Advances in Applied Mathematics 应用数学进展, 2024, 13(5), 1964-1970
Published Online May 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.135184

Hans汉斯

双边数据风险差的一致性检验

张延欣

南京邮电大学理学院，江苏 南京

收稿日期：2024年4月16日；录用日期：2024年5月9日；发布日期：2024年5月16日

摘　要

本文探讨了多组双边数据风险差的一致性假设检验问题及其检验过程，在一致性检验中，当样本量较小时，Wald统计量和Score统计量的第一类错误率接近于预设的显著性水平0.05，然而似然比统计量显示出了比较膨胀的第一类错误率。当样本量较大时，Score统计量检验效果更佳。随着样本量的增加，所有统计量检验效果趋于稳健。因此，在评估第一类错误率性能时，对于多组双边数据，建议采用统计量Score统计量进行风险差的一致性检验。

关键词

双边数据，一致性检验，第一类错误率

Homogeneity Test of Risk Difference for Bilateral Data

Yanxin Zhang

College of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Apr. 16th, 2024; accepted: May 9th, 2024; published: May 16th, 2024

Abstract

This article explores the homogeneity testing of risk difference for multiple sets of bilateral data and its testing process. In the homogeneity test, when the sample size is small, the Type I error rates of Wald statistic and Score statistic are close to the preset significance level of 0.05, while Likelihood ratio statistic exhibits a relatively inflated Type I error rate. However, when the sample size is larger, the testing effect of Score statistic is better. As the sample size increases, the testing effects of all statistical measures tend to become more robust. Therefore, when evaluating the performance of Type I error rates for multiple sets of bilateral data, it is recommended to adopt Score statistic for the homogeneity test of risk difference.

文章引用: 张延欣. 双边数据风险差的一致性检验[J]. 应用数学进展, 2024, 13(5): 1964-1970.
DOI: 10.12677/aam.2024.135184

Keywords

Bilateral Data, Homogeneity Test, Type I Error Rate

1. 引言

在医学临床试验中，若研究对象是患者身体成对器官，如耳朵、眼睛等，便会产生相关双边数据，研究者通常将器官患病称之为出现响应。相比较于肝脏、心脏这些单一器官，成对器官产生的双边数据更加复杂。Morris [1]发现对双边配对数据的统计检验如果忽略了内相关性的存在，可能会导致检验具有夸大的显著性水平。针对双边数据的研究，Ronser [2]提出了一个假设当一侧器官出现响应时，另外一侧器官也出现响应的条件概率与无条件概率成正比的模型。在此模型基础上，研究者提出了 Donner 模型[3]和 Dallal 模型[4]等适用于研究双边数据的模型。一般来说，风险差经常被用于衡量分组双边数据组间响应率的差异性[5]。Zhang 等人[6]认为，检测风险差一致性是临床试验中一个至关重要的问题。Lui 等[7]研究了在缺失样本数据中两组双侧数据的风险差的一致性检验问题。Shen 等[8]研究并推导了两组双边数据风险差一致性检验的三种检验方法。通常而言，鉴于对照组变量的差异性，多个观察组的设置较为常见。因此，在涉及相关配对的双边数据研究中，考虑包含多个观察组和一个对照组的情境是极具意义的[9]。综上所述，双边数据风险差的一致性假设检验问题具有重要的研究价值。本文不仅关注理论层面的探讨，更致力于解决实际应用中的问题。通过对双边数据风险差的一致性进行有效检验，更准确地评估风险差异，为决策提供科学依据。因此，本研究的开展具有重要的理论意义和实际应用价值。

2. Dallal 模型

设 m_i 为第 i ($i=1,2,\cdots,g$)组中的患者数量， m_{hi} 为第 i 组中有 h ($h=0,1,2$)个响应的患者数量， p_{hi} 为第 i 组中无、单边及双边响应的概率，其中 $m_i=m_{0i}+m_{1i}+m_{2i}$， $p_{0i}+p_{1i}+p_{2i}=1$，具体数据结构如表 1 所示。

