·论著·方法学研究·

极小样本两独立定量资料假设检验方法比较

郭轶斌[1]，李佳迅[2]，吴 骋[1]，郭 威[1]，何 倩[1]

1. 海军军医大学卫生勤务学系军队卫生统计学教研室（上海 200433）
2. 海军军医大学基础医学院（上海 200433）

【摘要】目的 探索极小样本两独立定量资料假设检验方法的表现性能。**方法** 使用蒙特卡洛方法产生不同均数差、分布和样本量的数据，分别使用 *t* 检验、Wilcoxon 秩和检验和 Bootstrap 法进行假设检验，并估计每种情形下的统计效率。**结果** 当样本量极小时，Wilcoxon 秩和检验的统计效率极低。当数据呈偏态分布时，Bootstrap 置信区间法容易犯 II 类错误。当均数差较大时，该法仍有较高的统计效率。不论数据是否服从正态分布，当样本量极小时，*t* 检验的表现优于 Wilcoxon 秩和检验。**结论** 根据本模拟研究结果，当数据服从正态分布时，建议使用 *t* 检验对极小样本进行统计推断。当数据不服从正态分布时，建议使用 Bootstrap 置信区间法对极小样本进行统计推断。

【关键词】极小样本；定量资料；数据模拟；假设检验；非参数检验

The comparison of hypothesis testing methods for two independent quantitative data with extremely small samples

Yi-Bin GUO[1], Jia-Xun LI[2], Cheng WU[1], Wei GUO[1], Qian HE[1]

1. Department of Military Health Statistics, Naval Medical University, Shanghai 200433, China
2. School of Basic Medicine, Naval Medical University, Shanghai 200433, China
Corresponding author: Yi-Bin GUO, Email: guoyibin@smmu.edu.cn

【Abstract】Objective To explore the performance of hypothesis testing methods for two independent quantitative data hypothesis tests with extremely small samples. **Methods** Monte Carlo method was used to generate data with different mean difference, distribution and sample size. *T*-test, Wilcoxon rank sum test and Bootstrap method were used to test the hypothesis, and the statistical efficiency was estimated in different scenarios. **Results** When the sample size was extremely small, the statistical efficiency of Wilcoxon rank sum test was very low. The Bootstrap confidence interval method was prone to make type II errors when the data were skew distributed. When the mean difference was large, the method still had high statistical efficiency. Whether the data followed normal distribution or not, when the sample size was extremely small, *t*-test performed better than Wilcoxon rank sum test. **Conclusion** According to the results of this simulation study, when the data follow the normal distribution, it is suggested to use *t*-test to analyze the extremely small samples. When the data does not follow the normal distribution, it is suggested to use the Bootstrap confidence interval method to analyze the extremely small samples.

DOI: 10.12173/j.issn.1004-4337.202302073
基金项目：2021 年海军军医大学校级课题（2021QN15）
通信作者：郭轶斌，博士，Email：guoyibin@smmu.edu.cn

【Keywords】Extremely small sample; Quantitative data; Data simulation; Hypothesis test; Non-parametric test

