率，一致性和评价倾向（手松手紧）进行了理论推导并给出了可行性分析。对于正确率的评价，具有较广的试用范围，可以进行转写正确率的估计，正确率的假设检验，置信区间估计.该方案只要在评价者的测试次数大于10次即可行，在30次以上有较高的准确性。对于一致性的评价，本文针对两类情况给出方法。当评价结果呈现多种情况且可以量化（连续型的评价结果），可使用基于空间中距离的聚类方法，选出一致性较高的类别。当评价结果近似分类变量（例如 对 or 错），则可以使用卡方检验一致性，在排除一致性较低的个体后，使用kappa系数度量剩余个体间的一致性情况。而评价的倾向性则试用范围较窄，在衡量评价者是否过高估计被评价人的普通话水平或者过低估计时，可以使用假设检验的方法甄别出有过高或过低估计倾向的评价者。

问题陈述

现实生活中，对于某一指标进行评价（例如普通话的标准程度），常常需要选取一定数量的人对该指标进行评价。

本文的目标在于要从理性科学的角度对评价人的正确率一致性进行评价，从而选出具有较高正确率并且找出这些评价人中一致性较好的群体。从而确定优质评价人选。

需要回答的问题：

1. 谁的评价正确率较高？是否和其他人有显著的正确率差异？
2. 哪些人的评价具有较高的一致性？
3. 评价者是否有过高评价或者过低评价的倾向？

数学模型

1. **数值处理**

为通过数值指标来衡量一个评价质量的好坏，需要对结果进行数值化处理，以方便计算。

数值化的方法有很多，对普通话的数值化处理，常以线性的方式处理：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 普通话等级 | 一甲 | 一乙 | 二甲 | 二乙 | 三甲 | 三乙 |
| 分值 | 5 | 4 | 3 | 2 | 1 | 0 |

此时可称被评判人的真实成绩为向量

1. **正确率**

每一次评价的结果要么正确要么错误，直接的正确率计算为一个人评价的正确次数比上总次数.

此时为正确个数，为评价次数。

1. **正确率差异的假设检验**

经过测量后，会得到多个测试人的正确率：. 此时我们会选取正确率最高的个体的正确率, 接下来需要检验 是否与具有显著性差异。相当于检验m人的正确率是否显著高于于其他人的正确率。

假设检验的方法：

双样本t-检验. 检验第i号人的正确率.是否显著性低于m号人的正确率. 原假设与备择假设如下：

假设检验统计量及p值计算：

如果p值大于显著性水平 (例如0.1), 我们不能有十足的把握认为两人的正确率有显著性差异。反之，我们认为具有显著性的差异，也就是认为第i号人的正确率是明显小于第m号人的正确率的.

不妨假设前k个人（集合A）与第m个人是没有显著的正确率差异的，这k个人的正确率如下：

此时我们认为这k个人具有相近且较好的正确率。

可行性试验：显著性水平0.1

评价五个人，正确率0.8 与0.6没有显著性差异

评价十个人，正确率0.8 与0.6有显著性差异

评价二十个人，正确率0.8 与0.65有显著性差异

评价三十个人，正确率0.8与0.7有显著性差异

可行性：该正确率假设检验要能达到筛选和评判人的地步，需要十人以上，最好能达到二十人以上。

1. **正确性**

在正确率相接近的情况下，我们希望一个人的对普通话评价的结果，应该和真实结果相接近，从而有更高的正确性。为了容易理解，可以参照下述例子：

甲，乙两人分别为A，B两个人的普通话水平打分。可见下表：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 甲 | 乙 | 真实分数 |
| A | 5分 | 4分 | 3分 |
| B | 0分 | 1分 | 2分 |

