Statistics and Application 统计学与应用, 2025, 14(1), 31-40
Published Online January 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.141004

Hans汉斯

基于异质性检验方法的统计模拟研究

练锴雯

南方医科大学公共卫生学院，广东 广州

收稿日期：2024年12月8日；录用日期：2025年1月3日；发布日期：2025年1月13日

摘　要

目的：在流行病学研究中，分层因素可能会影响研究因素与结局变量之间的关系。因此异质性检验在分析暴露与结局的关联性时尤为重要，以确保研究结果的科学性。常用的异质性检验方法包括Breslow-Day法、Tarone法和Logistic回归分析法，但它们在不同分层数据下的I类错误控制和检验效能表现存在差异。本研究旨在通过蒙特卡洛模拟实验评估Breslow-Day法、Tarone法和Logistic回归法在不同样本量和分层因素下的表现，以找出在分层异质性检验中最稳健的统计方法，为流行病学研究提供参考。方法：利用R软件生成模拟数据，分别针对单分层(性别)和双分层(性别和年龄)条件下多次模拟不同样本量的数据，评估各方法在一类错误和检验效能方面的表现。结果：在单分层情况下，Breslow-Day和Tarone法在小样本时的I类错误控制较弱，随着样本量增大逐渐稳定；Logistic回归法在小样本条件下控制能力较好，表现更为稳健。在双分层情况下，Breslow-Day和Tarone法不再适用，Logistic回归法在校正后可有效控制I类错误。结论：本研究揭示了不同异质性检验方法的适用条件和表现差异，为流行病学和分层分析的实践提供了数据支持。Breslow-Day和Tarone法适用于简单分层和小样本条件，而Logistic回归法在大样本和多分层条件下表现更优，需适当校正以控制误差。这为后续研究和应用提供了更为科学的异质性检验方法选择依据。

关键词

异质性检验，蒙特卡洛模拟，校正方法

Statistical Simulation Study Based on Heterogeneity Test Methods

Kaiwen Lian

School of Public Health, Southern Medical University, Guangzhou Guangdong

Received: Dec. 8th, 2024; accepted: Jan. 3rd, 2025; published: Jan. 13th, 2025

Abstract

Objective: In epidemiological research, stratification factors may influence the relationship between

文章引用: 练锴雯. 基于异质性检验方法的统计模拟研究[J]. 统计学与应用, 2025, 14(1): 31-40.
DOI: 10.12677/sa.2025.141004

study factors and outcome variables. Therefore, heterogeneity testing is crucial for analyzing the association between exposure and outcomes to ensure the scientific validity of research findings. Common heterogeneity test methods include the Breslow-Day test, Tarone's test, and Logistic regression analysis, but their performance in terms of Type I error control and test power varies under different stratified data conditions. This study aims to evaluate the performance of the Breslow-Day test, Tarone's test, and Logistic regression method under varying sample sizes and stratification factors through Monte Carlo simulation experiments, to identify the most robust statistical method for stratified heterogeneity testing, and provide references for epidemiological research. Methods: Simulated data were generated using R software under single-stratification (gender) and double-stratification (gender and age) conditions. Multiple datasets with varying sample sizes were simulated to evaluate the performance of each method in terms of Type I error control and test power. Results: Under single-stratification, the Breslow-Day and Tarone tests showed weak Type I error control for small samples but stabilized as the sample size increased. The Logistic regression method demonstrated better robustness and control in small sample conditions. Under double-stratification, the Breslow-Day and Tarone tests became inapplicable, while the Logistic regression method, after adjustment, effectively controlled Type I errors. Conclusion: This study highlights the applicability and performance differences of various heterogeneity testing methods, providing data-driven support for practice in epidemiology and stratified analysis. The Breslow-Day and Tarone tests are suitable for simple stratification and small sample conditions, whereas the Logistic regression method performs better in large samples and multi-stratification conditions, with necessary adjustments to control errors. These findings offer a scientific basis for selecting heterogeneity testing methods in future studies and applications.