Table 1. Data structure
表 1. 数据结构

响应数(h)	组数(g)			
	1	2	...	g
0	m_{01}	m_{02}	...	m_{0g}
1	m_{11}	m_{12}	...	m_{1g}
2	m_{21}	m_{22}	...	m_{2g}
合计	m_1	m_2	...	m_g

记 X_{ijk} 是第 i 组中第 j($j=1,\cdots,m_i$)个患者的第 k ($k=1,2$)只器官响应情况的指标，若无响应，则记 $X_{ijk}=0$，否则 $X_{ijk}=1$。Dallal 模型中包含两个假设：1) 第 i 组患者一侧器官有响应的概率为 $P\left(X_{ijk}=1\right)=\lambda_i$ $\left(0\le\lambda_i\le1\right)$；2) 患者一侧器官有响应，另一侧器官也有响应的概率为 $P\left(X_{ijk}=1\mid X_{ij(3-k)}=1\right)=\gamma_i$。基于假设，

可计算出第 i 组中无、单边及双边响应的概率分别为：

$$p_{0i}=1-(2-\gamma_i)\lambda_i,\ p_{1i}=2(1-\gamma_i)\lambda_i,\ p_{2i}=\gamma_i\lambda_i.$$

3. 风险差的一致性检验

假设第一组为对照组，其余组实验组，那么风险差 $\Delta_i=\lambda_i-\lambda_1\ (i=2,3,\cdots,g)$。一致性检验考虑的问题是各实验组与对照组之间的风险差是否相等，即

$$H_0:\Delta_2=\Delta_3=\cdots=\Delta_g=\Delta \quad \text{vs} \quad H_a:\Delta_r\neq\Delta_s\ (r\neq s)$$

如果不能否定原假设 H_0，则认为各实验组与对照组之间的响应率无显著性差异。

3.1. H_a 和 H_0 下的极大似然估计

记 $M=(m_1,m_2,\cdots,m_g)$，$\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_g)$，$\gamma=(\gamma_1,\gamma_2,\cdots,\gamma_g)$，根据数据结构，$H_a$ 下的对数似然函数可以写为：

$$l_{11}(\lambda,\gamma\mid M)=\sum_{i=1}^{g}\left(m_{0i}\log\left[1+\lambda_i(\gamma_i-2)\right]+m_{1i}\log\left[2\lambda_i(1-\gamma_i)\right]+m_{2i}\log\lambda_i\gamma_i\right)+\log C \tag{1}$$

其中 C 为一个常数，设 $\lambda_i,\gamma_i\ (i=1,2,\cdots,g)$ 和 $\Delta_i\ (i=2,3,\cdots,g)$ 在 H_a 下的极大似然估计分别为 $\hat{\lambda}_i,\hat{\gamma}_i$ 和 $\hat{\Delta}_i$，则 $\hat{\lambda}_i$ 和 $\hat{\gamma}_i$ 的值是偏导方程组 $\partial l_{11}/\partial\lambda_i=0,\partial l_{11}/\partial\gamma_i=0$ 的解，求解方程组可得出 $\hat{\lambda}_i=(m_{1i}+2m_{2i})/2m_i$，$\hat{\gamma}_i=2m_{2i}/(m_{1i}+2m_{2i})$。又因为风险差 $\Delta_i=\lambda_i-\lambda_1$，故风险差的估计值 $\hat{\Delta}_i=(m_{1i}+m_{2i})/2m_i-(m_{11}+m_{21})/2m_1$。在原假设 H_0 条件下，有 $\lambda_i=\lambda_1+\Delta\ (i=2,3,\cdots,g)$，则对数似然函数 l_{11} 等价于：

$$\begin{aligned}l_{10}(\lambda_1,\Delta,\gamma\mid M)=&\sum_{i=2}^{g}\left(m_{0i}\log\left[1+(\lambda_1+\Delta)(\gamma_i-2)\right]+m_{1i}\log\left[2(\lambda_1+\Delta)(1-\gamma_i)\right]+m_{2i}\log(\lambda_1+\Delta)\gamma_i\right)\\&+m_{01}\log\left[1+\lambda_1(\gamma_1-2)\right]+m_{11}\log\left[2\lambda_1(1-\gamma_1)\right]+m_{21}\log\lambda_1\gamma_1+\log C,\end{aligned} \tag{2}$$