在上述例子中可以看出，甲乙两人的评价正确率均为零，但是乙的评价显然更好，因为她的评价与真实的情况更加接近。我们把这个所谓的和真实值的接近程度，称为正确性。

正确性的数学度量：

如果一个人（ZM）的评价一致性程度很好，那么与这两个向量的距离应该很近，也就是：

应该较小。（上述距离选取为欧式距离）

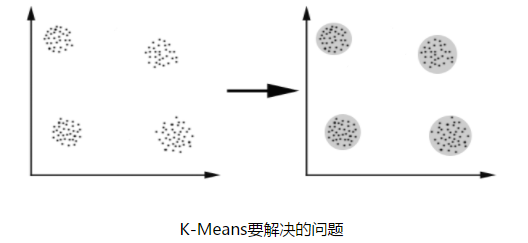
正确性衡量指标：

该值越小，正确性越高。

1. **一致性（连续型数据）**

一致性选择的目的是在于发现评价人之间是否有近似的评价或是相似的看法。因此需要通过统计方法选出具有较大相似性的组，然后择有较高正确性的那一组。

使用的方法类似聚类方法，但我们只需要考虑正确率较高的类别，而不在乎正确性较差的是否能分到同一类中。但聚类的方法可以类似于K-means：



方法过程：

（1）从 n个数据对象选择正确率最高的任意对象作为第一个初始聚类中心；

（2）选取与正确率最高对象最近的对象，加入类中。

（3）根据聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；计算类中个体与中心值的正确性指标

（4）循环（3）直到超过一定数值，例如0.3；或者当类中的个体数超过

此时我们选出类中个体，即为一致性较高且有较高正确性的个体。

可行性：只要有了每个人评价人的评价结果，就能做相应的分析，可行。

1. **分类型数据的一致性**

假设检验排除一致性极差的个体：

在对只有正确或错误的数据进行一致性判断时，可以使用卡方检验，因为每一个被评价对象都受到多个评价人的评价，可以认为这样的评价是paired-data. 因此使用Paired Chi-Square Test.

Contingency Table

|  |  |  |  |
| --- | --- | --- | --- |
|  | 评价者1：0 | 评价者1：1 | Total |
| 评价者2：0 |  |  |  |
| 评价者2：1 |  |  |  |
| Total |  |  | N |

利用卡方检验统计量，计算p值，在显著性水平下拒绝或接受原假设。

p值计算方法：

利用kappa 值判断剩余个体间的一致性程度：

运用Kappa值

Kappa值判断标准：

Kappa≥0.75，说明两种方法诊断结果一致性较好；

0.4≤Kappa<0.75，说明两种方法诊断结果一致性一般；

Kappa<0.4，说明两种方法诊断结果一致性较差。

1. **评价者的倾向**

这个部分与上面的内容属于平行的关系。目的观测评价者的评价是否存在把人过好评价或是过坏评价的趋势。因此，每个评价者的结果可分为三类：{评价正确}，{评价过高}和{评价过低}。因此上述三个时间的发生概率模型如下：

相关假设检验：

可行性试验：显著性水平0.1

10次评价，正确率，正评价与负评价相差两次，则评价倾向显著。

10次评价，正确率，正评价与负评价相差一次，则评价倾向不显著。

50次评价，正确率，正评价与负评价相差五次，则有较为显著的评价倾向。

10次评价，正确率，正评价与负评价相差一次，则评价倾向不显著。

10次评价，正确率，正评价与负评价相差两次，则评价倾向显著。

10次评价，正确率，正评价与负评价相差两次，则评价倾向不显著。

10次评价，正确率，正评价与负评价相差四次，则评价倾向显著。

项目可行性结果

首先需要有易于量化的评价指标，例如普通话的等级。这样才能有效数值化处理。

按照正确率的高低，通过假设检验，筛去正确率显著低于最高正确率的人。该过程需要每个评价人有十次以上评价较好，二十次评价为优。

一致性评价过程通过距离总能选出与正确率最高个体有较好一致性的个体们。该过程对数据数量的要求没给出具体要求，因为无论数据的多少，可以进行相应的距离计算和一致性分类。但在评价人较少时（少于3），无法确定一致性。就像两个人的话无法说明谁的一致性更好，而评价人越多时（大于5），一致性的结果会更好且更合理。