Keywords

Heterogeneity Test, Monte Carlo Simulation, Adjustment Methods

1. 引言

在流行病学研究中，分层因素(如性别、年龄、社会经济地位等)可能会影响研究因素与结局变量之间的关联性。某些分层因素在不同人群中会改变暴露因素的效应，导致研究因素与结局的关系在不同亚群中表现出不同的强度或方向。所以在分析暴露与结局的关联性时，通常需要考虑分层因素的潜在作用，以提高研究结论的准确性和科学性。分层分析在控制潜在混杂因素、提高效应估计的准确性方面至关重要。因此为了判断是否可以合并不同分层的效应估计，需要进行层间同质性检验，即确定在不同分层水平上，研究因素与结局变量的关联效应是否一致(即判断 OR 值是否相等)。目前常用的异质性检验方法包括 Breslow-Day 法、Tarone's 法以及 Logistic 回归分析法。其中，Breslow-Day 和 Tarone's 检验都是基于卡方分布的统计方法，对每个分层进行效应一致性检验；而 logistic 回归分析法则通过纳入分层因素、研究因素及其交互项来评估交互效应的显著性。这些方法在面对不同分层数据下的 I 类错误控制和检验效能存在差异，因此选择适当的方法进行异质性检验尤为重要。本研究以吸烟与患肺癌关系为例，通过蒙特卡洛模拟生成大量的分层模拟数据，以分析性别、年龄等分层因素在吸烟与肺癌关系中的异质性检验表现。进而比较 Breslow-Day 法、Tarone's 法和 Logistic 回归分析法在一类错误控制和检验效能方面的表现，以期找出在分层异质性检验中最稳健的统计方法，为流行病学研究提供参考依据。

2. 模型及理论基础

2.1. Breslow-Day 检验

Breslow-Day 检验[1]可以判断当分层因素存在时，暴露与疾病之间的关联产生的异质性检验。Breslow-Day 检验的原假设为不同分层间的比值比(*OR*)一致，备择假设为至少一个分层的 *OR* 与其他分层显著不同。当我们得到的 p 值低于设定的显著性水平时，就可以拒绝原假设，认为在不同分层间存在效应异质性。这种情况下，通常建议分层报告结果，而不是合并效应。在流行病学分层分析中，不同层次的比值比可以通过 2 × 2 列联表来表示。

假设某一研究有 K 个分层，对于每个分层 k，可以构建 2 × 2 列联表。在本次研究背景中，暴露即为是否吸烟，结局即为是否患肺癌，分层因素即为性别、年龄。在列联表中，第 k 层的比值比(*OR*)为，

$$OR_k = \frac{a_k \cdot d_k}{b_k \cdot c_k} \tag{1}$$

Breslow-Day 检验通过检验这些分层 *OR* 值的差异来判断异质性。基于各层的期望频数来计算检验统计量 Q 的公式为，

$$Q = \sum_{k=1}^{K} \frac{\left(O_k - E_k\right)^2}{V_k} \tag{2}$$

其中，O_k 是第 k 层的观测比值比；E_k 是假设比值比一致下的期望比值比；V_k 是第 k 层的比值比的方差。

该统计量 Q 服从 $K-1$ 自由度的卡方分布。若 Q 的值较大，则表明各分层的比值比差异显著，存在效应异质性。通过 Breslow-Day 检验，流行病学研究可以更精确地评估暴露因素在不同亚组中对结局的影响差异。这有助于识别特定人群中的显著风险，确保模型在这些亚组中的适用性，从而提高分析结果的科学性和可靠性。

2.2. Tarone 检验

Tarone 检验[2]是对 Breslow-Day 检验的一种修正方法，也是用于评估多个 2×2 列联表中暴露因素和结局的关联是否具有一致性的方法(即检验各表中的比值比 *OR* 是否相等)。TArone's test 主要通过引入连续性校正来修正统计量，使得其更准确地处理样本量较小时的情境。

假设有 K 个 2 × 2 列联表，每个表格表示一个亚组的暴露与结局的关联关系，且每个列联表的比值比为 θ_k，Tarone's Test 的统计量 Q_T 表示为，

$$Q_T = \sum_{K=1}^{K} \frac{\left(O_k - E_k\right)^2}{V_k} - \delta \tag{3}$$

其中，E_K 为第 K 个列联表中暴露组患病的期望值；V_k 是 O_k 的方差；δ 是 Tarone's Test 中的连续性校正项。

$$\delta = \frac{\left(\sum_{k=1}^{K} \frac{O_k - E_k}{V_k}\right)^2}{\sum_{K=1}^{K} \frac{1}{V_k}} \tag{4}$$