设 λ_1,γ_i 和 Δ 在 H_0 下极大似然估计分别为 $\tilde{\lambda}_1,\tilde{\gamma}_i$ 和 $\tilde{\Delta}$。令 l_{10} 关于 λ_1,γ_i 和 Δ 的偏导均为 0，并求解方程组。然而上述方程组没有精确解，故选用费舍尔评分迭代算法计算 $\tilde{\lambda}_1,\tilde{\gamma}_i$ 和 Δ 的近似值，算法过程可简单描述为以下 4 步：

1) 定义各参数的初始值为：

$$\lambda_1^{(0)}=\hat{\lambda}_1,\ \gamma_i^{(0)}=\hat{\gamma}_i,\ \Delta^{(0)}=\frac{1}{g-1}\sum_{i=2}^{g}\Delta_i,$$

2) 第$(t+1)$次迭代，$\tilde{\Delta}^{(t+1)}$ 的估计值更新为：

$$\Delta^{(t+1)}=\Delta^{(t)}-\left(\frac{\partial^2 l_{10}}{\partial\Delta^2}\right)^{-1}\frac{\partial l_{10}}{\partial\Delta},$$

3) 第$(t+1)$次迭代，$\lambda_1^{(t+1)}$ 和 $\gamma_i^{(t+1)}$ 的估计值更新为：

$$\begin{bmatrix}\lambda_1^{(t+1)}\\ \gamma_1^{(t+1)}\\ \vdots\\ \gamma_g^{(t+1)}\end{bmatrix}=\begin{bmatrix}\lambda_1^{(t)}\\ \gamma_1^{(t)}\\ \vdots\\ \gamma_g^{(t)}\end{bmatrix}+I\left(\lambda_1^{(t)},\gamma_1^{(t)},\cdots,\gamma_g^{(t)}\right)^{-1}\begin{bmatrix}\frac{\partial l_{10}}{\partial\lambda_1}\\ \frac{\partial l_{10}}{\partial\gamma_1}\\ \vdots\\ \frac{\partial l_{10}}{\partial\gamma_g}\end{bmatrix}\Bigg|_{\Delta=\Delta^{(t+1)}}$$

其中 I 是费舍尔信息矩阵。

4) 重复步骤 1)~3)，直到所有参数的估计值趋于收敛。

3.2. 检验统计量

构造以下三个常见的检验统计量：似然比检验统计量、Wald 检验统计量和 Score 检验统计量。

3.2.1. 似然比统计量

记 $\hat{\lambda}=\left(\hat{\lambda}_1,\hat{\lambda}_2,\cdots,\hat{\lambda}_g\right)$，$\hat{\gamma}=\left(\hat{\gamma}_1,\hat{\gamma}_2,\cdots,\hat{\gamma}_g\right)$，$\tilde{\lambda}=\left(\tilde{\lambda}_1,\tilde{\lambda}_2,\cdots,\tilde{\lambda}_g\right)$，$\tilde{\gamma}=\left(\tilde{\gamma}_1,\tilde{\gamma}_2,\cdots,\tilde{\gamma}_g\right)$。为了检验假设 H_0，构造似然比统计量 T_L^H 为：

$$T_L=2\left[l_{11}\left(\hat{\lambda},\hat{\gamma}\mid M\right)-l_{10}\left(\tilde{\lambda}_1,\tilde{\Delta},\tilde{\gamma}\mid M\right)\right]$$