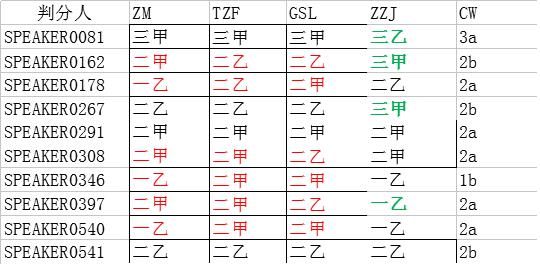
对评价倾向的测试，大约十次的评价数据，正负评价相差占总评价次数的百分之二十时，能有较为显著评价倾向。大致需要三十次的评价，正负评价相差占总评价次数百分之十才能有较为显著的评价倾向。

综上，在评价次数少于10时，各项评价的误差较大，不利于得出显著性的结论，或者说结论有较差的指导意义。十次以上的评价会有较好的结果。三十次的评价可以达到较为理想的评价结果。

普通话实验（一次实验尝试，仅供参考）

1. **问题描述**

如下是四个人ZM, TZF, GSL, ZZJ对十个人的普通话能力评价结果，CW是正确值，需要求解四人中正确率和一致性较好一人或是几人。



1. **数值处理**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 对象 | ZM | 评分 | TZF | 评分 | GSL | 评分 | ZZJ | 评分 | CW | 评分 |
| 1 | 三甲 | 1 | 三甲 | 1 | 三甲 | 1 | 三乙 | 0 | 三甲 | 1 |
| 2 | 二甲 | 3 | 二乙 | 2 | 二乙 | 2 | 三甲 | 1 | 二乙 | 2 |
| 3 | 一乙 | 4 | 二乙 | 2 | 二甲 | 3 | 二乙 | 2 | 二甲 | 3 |
| 4 | 二乙 | 2 | 二乙 | 2 | 二乙 | 2 | 三甲 | 1 | 二乙 | 2 |
| 5 | 二甲 | 3 | 二甲 | 3 | 二甲 | 3 | 二甲 | 3 | 二甲 | 3 |
| 6 | 二甲 | 3 | 二甲 | 3 | 二乙 | 2 | 二甲 | 3 | 二甲 | 3 |
| 7 | 一乙 | 4 | 二甲 | 3 | 二甲 | 3 | 一乙 | 4 | 一乙 | 4 |
| 8 | 二甲 | 3 | 二甲 | 3 | 二乙 | 2 | 一乙 | 4 | 二甲 | 3 |
| 9 | 一乙 | 4 | 二甲 | 3 | 二甲 | 3 | 一乙 | 4 | 二甲 | 3 |
| 10 | 二乙 | 2 | 二乙 | 2 | 二乙 | 2 | 二乙 | 2 | 二乙 | 2 |

1. **正确率计算结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 人 | ZM | TZF | GSL | ZZJ |
| 正确率 | 0.7 | 0.8 | 0.7 | 0.4 |

1. **正确率假设检验结果**

将ZM, GSL, ZZJ 分别与正确率最高的TZF做假设检验，检验p值和结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 人 | ZM | GSL | ZZJ |
| 正确率 | 0.7 | 0.7 | 0.4 |
| p-值 | 0.3016 | 0.3016 | 0.0228 |
| 显著性水平0.1下结论 | 与TZF无显著性差异 | 与TZF无显著性差异 | 显著低于TZF |

因此该四人中，正确率较高的为前三人.

1. **一致性**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 人 | ZM | TZF | GSL | ZZJ | CW |
| 进入类顺序 | 4 | 2 | 3 | 5 | 1 |

1. https://baike.baidu.com/item/K-means/4934806?fr=aladdin
2. https://baike.baidu.com/item/K-means/4934806?fr=aladdin