Tarone's Test 的校正项通过对各层差异的加权和进行平方并除以方差的倒数和，使得统计量在小样本下更稳定。当假设所有亚组中的 *OR* 相等时，统计量 Q_T 服从 χ^2 分布，自由度为 $K-1$。若统计量 Q_T 的计算结果大于给定的临界值，则拒绝原假设，认为不同亚组间的 *OR* 存在显著差异。

相比于传统的 Breslow-Day 检验，Tarone’s Test 通过引入连续性校正，更适合在小样本量或存在稀疏数据时使用。因此 Tarone’s Test 在样本规模较小或数据分布不均衡的情境中具有更高的准确性，有助于提升异质性检验的稳健性。

2.3. Logistic 回归

Logistic 回归分析是一种常用于研究二分类结局或多分类结局的变量和多个自变量(如暴露因素、分层因素)之间的关系的统计方法。同样在流行病学研究中 Logistic 回归也可用于检验暴露与结局的关联是否在不同分层因素下具有一致性，即检验是否存在交互效应。Logistic 回归用于分层异质性检验时是将分层因素、研究因素、分层因素与研究因素的交互项纳入模型，检验是否存在交互效应。

假设结局变量为 Y，研究因素为 X，分层因素为 Z，Logistic 回归模型为，

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1X+\beta_2Z+\beta_3(X*Z) \tag{5}$$

其中，$P(Y=1)$ 是患病(或结局为 1)的概率；β_0 是截距项；β_1 是研究因素 X 的系数，表示在未考虑分层因素 Z 的情况下，研究因素对患病概率的影响；β_2 是分层因素 Z 的系数，表示分层因素本身对患病概率的影响；β_3 是研究因素和分层因素的交互项系数，用于检测是否存在交互效应。如果交互项显著，则表明研究因素对结局的影响在分层因素的不同水平上有显著差异，即存在异质性。

Logistic 回归能够同时控制多个混杂因素和分层因素的影响，且便于在模型中引入交互项，从而灵活地评估不同因素之间的复杂关联。相比于简单的分层分析，Logistic 回归允许在一个模型中直接量化交互效应，提高了检验结果的精确性。

2.4. Bonferroni 校正

Bonferroni 校正[3] [4]是一种用于多重比较的统计校正方法，旨在降低因多次假设检验而导致的 I 类错误(即假阳性)累积风险。当我们进行多次假设检验时，每次检验都有一定的显著性水平，多个检验的总体 I 类错误率会累加，增加了错误拒绝原假设的可能性。Bonferroni 校正通过将显著性水平 α 按检验次数 m 进行分配，控制多重比较中的整体 I 类错误率。这样只有当某个检验的 p 值小于或等于校正后的显著性水平时，结果才被视为显著，从而控制整体 I 类错误率在原设定的 α 水平之内。当每次检验相互独立时，其公式为，

$$\alpha_{adjusted}=\frac{\alpha}{m} \tag{6}$$

其中，α 是预设的整体显著性水平，而 m 是独立检验的次数。通过降低每个检验的阈值至 $\frac{\alpha}{m}$，可以减小多重比较带来的累积 I 类错误概率，确保结果的稳健性和有效性。

在流行病学中，Bonferroni 校正常用于大规模数据分析中，例如人群中不同暴露因素的比较分析等，以减少多重检验带来的假阳性发现风险，使结果更具科学性和实际意义。

2.5. Benjamini-Hochberg 校正

Benjamini-Hochberg 校正[5] [6]是一种控制多重检验中“假发现率”(False Discovery Rate, FDR)的方法。与控制整体显著性水平的 Bonferroni 校正不同，BH 校正旨在控制多个检验中错误拒绝零假设的比例，适用于在大量比较中减少假阳性但又不严格控制 I 类错误的场合。BH 校正的基本步骤如下：