原假设 H_0 下，T_L^H 渐近服从自由度为 g-2 的 $\mathcal{X}^2$ 分布。

3.2.2. Wald 统计量

记 $\beta_1=\left(\lambda_1,\Delta_2,\Delta_3,\cdots,\Delta_g,\gamma_1,\gamma_2,\cdots,\gamma_g\right)_{1\times 2g}$，设 $\hat{\beta}_1$ 是 β_1 在 H_a 下的全局极大似然估计。假设 H_0 等价于 $\Delta_2-\Delta_3=\Delta_3-\Delta_4=\cdots=\Delta_{g-1}-\Delta_g$，即 $C_1\beta_1^{\mathrm{T}}=0$，其中

$$C_1=\begin{bmatrix}0&1&-1&0&0&\cdots&\cdots&0\\0&0&1&-1&0&\cdots&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\ddots&\vdots&\vdots&\vdots\\0&0&0&0&1&-1&\cdots&0\end{bmatrix}_{(g-2)\times 2g}$$

构造 Wald 检验统计量 T_W^H 为：

$$T_W^H=\left(\beta_1C_1^{\mathrm{T}}\right)\left(C_1P^{-1}C_1^{\mathrm{T}}\right)^{-1}\left(C_1^{\mathrm{T}}\beta_1^{\mathrm{T}}\right)\Big|_{\beta_1=\hat{\beta}_1},$$

其中 P 是费舍尔信息矩阵，原假设 H_0 下，T_W^H 渐近服从自由度为 g-2 的 $\mathcal{X}^2$ 分布。

3.2.3. Score 统计量

记 $\beta_2=\left(\lambda_1,\Delta,\gamma_1,\cdots,\gamma_g\right)_{1\times(g+2)}$，设 $\tilde{\beta}_2$ 是 β_2 在 H_0 下的极大似然估计。定义 Score 检验统计量 T_S 为：

$$T_S^H=U_1P^{-1}U_1^{\mathrm{T}}\Big|_{\beta_2=\tilde{\beta}_2},$$

其中 $U_1=\left(\frac{\partial l_{10}}{\partial\lambda_1},\frac{\partial l_{10}}{\partial\Delta_2},\frac{\partial l_{10}}{\partial\Delta_3},\cdots,\frac{\partial l_{10}}{\partial\Delta_g},\frac{\partial l_{10}}{\partial\gamma_1},\frac{\partial l_{10}}{\partial\gamma_2},\cdots,\frac{\partial l_{10}}{\partial\gamma_g}\right)_{1\times 2g}$，$P$ 是费舍尔信息矩阵，原假设 H_0 下，T_S^H 渐近服从自由度为 g-2 的 $\mathcal{X}^2$ 分布。

4. 数值模拟研究

在 Monte-Carlo 模拟中，设置每组的样本数量为 $m=50,100,150$，组数 $g=3,5$。记 $\lambda=\left(\lambda_1,\lambda_2,\cdots,\lambda_g\right)$，$\gamma=\left(\gamma_1,\gamma_2,\cdots,\gamma_g\right)$，在 λ_i 和 $\gamma_i\left(i=1,2,\cdots,g\right)$ 的选择上，必须确保响应率 p_{hi} 的取值在 0.1 到 0.9 之间，否则可能会导致出现数据为 0 的情况，从而产生不准确的结果，具体参数设置如表 2 所示。

为了评估三个检验统计量在检验中的性能，计算并比较了其在不同参数设置下的第一类错误率。在表 2 的每组参数配置下，随机模拟生成 10,000 个样本，并统计其中检验 p 值小于显著性水平的次数，通过拒绝次数除以 10,000，计算出经验第一类错误率。如果检验的第一类错误率小于 0.04 或大于 0.06，则意味检验表现过于保守或膨胀，否则是稳健的[10]。

Table 2. Parameter configuration
表 2. 参数配置

参数		组数	
		$g = 3$	$g = 5$
λ_1	i	0.2	0.2
	ii	0.3	0.3
γ	a	(0.4, 0.6, 0.4)	(0.4, 0.6, 0.4, 0.6, 0.4)
	b	(0.5, 0.6, 0.5)	(0.5, 0.6, 0.5, 0.6, 0.5)
	c	(0.6, 0.6, 0.6)	(0.6, 0.6, 0.6, 0.6, 0.6)