在基础医学实验研究中，研究对象以细胞、动物为主，一些实验细胞或动物模型不仅构造困难，而且花费较大，如巴马小型猪或恒河猴等，不仅动物本身费用较高，同时因伦理限制无法纳入太多。因此，部分动物实验的样本量极小，如每组小于 10 例 [1–3]。统计学上为了保证一定的统计检验效率，常要求样本例数不能过小。此外，在使用如独立样本 *t* 检验等参数检验方法时，还要求样本服从正态分布和方差齐性的假设 [4]。但在极小样本的情况下，即使是不服从正态分布的样本，在统计检验效率很低的情况下也无法拒绝 H_0 假设（样本服从正态分布或满足方差齐性）。当独立定量资料样本不满足正态分布或方差齐性假设时，可以使用对数据分布不敏感的非参数检验，对于两组独立定量资料，可以使用 Wilcoxon 秩和检验或 Mann–Whitney *U* 检验来比较两个样本所代表的总体分布位置是否相同 [5–6]。但这两种方法是将样本的原始数据编秩后再进行后续的假设检验，当资料服从参数检验的条件时，会导致样本大量变异的信息损失，进而影响统计检验效率，增加犯 II 类错误的概率 [7]。当样本量小于 4 时，使用 Wilcoxon 秩和检验的 *P* 值均大于 0.05。Siegel 认为样本量小于 6 时，不能使用 *t* 检验 [8]。祝国强等认为在对非正态极小样本的定量资料进行统计推断时，不适合使用 *t* 检验，推荐使用 Wilcoxon 秩和检验 [9]。林正大等认为在大样本或偏离对称性较远的情况下，Wilcoxon 秩和检验更优 [10]。对于统计学的频率学派来说，假设检验和置信区间（Confidence Interval, CI）是一对相伴相随的概念，在同一置信度 / 检验水准下，参数的置信区间未跨过拒绝域，假设检验则不能拒绝 H_0。Bootstrap 法是一种可以用来稳健地估计置信区间的非参数方法，其通过对原始样本数据进行有放回抽样得到统计量的经验分布，从而估计统计量对应总体参数的置信区间 [11]。在极小样本时，Bootstrap 法能否达到其在大样本中的稳健性，以及该方法估计的置信区间的精度也值得进一步探索。

本研究采用蒙特卡洛数据模拟方法，比较两独立样本 *t* 检验、Wilcoxon 秩和检验和 Bootstrap 置信区间法在解决极小样本两独立定量资料比较中的表现，以期为相关实验性研究提供方法学参考。

1 资料与方法

1.1 模拟数据的生成

通过蒙特卡洛数据模拟方法生成模拟数据，主要有以下几个模拟情景。样本含量：本研究主要模拟极小样本量下的统计方法表现性能，共模拟 5 种样本量——每组各 2、3、5、10 和 20。均数差：共设置 5 种均数差——0、0.5、1、2 和 3。从均数相同的两总体中抽样，两总体均数差为 0，H_0 成立，且均数差的置信区间包含 0，认为两样本来自同一总体，两组样本均数的不同由抽样误差造成，当统计检验方法拒绝 H_0 时则认为发生 I 类错误。当两样本均数差不为 0 时，两样本不是来自同一样本，若统计检验方法未能拒绝 H_0，则认为发生 II 类错误。样本分布：共设置 3 种总体分布，第 1 种为两样本均服从总体方差为 12 的正态分布，总体均数根据均数差确定（其中一组为 0，即第一组的总体为标准正态分布）；第 2 种（偏态分布一）为两样本服从偏度系数为 1.5，峰度系数为 3.0 的偏态分布；第 3 种（偏态分布二）为两样本服从偏度系数为 1.0，峰度系数为 2.0 的偏态分布 [12]。

对以上三个因素的不同水平进行全排列构建 75 种（5 种样本量 ×5 种均数差 ×3 种总体分布）情景，每种生成 10 000 个模拟数据集。

1.2 检验方法

基于 Bootstrap 法估计均数差的置信区间。采用 Bootstrap 重抽样技术对模拟数据集进行 1 000 次重抽样构建两样本均数差的经验分布。通过估计经验分布的第 2.5% 和第 97.5% 分位数确定均数差的 95%CI。当 95%CI 下限大于 0 或上限小于 0 时，认为两组均数差异有统计学意义，两样本对应的总体均数不同。

参数和非参数假设检验法。采用两独立样本 *t* 检验和 Wilcoxon 秩和检验对两总体均数是否相同进行假设检验。与 Bootstrap 法估计的 95%CI 相对应，假设检验的检验水准 α=0.05，均为双侧检验。

1.3 评价标准

在均数差为 0 时，若 *t* 检验和 Wilcoxon 秩和检验的 *P* 值小于 α，或 Bootstrap 法估计的均数差 95%CI 未跨过 0，认为发生 I 类错误。在均数差不为 0 时，以上情形认为成功检验出统计学差异，即未发生 II 类错误。

分别使用 *t* 检验、Wilcoxon秩和检验和 Bootstrap 置信区间法对 75 种情景下，每种情景的 10 000 个模拟数据集进行分析。计算并比较 3 种方法在不同数据情景下的 I 类错误发生率和 100%–II 类错误发生率（统计效率）。