1) 将所有检验的 p 值按从小到大的顺序排序: $p_{(1)},p_{(2)},p_{(3)},\cdots,p_{(m)}$，其中 m 为检验的总数。

2) 找到最大的 k，使得 $p_{(k)} \le \frac{k}{m}\alpha$，其中 α 是预设的显著性水平。

3) 将前 k 个 p 值所对应的检验结果判为显著，其他结果不显著。

即，BH 校正公式为：

$$p_{(i)} \le \frac{i}{m}\alpha \tag{7}$$

其中，i 表示排序后的第 i 个 p 值位置，m 是检验总数，α 是整体假发现率阈值。

在流行病学和公共卫生研究中，BH 校正常用于多因素暴露的统计分析中，例如当研究多种环境因素对健康结果的影响时，通过控制假发现率来确定显著的因果关系。这种校正比 Bonferroni 方法更灵活，能更好地保持统计效能，因此在大型数据分析中被广泛应用。

3. 模拟方法

本文以吸烟与肺癌的关系出发，研究在单分层、双分层存在时上述方法的 I 类错误率和检验效能。本文设计了不同的实验参数，并通过多次蒙特卡洛模拟生成数据，以评估这些方法在异质性检验中的表现。本文所进行的实验均基于 R 语言 4.1.1 进行。

检验的原假设及备择假设分别为，

$$H_0 : OR_1 = OR_2 = \cdots = OR_k \tag{8}$$

$$H_1 : \text{至少有两个 } OR_i \ne OR_j \tag{9}$$

其中，k 为分层数。

3.1. 单分层因素

当只存在一个分层因素时，研究三种方法的一类错误和检验效能。假设以性别作为分层因素，即吸烟与患肺癌的关系只存在性别为分层时，生成模拟数据的参数设定如下表 1 所示(以下吸烟与患肺癌的参数的设置均为率)。

Table 1. Parameter design for gender stratification
表 1. 性别分层时参数的设计

变量	H_0	H_1
Gender	$\text{Male} : \text{Female} = 1:1$	$\text{Male} : \text{Female} = 1:1$
Smoking	$\text{Smoking}_{\text{male}} : \text{Smoking}_{\text{female}} = 0.3:0.3$	$\text{Smoking}_{\text{male}} : \text{Smoking}_{\text{female}} = 0.5:0.1$
Cancer	$\text{Cancer}_{\text{smoking}} : \text{Cancer}_{\text{unsmoking}} = 0.3:0.03$	$\text{Cancer}_{\text{smoking}} : \text{Cancer}_{\text{unsmoking}} = 0.3:0.03$

在设定参数后我们通过二项分布生成吸烟与肺癌数据，并按照不同的性别和吸烟状态进行分层。在每次模拟中，我们根据上述参数设置生成男性和女性样本数据，分别记录吸烟和患肺癌的情况。此外为了测试不同样本量对检验方法表现的影响，本次实验将样本量分别设为 500，1000、2000、3000、4000 和 5000。