在原假设 H_0 条件下，计算上述统计量在一致性检验中犯第一类错误的概率，取 $\Delta = 0.05, 0.075, 0.1$ 且 $\Delta_2 = \Delta_3 = \cdots = \Delta_g = \Delta$。在表 2 中的每个参数配置下随机生成 10,000 个样本，通过计算在显著性水平 $\alpha = 0.05$ 下的拒绝 H_0 的比例求得第一类错误率，结果如表 3 和表 4 所示。结果表明，T_W^H 和 T_S^H 的第一类错误率接近于显著性水平，而 T_L^H 的第一类错误率在 $m = 50$ 时表现非常膨胀。同时可以发现，一致性检验统计量 $T_K^H\ (K = L, W, S)$ 的第一类错误率均随着样本量的增大而趋于稳健。

此外，在 H_0 假设下，取 $m = 50, 100, 150$，在随机生成 1000 组参数 (λ, γ)。对于每种参数设置，每个检验重复 10,000 次，然后计算第一类错误率。通过图 1 中的一组箱线图，比较了 $m = 50, 100, 150$ 情况下，上述统计量在第一类错误率方面的表现。结果表明：在一致性检验中，当 $m = 50$ 时，T_W^H 和 T_S^H 的第一类错误率接近于显著性水平 0.05，而 T_L^H 则产生了较为膨胀的第一类错误率。当样本量数较大时，T_S^H 检验效果更好。所有统计量 $T_K^H\ (K = L, W, S)$ 随着样本量的增加也更加稳健。因此，基于三个统计量在第一类错误率的表现，对于多组相关配对数据，推荐构建统计量 T_S^H 进行风险差的一致性检验。

Table 3. Type I Error Rate of Each Statistic when $g = 3$
表 3. $g = 3$ 时各统计量第一类错误率

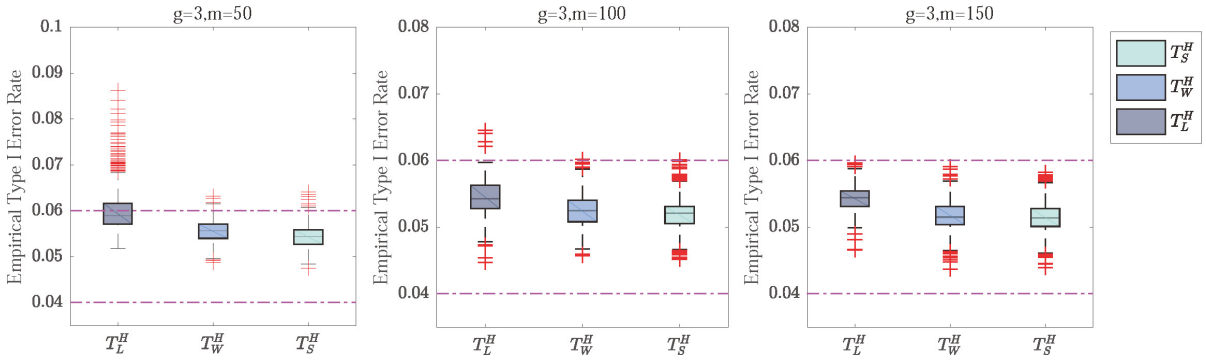
			$m = 50$			$m = 100$			$m = 150$		
Δ	λ_1	γ	T_L^H	T_W^H	T_S^H	T_L^H	T_W^H	T_S^H	T_L^H	T_W^H	T_S^H
0.05	0.2	a	0.069	0.056	0.055	0.065	0.054	0.054	0.057	0.054	0.054
		b	0.083	0.060	0.058	0.064	0.053	0.052	0.057	0.050	0.050
		c	0.081	0.058	0.058	0.058	0.048	0.046	0.063	0.057	0.056
	0.3	a	0.071	0.054	0.052	0.063	0.051	0.050	0.060	0.054	0.053
		b	0.077	0.055	0.053	0.067	0.056	0.054	0.056	0.049	0.049
		c	0.078	0.056	0.054	0.066	0.054	0.054	0.059	0.052	0.051
0.075	0.2	a	0.068	0.053	0.051	0.058	0.053	0.053	0.057	0.054	0.055
		b	0.070	0.055	0.054	0.057	0.052	0.052	0.054	0.052	0.052
		c	0.071	0.056	0.054	0.054	0.047	0.046	0.058	0.055	0.055
	0.3	a	0.074	0.056	0.056	0.061	0.054	0.055	0.054	0.052	0.052
		b	0.069	0.054	0.053	0.060	0.053	0.053	0.054	0.051	0.052