1.4 统计软件和硬件

本研究使用的统计软件为 R 4.1.3，数据模拟的平台为塔式服务器，处理器型号为 Intel Xeon Gold 6230，内存为 384GB。

2 结果

2.1 I类错误

大样本时 I 类错误的发生与样本量无关，其仅与检验水准 α 有关，但根据本研究的模拟结果，*t* 检验和 Wilcoxon 秩和检验的 I 类错误发生率均小于检验水准（图 1a 和图 1b）。当样本量 *n*=2、*n*=3 时，Wilcoxon 秩和检验的 I 类错误发生率为 0。这是由 Wilcoxon 秩和检验方法特性造成的[8]。对于 *t* 检验来说，极小样本时的 I 类错误发生率小于检验水准 α，尤其是当数据分布为本研究设定的两种偏态分布时更为明显，这可能与此种情形下不适用 *t* 检验有关。但 Bootstrap 置信区间法的 I 类错误发生率较高，当数据服从正态分布时，I 类错误发生率随着样本量的增加而下降，当数据为偏态分布时，I 类错误发生率随着样本量的增加而上升（图 1c）。

2.2 统计效率

三种总体分布下（正态分布、偏态分布一和偏态分布二）分别使用三种方法（*t* 检验、Wilcoxon 秩和检验和 Bootstrap 置信区间法）的统计效率分别如图 2a、图 2b 和图 2c 所示。当均数差较小时，无论使用哪种方法，统计效率都很低，Bootstrap 置信区间法表现略优于另外两种假设检验的方法；当均数差较大时，即使样本量很小，Bootstrap 置信区间法仍有较高的统计效率，说明此时犯 II 类错误的概率较低（图 2c）。

无论数据是否服从正态分布，当样本量极小时（*n*=2、*n*=3），*t* 检验的表现优于 Wilcoxon 秩和检验。但当样本量较大且均数差也较大时，*t* 检验与 Wilcoxon 秩和检验统计效率差异不大（图 2a、图 2b）。

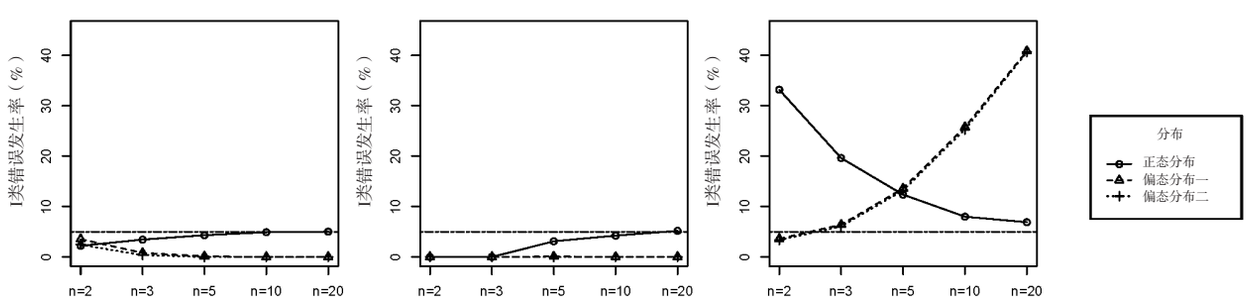

图1a *t*检验　图1b Wilcoxon秩和检验　图1c Bootstrap置信区间法

图1 三种方法的I类错误发生率（%）

Figure 1. Type I error rate of three methods (%)

3 结论

本研究通过数据模拟的方法，探索了采用两独立样本 *t* 检验、Wilcoxon 秩和检验和 Bootstrap 置信区间法对极小样本两独立定量资料进行统计推断时统计效率的差异。由模拟结果可见，相较于 Wilcoxon 秩和检验，*t* 检验在样本量极小时（*n*=2、*n*=3）仍有一定的统计效率，且对总体数据分布不是很敏感。当数据服从本研究设定的两种偏态分布时，*t* 检验的表现不差于 Wilcoxon 秩和检验。在样本量极小时，Bootstrap 置信区间法可以增加统计效率，但在两组样本均数差为 0（即两组样本来自同一总体），且数据服从正态分布时，犯 I 类错误的概率较高。