对于每种样本量，我们生成不同性别和吸烟状态下的患癌数据，通过共 10000 次的模拟，获得在不同假设下的检验统计量，并评估各检验方法的 I 类错误率和检验效能。

3.2. 双分层因素

当存在两个分层因素时，研究三种检验方法的一类错误和检验效能。假设两个分层因素末年龄和性

别，即存在年龄和性别两个分层因素时研究吸烟与肺癌关系，生成模拟数据的参数设定如下表 2 所示(以下吸烟与患肺癌的参数的设置均为率)。

假设吸烟人群只存在于 20 岁至 80 岁之间，将年龄分为三组，20 岁至 40 岁为一组，40 岁至 60 岁为一组，60 岁至 80 岁为一组。

Table 2. Parameter design for gender and age stratification
表 2. 性别与年龄分层时参数的设计

变量	H_0	H_1
Age	$Group1:Gropu2:Gropu3=1:1:1$	$Group1:Gropu2:Gropu3=1:1:1$
Male	$Male1:Male2:Male3=1:1:1$	$Male1:Male2:Male3=1:1:1$
Female	$Female1:Female2:Female3=1:1:1$	$Female1:Female2:Female3=1:1:1$
$Smoking_{Male}$	$Smoking_{male1}:Smoking_{male2}:Smoking_{male3}=0.3:0.3:0.3$	$Smoking_{male1}:Smoking_{male2}:Smoking_{male3}=0.3:0.25:0.2$
$Smoking_{Female}$	$Smoking_{Female1}:Smoking_{Female2}:Smoking_{Female3}=0.3:0.3:0.3$	$Smoking_{Female1}:Smoking_{Female2}:Smoking_{Female3}=0.6:0.5:0.1$
Cancer	$Cancer_{smoking}:Cancer_{unsmoking}=0.3:0.03$	$Cancer_{smoking}:Cancer_{unsmoking}=0.3:0.03$

在设定参数后我们通过二项分布生成吸烟与肺癌数据，并按照不同的性别及年龄对吸烟状态进行分层。在每次模拟中，我们根据上述参数设置生成不同年龄组的男性和女性样本数据，分别记录吸烟和患肺癌的情况。此外为了测试不同样本量对检验方法表现的影响，本次实验将样本量分别设为 500、1000、2000、3000、4000 和 5000。

对于每种样本量，我们生成不同性别和吸烟状态下的患癌数据，通过共 10,000 次的模拟，获得在不同假设下的检验统计量，并评估各检验方法的 I 类错误率和检验效能。

3.3. 结果

单因素分层时三种检验方法的一类错误的模拟结果如表 3 所示，随样本量增加的变化情况如图 1 所示。从模拟结果中可以看出当样本量较小时 Breslow-Day 检验和 Tarone 检验方法，未能控制住一类错误，但随着样本量的增大其一类错误能较好的控制在设定的 0.05 附近。而对于 Logistic 回归控制一类错误的能力较好，并且其一直稳定在 0.05 附近。

检验效能模拟结果如表 4 所示，随样本量增加的变化情况如图 2 所示，由表可以得知对于 Breslow-Day 检验、Tarone 检验在小样本时检验效能较低，Logistic 回归在小样本时检验效能较差，而当样本量增大时 Breslow-Day 检验、Tarone 检验的检验效能变化较快，在样本量达到一定时检验效能能够达到 1。

Table 3. Type I error simulation results for gender stratification
表 3. 性别分层时一类错误模拟结果

Test	Sample					
	500	1000	2000	3000	4000	5000
Breslow-Day	0.0549	0.0510	0.0495	0.0489	0.0506	0.0497
Tarone	0.0547	0.0507	0.0493	0.0488	0.0503	0.0496
Logistic	0.0359	0.0455	0.0479	0.0479	0.0496	0.0489

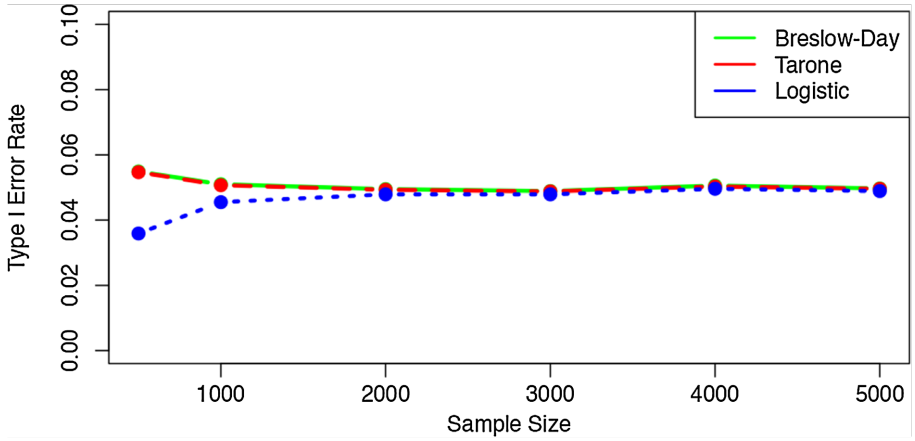

Figure 1. Type I error for gender stratification
图 1. 性别分层时一类错误

Table 4. Simulation results of test power for gender stratification
表 4. 性别分层时检验效能模拟结果

