双样本精确置换检验算法的一个优化

0. 引言

置换检验（Permutation test, PT）的思想可以追溯到上世纪30年代Fisher等人的工作。Fisher（1935）在他的著作《实验设计》中，提及到了一种叫做“随机化检验（Randomization test）”的统计方法，该方法系置换检验的雏形。Pitman（1937）进一步发展了该方法，并将其引至双样本假设检验等领域。随后，Lehmann和Stein（1949）证明了置换检验是在Neyman-Pearson情形下的最优检验方法。除此之外，作为非参数统计中的重要方法，置换检验具有并不需要假定总体分布，对离群值不敏感等优点。但由于其计算过于繁琐，计算量巨大，因此置换检验至今仍然没有成为假设检验中的主力军。本文提出一种对双样本情况下精确置换检验（Exact permutation test, EPT）算法的优化，并将其与传统EPT方法，随机置换检验（Randomized permutation test, RPT）方法、经典参数方法以及经典非参数方法进行对比。

1. 置换检验

置换检验的基本思想与传统假设检验方法基本一致，即假设样本来自于同一总体，那么给定的统计量在一次实验中不太可能处于理论抽样分布的极端位置。在置换检验中，理论抽样分布通过按一定规模对全部样本进行不同组合并计算对应的统计量形成。根据形成理论抽样分布是否为精确分布可以将置换检验分为精确置换检验（EPT）和随机置换检验（RPT）两大类。其中EPT要求不重不漏地计算出给定样本所有可能的组合并逐一计算统计量，进而得到理论抽样分布，RPT则通过反复的对样本进行随机抽样，并对每个随机抽样获得的样本进行计算得到理论抽样分布的近似分布。

举一个例子，设目前手头有两个样本X，Y，样本量分别为，我们需要检验这两个样本是否为同一总体中抽样获得。那么，在计算理论抽样分布时，EPT要求我们首先混合两个样本，形成一个样本量为20的样本Z，而后对混合样本Z按照与原有样本的样本量保持一致的原则进行划分，易知本例中的划分方法共有种，进一步对每一种划分情况都计算对应的统计量，最终形成精确的理论抽样分布。而RPT方法则是对混合样本Z按照给定的样本量大小不断进行k次随机抽样，对每一次抽样结果的统计量进行计算，最终形成近似的理论抽样分布。

2. 对EPT算法的优化

从前面的介绍中我们容易发现，传统的EPT方法要求我们不重不漏地计算所有的混合样本划分情形以形成理论抽样分布，而后借助理论抽样分布计算置信度。这样做不仅需要在计算理论抽样分布时对全部的组合可能进行计算，而且还需要对各个组合的结果进行比较大小。这样，对于两组样本量均为n的样本来说，传统的EPT方法需要完成次加法运算，和次比较运算才能得到结果，其中还不包括计算分组结果的计算量。这使得传统的EPT方法在样本量不是很大的情况下就已经需要消耗大量的时间进行计算，直接限制了EPT方法的使用。因此，我们对传统EPT方法进行了一定的改进，在使用平均数构建统计量的情况下减少了大量不必要的计算。

在说明该算法之前，我们需要先说明若干引理。

引理1：在双样本精确置换检验中，选择样本均数差作为检验统计量等价于选择其中任一样本和作为检验统计量。

注意到针对给定的两样本与，由于EPT的理论抽样分布是通过对该二样本混合形成的样本进行重新划分得到的，因此针对所有的可能组合而言，样本的观测值总和是一定的，我们不妨设两样本的观测值的总和为S，样本1观测值的和为S1，样本量为n1，样本2的样本量为n2。显然，样本2观测值的和为。因此，我们有：

即S1与之间存在线性关系。显然，此时有：

即对进行检验等价于对进行检验。

引理2：在双样本精确置换检验中，当我们选择样本均数差作为检验统计量时，任何双边检验均可以转化为一个左边检验和一个右边检验。

注意到假设检验的实质即为针对某一个特定的统计量，计算出零假设成立的前提下出现实验结果或比其更极端结果的条件概率（即）。因此，在进行样本均数差的双边检验时，我们需要计算的实际上是：

考虑到，显然我们有：

即对进行双边检验等价于对其进行一次左边检验和一次右边检验。

综合引理1与引理2，容易有：

即分别对应右边检验，左边检验，双边检验所检验的统计量及其临界条件。从该结论可以看出，计算EPT检验的核心步骤是计算样本1与样本2的混合样本的全部n1组合中满足或者的组合的个数。