综上，根据本模拟研究结果，当数据服从正态分布时，建议使用 *t* 检验对极小样本进行

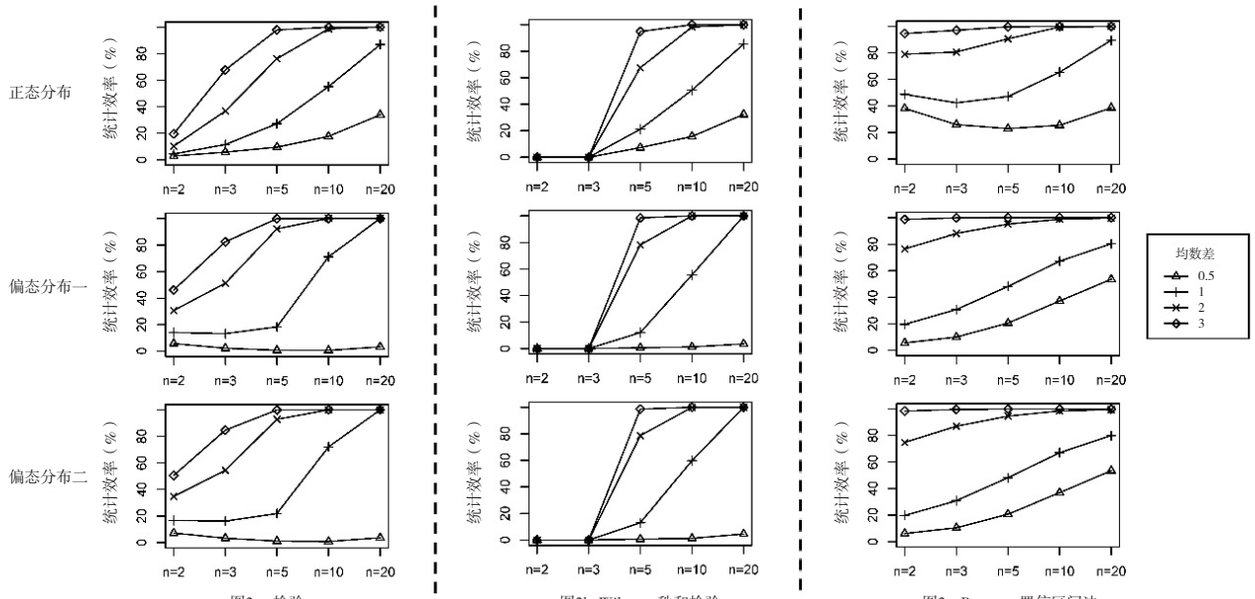

图2 不同情形下三种方法的统计效率（%）

Figure 2. Power of three methods in different scenarios (%)

统计推断；当数据不服从正态分布时，建议使用 Bootstrap 置信区间法对极小样本进行统计推断。由于对于极小样本统计效率太低，当样本量极小时，无论数据服从何种分布，均不建议使用 Wilcoxon 秩和检验进行统计推断。

参考文献

1 白雪，孟宪玉，朱菊茹，等．2 型糖尿病小型猪模型制备方法的研究进展 [J]. 中华实用诊断与治疗杂志，2019, 33(7): 717–719. [Bai X, Meng XY, Zhu JR, et al. Preparation of minipig models of type 2 diabetes mellitus[J]. Journal of Chinese Practical Diagnosis and Therapy, 2019, 33(7): 717–719.] DOI: 10.13507/j.issn.1674–3474.2019.07.026.

2 顾鹏，陈傍柱，徐涛，等．Hpd 基因修饰制备高酪氨酸血症 III 型巴马小型猪模型 [J]. 中国比较医学杂志，2019, 29(5): 11–16. [Gu P, Chen BZ, Xu T, et al. Generation of a Bama minipig model of hereditary tyrosinemia type III by modifying the Hpd gene[J]. Chinese Journal of Comparative Medicine, 2019, 29(5): 11–16.] DOI: 10.3969/j.issn.1671–7856.2019.05.002.