Test	Sample					
	500	1000	2000	3000	4000	5000
Breslow-Day	0.6838	0.9440	0.9986	1.0000	1.0000	1.0000
Tarone	0.6544	0.9343	0.9984	1.0000	1.0000	1.0000
Logistic	0.1918	0.6829	0.9521	0.9903	0.9980	0.9995

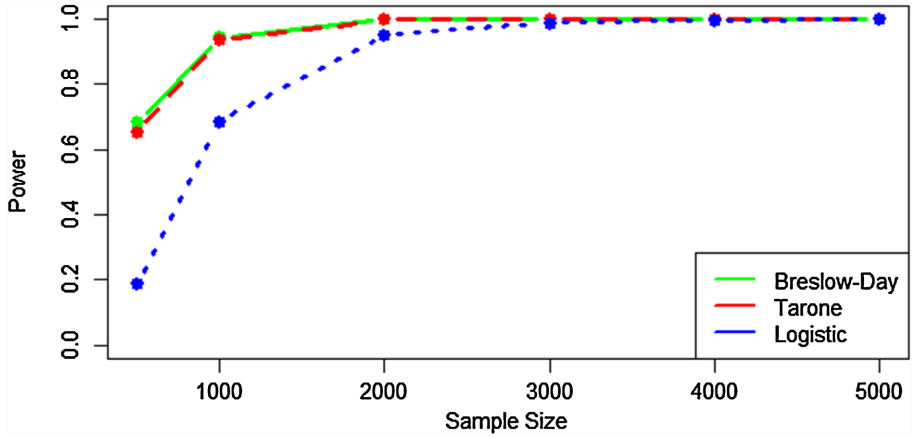

Figure 2. Test power for gender stratification
图 2. 性别分层时检验效能

而对于 Logistic 回归在样本量增加时检验效能增加较慢，在样本量较大时其检验效能也能达到较高的水平。

当存在两个分层因素时，对于 Breslow-Day 检验以及 Tarone 检验并不适用，因此本文只论证 Logistic 回归分析方法的一类错误和检验效能。因为存在多重比较，因此我们引入 Bonferroni 校正和 Benjamini-Hochberg(BH)校正的方法校正 P 值，然后将未校正和两种校正方法的一类错误和检验效能进行比较。

一类错误的模拟结果如表 5 所示，变化情况如图 3 所示。从模拟结果中可以看出对于未校正的情况，一类错误出现膨胀，对于校正后的结果，Bonferroni 当样本量较小时其会出现一类错误保守的情

况，随着样本量的增加其一类错误能够较接近 0.05，对于 Benjamini-Hochberg 校正的结果，其一类错误接近于 0.05。

Table 5. Type I error simulation results for gender and age stratification
表 5. 性别与年龄分层时一类错误模拟结果

Correction	Sample					
	500	1000	2000	3000	4000	5000
Original	0.250	0.271	0.261	0.252	0.278	0.263
Bonferroni	0.036	0.039	0.036	0.044	0.037	0.045
BH	0.038	0.040	0.044	0.047	0.040	0.046

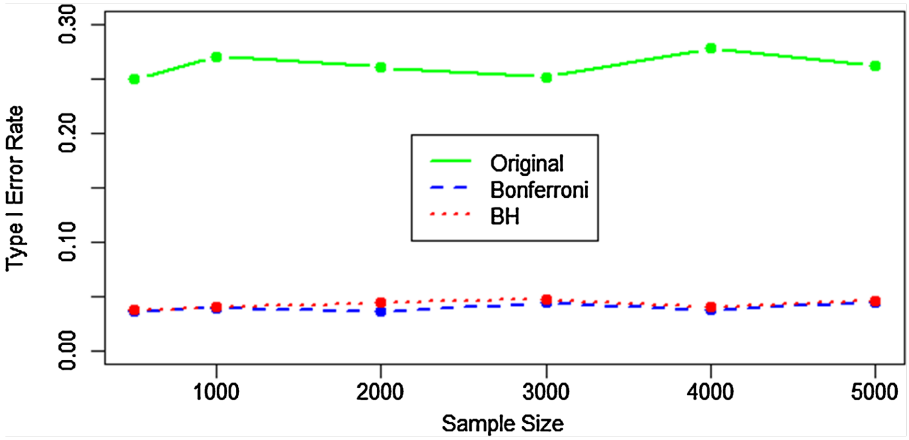

Figure 3. Type I error for gender and age stratification
图 3. 性别与年龄分层时的一类错误

检验效能模拟结果如表 6 所示，变化情况如图 4 所示。从模拟结果中可以看出对于未校正的情况的结果与校正后的结果相差不大，其检验效能均在较高的水平，且当样本量较大时，检验效能都可以达到 1。