续表

		c	0.070	0.056	0.054	0.059	0.054	0.053	0.061	0.058	0.058
0.10	0.2	a	0.064	0.055	0.055	0.056	0.054	0.054	0.050	0.049	0.050
		b	0.066	0.057	0.055	0.056	0.053	0.054	0.048	0.046	0.047
		c	0.066	0.058	0.057	0.054	0.052	0.052	0.055	0.052	0.054
	0.3	a	0.074	0.056	0.062	0.057	0.052	0.054	0.056	0.052	0.054
		b	0.071	0.059	0.061	0.055	0.050	0.052	0.054	0.051	0.053
		c	0.070	0.059	0.059	0.058	0.053	0.055	0.049	0.047	0.048

Table 4. Type I error rate of each statistic when $g = 5$
表 4. $g = 5$ 时各统计量的第一类错误率

			$m = 50$			$m = 100$			$m = 150$		
Δ	λ_1	γ	T_L^H	T_W^H	T_S^H	T_L^H	T_W^H	T_S^H	T_L^H	T_W^H	T_S^H
0.05	0.2	a	0.070	0.066	0.057	0.061	0.059	0.053	0.057	0.057	0.053
		b	0.069	0.063	0.056	0.059	0.057	0.052	0.054	0.053	0.051
		c	0.071	0.063	0.057	0.059	0.059	0.052	0.055	0.056	0.052
	0.3	a	0.069	0.064	0.054	0.060	0.057	0.053	0.051	0.048	0.047
		b	0.071	0.064	0.055	0.061	0.057	0.052	0.061	0.059	0.056
		c	0.071	0.064	0.056	0.057	0.056	0.049	0.060	0.059	0.055
0.075	0.2	a	0.063	0.061	0.053	0.059	0.059	0.056	0.055	0.056	0.053
		b	0.067	0.066	0.057	0.055	0.054	0.052	0.055	0.055	0.053
		c	0.068	0.068	0.059	0.053	0.054	0.050	0.058	0.057	0.056
	0.3	a	0.068	0.063	0.057	0.055	0.053	0.051	0.055	0.054	0.053
		b	0.070	0.064	0.058	0.057	0.056	0.052	0.057	0.056	0.054
		c	0.070	0.066	0.057	0.059	0.057	0.054	0.054	0.053	0.052
0.10	0.2	a	0.060	0.062	0.056	0.054	0.054	0.053	0.060	0.060	0.059
		b	0.064	0.066	0.059	0.057	0.059	0.056	0.057	0.057	0.056
		c	0.062	0.064	0.055	0.057	0.057	0.054	0.054	0.055	0.053
	0.3	a	0.080	0.065	0.071	0.061	0.057	0.057	0.059	0.055	0.060
		b	0.070	0.062	0.061	0.061	0.058	0.056	0.055	0.053	0.054
		c	0.072	0.066	0.061	0.056	0.054	0.052	0.055	0.055	0.054