3 周莹，杜湧瑞，Zelinski Mary B, 等．慢性炎症在 X 射线诱发恒河猴卵巢组织持续损伤中的潜在作用 [J]. 天津医科大学学报，2022, 28(2): 160–164. [Zhou Y, Du YR, Zelinski MB, et al. Potential role of chronic inflammation in persistent ovarian injury exposed to X–ray targeted irradiation in rhesus monkeys[J]. Journal of Tianjin Medical University, 2022, 28(2): 160–164.] https://d.wanfangdata.com.cn/periodical/ChlQZXJpb2RpY2FsQ0hJTmV3UzIwMjMwNDI2EhR0aWFuanlrZHh4YjIwMjIwMjAxMBoIaWU5NTVndmI%3D.

4 Bland JM, Altman DG. Analysis of continuous data from small samples[J]. BMJ, 2009, 338: a3166. DOI: 10.1136/bmj.a3166.

5 祝国强．医药数理统计方法 (第 2 版)(BZ)[M]. 北京：高等教育出版社．2009. [Zhu GQ. Mathematical and Statistical Methods in Medicine (2nd edition) (BZ)[M]. Beijing: Higher Education Press. 2009.]

6 娄冬华．秩和检验的统计思想 [J]. 中国卫生统计，2005, 22(4): 264–265, 267. [Lou DH. Statistical thought of rank sum test[J]. Chinese Journal of Health Statistics, 2005, 22(4): 264–265, 267.] DOI: 10.3969/j.issn.1002–3674.2005.04.028.

7 王俊，吴熙．实际应用中方差分析与秩和检验结果比较 [J]. 中国卫生统计，2008, 25(1): 55, 58. [Wang J, Wu X. The results of variance analysis and rank sum test were compared in practical application[J]. Chinese Journal of Health Statistics, 2008, 25(1): 55, 58.] DOI: 10.3969/j.issn.1002–3674.2008.01.019.

8 Siegel S. Nonparametric statistics for the behavioral sciences(1st ed)[M]. Tokyo: McGraw–Hill Kogakusha. 1956.

9 祝国强，杭国明，滕海英，等．谈谈两总体比较的非参数检验方法 [J]. 数理医药学杂志，2011, 24(5): 524–525. [Zhu GQ, Hang GM, Teng HY, et al. On two types of non–parametric tests[J]. Journal of Mathematical

Medicine, 2011, 24(5): 524–525.] DOI: 10.3969/j.issn.1004–4337.2011.05.006.

10 林正大, 刘平, 黄士铮, 等. 方差非齐及小样本下总体均值差检验的探讨[J]. 上海师范大学学报(自然科学版), 1995, 24(4): 19–23. [Lin ZD, Liu P, Huang SZ, et al. Inquisition into the testing of population mean difference under small sample and non–homo geneity variance[J].Journal of Shanghai Normal University (Natural Sciences), 1995, 24(4): 19–23.] DOI: CNKI:SUN:SHDZ.0.1995–04–003.

11 Efron, Bradley. The jackknife, the bootstrap and other resampling plans[J]. Society for Industrial and Applied Mathematics, 1982. DOI: 10.1137/1.9781611970319.ch3.

12 Fleishman A. A method for simulating non–normal distributions[J]. Psychometrika, 1978, 43(4): 521–532. DOI: 10.1007/bf02293811.

收稿日期：2023 年 02 月 15 日 修回日期：2023 年 03 月 27 日
本文编辑：李 阳 黄 笛

引用本文：郭轶斌, 李佳迅, 吴骋, 等. 极小样本两独立定量资料假设检验方法比较[J]. 数理医药学杂志, 2023, 36(7): 481–485. DOI: 10.12173/j.issn.1004–4337.202302073
Guo YB, Li JX, Wu C, et al. The comparison of hypothesis testing methods for two independent quantitative data with extremely small samples[J]. Journal of Mathematical Medicine, 2023, 36(7): 481–485. DOI: 10.12173/j.issn.1004–4337.202302073