Table 6. Simulation results of test power for gender and age stratification
表 6. 性别与年龄分层时的检验效能模拟结果

Correction	Sample					
	500	1000	2000	3000	4000	5000
Original	0.984	1.000	1.000	1.000	1.000	1.000
Bonferroni	0.983	1.000	1.000	1.000	1.000	1.000
BH	0.983	1.000	1.000	1.000	1.000	1.000

4. 总结

4.1. 结论

本研究通过蒙特卡洛模拟和实例验证，系统评估了 Breslow-Day 检验、Tarone 检验和 Logistic 回归分析法在异质性检验中的表现，重点考察了其在一类错误控制和检验效能方面的差异。研究结果表明，

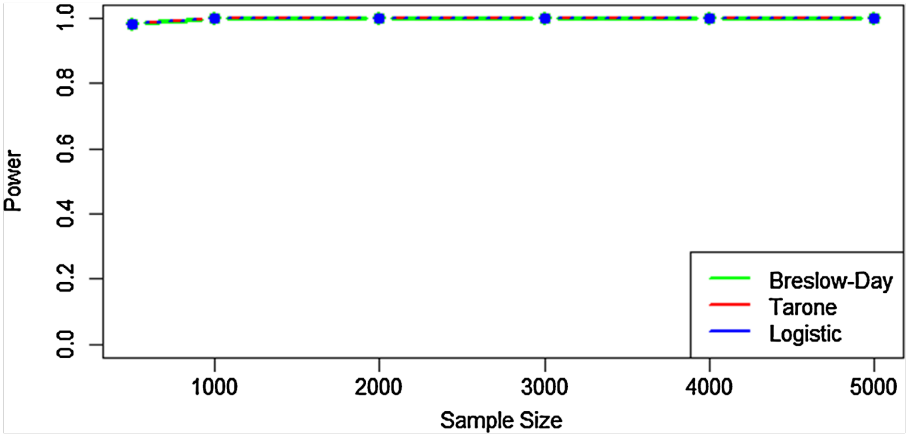

Figure 4. Test power for gender and age stratification
图 4. 性别与年龄分层时的检验效能

不同方法在分层数量和样本量的变化下表现存在一定的差异。

在单分层因素的情境下，Breslow-Day 检验和 Tarone 检验在小样本条件下的一类错误控制能力较弱，可能无法严格控制在 0.05 附近，表现出一定的保守性。随着样本量的增加，这两种方法的一类错误控制逐渐趋于稳定，检验效能也随之提高。然而 Logistic 回归分析在小样本条件下表现更为稳健，不仅一类错误控制能力优越，其整体检验效能也随样本量增加而显著提升。因此，在单分层因素条件下，Logistic 回归法表现相对更加稳健和可靠。

在多分层因素的情境下，Breslow-Day 和 Tarone 检验因假设限制而无法很好地适用。本研究针对这一情况对 Logistic 回归分析法进行了深入分析，并结合 Bonferroni 和 Benjamini-Hochberg 校正进行了比较。结果显示，未校正的 Logistic 回归法在多分层情境下可能导致一类错误膨胀；在校正方法中，Bonferroni 校正在小样本条件下表现出较强的保守性，尽管可以有效控制一类错误，但在检验效能上有所下降。而 Benjamini-Hochberg 校正在一类错误控制和检验效能之间表现出更为平衡的优势，尤其在中等样本量及以上的条件下效果更佳。