引理3：对项数为n的任意单调不减的有限数列的两个长度为r的子数列和而言，如果满足对均有，那么一定可以断言；对项数为n的任意单调不增的有限数列的两个长度为r的子数列和而言，若满足上述条件，那么一定可以断言。其中表示数列的第i项在原数列中的位置，表示数列的第i项在原数列中的位置。

这是显然的，注意到对单调不减数列我们有，因此，对任意的，均有，由同向不等式的可加性可知引理3成立。

此外，参考按字典序生成组合的算法，我们对从1开始，长度为n的正整数集按字典序的下一个r组合进行定义（以下简称字典序）：令全体集合中的一个r集合为，该集合按字典序的下一个r集合则为，其中。容易知道，从集合开始，按上述顺序可以不重不漏地给出S集合中的全体r集合，而针对长度为n的任意组合，可以将其中的元素编号，上述算法依然适用。

结合上述定义和引理3，我们容易得到下述两个结论：

结论1：按字典序生成单调不减的有限数列的长度为r的子数列时，形如的子数列是所有含的子数列中和最小的子数列。按字典序生成单调不增的有限数列的长度为r的子数列时，形如的子数列是所有含的子数列中和最大的子数列。

按照引理3，需要说明上述结论成立，我们只需要说明在字典序中，形如的序列的每一项都不大于在字典序中排在其后的与其等长的子序列的对应项。实际上这也是显然的，注意到在字典序中，下一个序列至少有一项要大于原序列，并且小于该项的所有项均与原序列相等。进一步地，我们考虑满足形如的初始序列。按照前述规律，我们不妨设第k项是满足前述条件的项并令下一个序列为，于是我们有。考虑到，中所有的项均为正整数，进一步有，重复前述步骤，即可得知前述结论的正确性。

结论2：给定最大项为的单调不减有限数列的长度为r的子数列中，形如的子数列是其中和最大的。给定最小项为的单调不增有限数列的长度为r的子数列中，形如的子数列是其中和最小的。

考虑单调不减有限数列的定义（），结论2显然成立。

基于前述结论，我们可以对传统的EPT算法进行改进。改进的EPT算法如下：

1. 建立假设，确定检验水平。

2. 选择样本量较小的样本的全部样本值的和作为统计量，并计算出现有样本统计量。

3. 选择合适顺序将全部样本进行排序并标号（假定样本量较小的样本均值为μ1，若则采用逆序，若则采用顺序，则进行两次排序，并计算逆序情况下对应的，其中分别为两样本的样本量和两样本全部样本值的总和），并置参量，，置下标数组。

4. 按照下标数组给出的序号取出排序混合样本中对应的元素并计算其样本值的代数和T，若T满足给定条件（若则为，若则为，则混合样本顺序时为，混合样本逆序时为），则令，反之则令。

5. 计算下一个下标数组，具体而言：若第4步中满足给定条件，则找到使得成立的最大整数j；反之，若第4步中不满足给定条件，则找到使得成立的最大整数j，下一个下标数组即为。

6. 重复第4步和第5步，直到下标数组为止。

7. 根据公式计算出最终的概率（若则需要顺序逆序各计算一次并取两次的i值相加得到对应的i值）。

本算法通过对数据进行排序处理，在一定程度上利用数据的位置信息替代了其大小信息，从而实现了在生成组合序列的步骤上挑选出肯定不需要检验的组合，进而减少了传统EPT算法的计算量。

3. 数据实验

3.1实验1：不同信噪比下给出置信度的比较

这部分旨在说明改良的EPT算法并没有降低原始算法的精度，该部分比较了四种常见的双样本平均值检验的算法：EPT、RPT、独立样本T检验和改良的EPT方法。所有计算在R语言环境下完成（R版本3.2.4），测试平台相关信息见表。

|  |  |
| --- | --- |
| 项目 | 型号/参数 |
| CPU | Intel Core i5-2410M 2.30GHz |
| 内存 | 4G可用 |
| 操作系统 | Windows 10 专业版 |

测试通过R语言自带的随机数发生器生成待测样本，固定随机数种子为12345，分别生成信噪比为-4dB至10dB的正态分布随机样本、对数正态分布与均匀分布随机样本并进行左边检验，结果如下：

（dB是表示功率量之比的一种单位，等于功率强度之比的常用对数的10倍）

3.2实验2：不同信噪比下计算时间的比较

3.3实验3：不同样本容量下计算时间的比较

4. 分析与结论