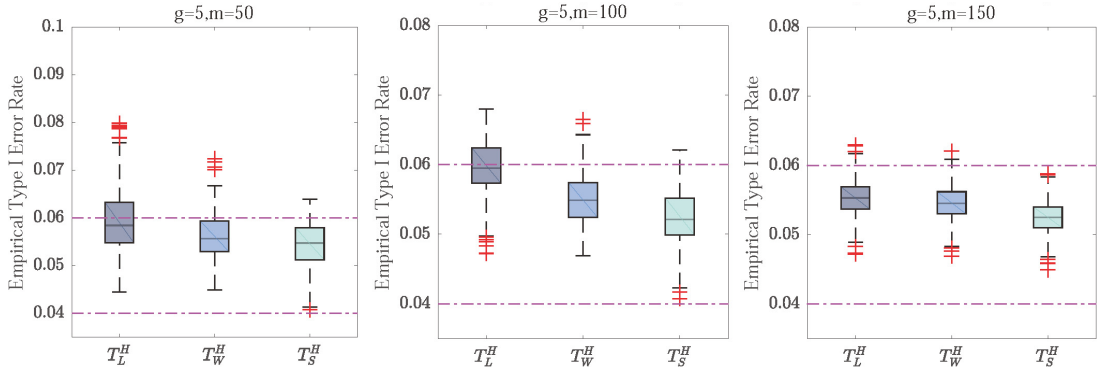


Figure 1. Box Plot of the Type I Error Rate of Each Statistic under 1000 Parameters
图 1. 1000 个参数下各统计量第一错误率箱线图

5. 结论

本文提出了双边数据风险差的一致性假设检验问题及其检验过程，模拟研究发现，当样本量较小时，Wald 统计量和 Score 统计量优于似然比统计量。当样本量较大时，Score 统计量检验效果更好。因此，针对第一类错误率性能的考量，对于多组相关配对数据，推荐构建 Score 统计量进行风险差的一致性检验。本研究还具有广阔的创新空间，未来可以深入研究其他统计量在风险差一致性检验中的表现，以寻找更优的检验方法。

参考文献

[1] Morris, R.W. (1993) Bilateral Procedures in Randomised Controlled Trials. *The Journal of Bone and Joint Surgery*, **75**, 675-676. https://doi.org/10.1302/0301-620X.75B5.8376419

[2] Rosner, B. (1982) Statistical Methods in Ophthalmology: An Adjustment for the Intraclass Correlation between Eyes. *Biometrics*, **38**, 105-114. https://doi.org/10.2307/2530293

[3] Donner, A. (1989) Statistical Methods in Opthalmology: An Adjusted Chi-Square Approach. *Biometrics*, **45**, 605-611. https://doi.org/10.2307/2531501

[4] Dallal, G.E. (1988) Paired Bernoulli Trials. *Biometrics*, **44**, 253-257. https://doi.org/10.2307/2531913

[5] Lipsitz, S.R., Dear, K.B.G. and Laird, N.M., *et al*. (1998) Tests for Homogeneity of the Risk Difference When Data Are Sparse. *Biometrics*, **54**, 148-160. https://doi.org/10.2307/2534003

[6] Zhang, L., Yang, H. and Cho, I. (2009) Test Homogeneity of Risk Difference across Subgroups in Clinical Trials. *Journal of Biopharmaceutical Statistics*, **19**, 67-76. https://doi.org/10.1080/10543400802527874

[7] Lui, K.J. (2005) A Simple Test of the Homogeneity of Risk Difference in Sparse Data: An Application to a Multicenter Study. *Biometrical Journal*, **47**, 654-661. https://doi.org/10.1002/bimj.200410150

[8] Shen, X. and Ma, C.X. (2018) Testing Homogeneity of Difference of Two Proportions for Stratified Correlated Paired Binary Data. *Journal of Applied Statistics*, **45**, 1410-1425. https://doi.org/10.1080/02664763.2017.1371679

[9] Kropf, S., Hothorn, L.A. and Lauter, J. (1997) Multivariate Many-to-One Procedures with Applications to Preclinical Trials. *Drug Information Association*, **31**, 433-447. https://doi.org/10.1177/009286159703100214

[10] Tang, M.L., Tang, N.S. and Ronser, B. (2006) Statistical Inference for Correlated Data in Ophthalmologic Studies. *Statistics in Medicine*, **25**, 2771-2783. https://doi.org/10.1002/sim.2425