对于上述差异的产生主要原因可能有几点，其一，Breslow-Day 检验和 Tarone 检验基于卡方统计量的渐近分布假设，该假设在小样本条件下无法很好地近似真实分布，从而可能导致一类错误偏低或检验效能下降。而 Logistic 回归分析法基于回归模型对回归系数的显著性进行直接检验，不依赖卡方统计量的渐近分布。通过参数估计的置信区间，Logistic 回归能够更准确地反映分层数据中效应的显著性。同时，Logistic 回归的最大似然估计方法在小样本条件下可以更有效地捕获数据特性，因而在一类错误控制和稳定性方面具有显著优势。其二，Logistic 回归法的灵活性使其能够自然地纳入分层变量间的交互效应，从而在多分层条件下表现更优。

综上所述，Breslow-Day 和 Tarone 检验在简单分层因素和小样本量的条件下具有一定应用价值，但在多分层或复杂数据条件下，Logistic 回归法因其更优的一类错误控制和检验效能，表现出显著优势。

4.2. 讨论

研究结果表明，Logistic 回归法在异质性检验中表现出较强的适用性，特别是在样本量有限或研究问题较为复杂的情况下，能够更稳健地控制一类错误并提高检验效能。然而，在实际应用中，研究者需要综合考虑具体研究情境，合理选择方法。为此对于未来的应用有几点建议，在分层因素数量方面，在单分层因素的条件下，可根据样本量选择适当的检验方法。当样本量较小时，Logistic 回归法由于其稳健性

和灵活性更为推荐；而在较大样本条件下，Breslow-Day 和 Tarone 检验可以作为辅助分析工具。对于多分层因素的条件，Breslow-Day 和 Tarone 检验已无法适用，应优先选择 Logistic 回归法。对于样本量的影响方面，在小样本条件下，Logistic 回归法因其对一类错误的良好控制能力而优于传统方法；在大样本条件下，Breslow-Day 和 Tarone 检验的检验效能有所提升，可作为补充方法进行验证分析。当存在多个分层因素需要多重比较校正时，若研究需要进行多重比较校正，应根据具体研究目标和数据特性选择适当的校正方法。对于严格控制一类错误的研究，Bonferroni 校正较为适用；而对于在控制错误率的同时关注检验效能的研究，Benjamini-Hochberg 校正更具优势。

尽管本研究通过模拟和实例验证为未来的研究提供了一定的依据，但仍存在一定的局限性。首先，模拟研究中参数设置可能与实际数据的复杂性存在差异，结果的广泛适用性需进一步验证。其次，研究主要探讨了传统的多重比较校正方法，未来可结合贝叶斯校正或随机效应模型等现代统计方法，进一步提升模型的适用性和稳健性。

总之，本研究为异质性检验方法的选择提供了有益的指导，尤其是强调了 Logistic 回归法在小样本和多分层因素条件下的优势。在未来的研究中，可进一步探索新方法的优化及其在更复杂数据情境中的应用潜力，以为异质性分析提供更全面的解决方案。

参考文献

[1] Aguerri, M.E., Galibert, M.S., Attorresi, H.F. and Prieto Marañón, P. (2007) Erroneous Detection of Nonuniform DIF Using the Breslow-Day Test in a Short Test. *Quality & Quantity*, **43**, 35-44. https://doi.org/10.1007/s11135-007-9130-2

[2] Tamura, R.N. and Young, S.S. (1986) The Incorporation of Historical Control Information in Tests of Proportions: Simulation Study of Tarone's Procedure. *Biometrics*, **42**, 343. https://doi.org/10.2307/2531054

[3] Armstrong, R.A. (2014) When to Use the Bonferroni Correction. *Ophthalmic and Physiological Optics*, **34**, 502-508. https://doi.org/10.1111/opo.12131

[4] Curtin, F. and Schulz, P. (1998) Multiple Correlations and Bonferroni's Correction. *Biological Psychiatry*, **44**, 775-777. https://doi.org/10.1016/s0006-3223(98)00043-2

[5] Ghosh, D. (2020) Wavelet-Based Benjamini-Hochberg Procedures for Multiple Testing under Dependence. *Mathematical Biosciences and Engineering*, **17**, 56-72. https://doi.org/10.3934/mbe.2020003

[6] Ferreira, J.A. (2007) The Benjamini-Hochberg Method in the Case of Discrete Test Statistics. *The International Journal of Biostatistics*, **3**, Article 11. https://doi.org/10.2202/1557-4679.1065