Murakami, Hidetoshi; Neuhäuser, Markus

Article — Published Version

A maximum statistic for the one-sided location-scale alternative in the two-stage design

Statistical Methods & Applications

Suggested Citation: Murakami, Hidetoshi; Neuhäuser, Markus (2024) : A maximum statistic for the one-sided location-scale alternative in the two-stage design, Statistical Methods & Applications, ISSN 1613-981X, Springer Berlin Heidelberg, Berlin/Heidelberg, Vol. 34, Iss. 1, pp. 91-112, https://doi.org/10.1007/s10260-024-00775-9

Statistical Methods & Applications (2025) 34:91–112
https://doi.org/10.1007/s10260-024-00775-9

ORIGINAL PAPER

A maximum statistic for the one-sided location-scale alternative in the two-stage design

Hidetoshi Murakami[1] · **Markus Neuhäuser**[2]

Accepted: 8 December 2024 / Published online: 23 December 2024

Abstract
An increase in location is typically accompanied by an increase in variability. Subsequently, the heteroscedasticity can indicate a treatment effect. Therefore, it may be appropriate to perform a location-scale test. A common statistic for a location-scale test is the sum of a location and scale statistic. As demonstrated by Neuhäuser (Biometri J 43:809–819, 2001), weighting the sum increases the power. Although weights cannot usually be reasonably selected a priori, a weighting is possible in an adaptive design using the information obtained in an interim analysis. Here, we propose an adaptive statistic that increases and stabilizes the power. The power performance in various situations for continuous and discrete distributions is investigated using Monte Carlo simulations, which reveal that the proposed statistic increases and stabilizes the power, thus rendering it a strong competitor to existing location-scale statistics. The new statistic is illustrated using real data.

Keywords Adaptive procedure · Interim analysis · Lepage statistic · Maximum statistic

1 Introduction

In randomized clinical trials and other areas, increasing treatment effects are frequently accompanied by increased variability. For example, an observed increase of variance in the experimental group compared with the control group (e.g., placebo group) typically reflects a variation in treatment response among patients in the former group; see Senn (2016, p.968). Then, the experimental group generally has a larger variance and larger mean when (i) a difference exists between groups

✉ Hidetoshi Murakami
hide-murakami@rs.tus.ac.jp

[1] Department of Applied Mathematics, Tokyo University of Science, 1-3 Kagurazaka, Shinjyuku-ku, Tokyo 162-8601, Japan

[2] Department of Mathematics and Technology, Koblenz University of Applied Sciences, Remagen, Germany

and (ii) increasing values indicate better efficacy. In a randomized clinical trial of patients with chronic obstructive pulmonary disease, the mean number of exacerbations per patient was 1.4 and 2.0 in groups treated with a constant dose and gradually reduced dose of prednisone, with standard deviations of 1.5 and 1.9, respectively; see Neuhäuser (2001). Additionally, Brunner et al. (2018, p.127) presented an example of ferritin among children with dwarfism. Furthermore, Zar (2010, p.131) reported the blood-clotting times (in min) of male adult rabbits treated with two different drugs. In these examples, differences exist in location and scale, and the group with the larger mean exhibits greater variability. Further examples using non-clinical data can be found in Neuhäuser (2012, p.62 and p.119). Thus, the treatment can simultaneously affect both location and variability. When homogeneous patients, or in general experimental units, are randomly assigned to different treatments, the equality of variances can be considered a characteristic of the null hypothesis, and apparent heteroscedasticity may indicate treatment differences.

In such a situation, a location-scale test may be appropriate. For instance, nonparametric location-scale statistics were proposed by Cucconi (1968) and Lepage (1971). Various modifications of the Lepage statistic have been proposed, see Pettitt (1976); Büning and Thadewald (2000); Neuhäuser (2000); Kössler (2006); Murakami (2007, 2016); Mukherjee and Marozzi (2019); Kössler and Mukherjee (2020); Mukherjee et al. (2021); Yamaguchi and Murakami (2023). Notably, the statistic of Pettitt (1976) is essentially equivalent to that of Cucconi (1968) for continuous distributions; see Nishino and Murakami (2019). Marozzi (2013b) compared several location-scale statistics, including the Cucconi and Lepage statistics.

Lepage's statistic—the sum of two linear rank statistics—combines the Wilcoxon rank-sum statistic for a location test with the Ansari–Bradley statistic for a dispersion test. When the sum of two linear rank statistics is used, a weighted combination of the two statistics is possible to improve the power (Lepage 1975; Smit et al. 1987). However, in practical applications, the weights cannot be reasonably selected a priori. Therefore, Neuhäuser (2001) proposed the use of a two-stage adaptive design based on Bauer and Köhne (1994). Adaptive designs conduct a sequence of experiments and analyze the data from each stage separately, allowing for preplanned opportunities to modify the trial's course based on accruing information, if necessary (Bauer and Köhne 1994; Pallmann et al. 2018; Dimairo et al. 2020). In our situation, the data from the first stage are analyzed using an unweighted statistic. Thereafter, using the information gained in the interim analysis, weights are selected to analyze the second stage.

The remainder of this paper is organized as follows: In Sect. 2, we revisit the two-sample linear rank statistic. In Sect. 3, we consider a modification to the adaptive one-sided Lepage statistic proposed by Neuhäuser (2001). In Sect. 4, we briefly describe the adaptive designs of Bauer and Köhne (1994) and Lehmacher and Wassmer (1999). In Sect. 5, we present the numerical results based on Monte Carlo simulations. In Sect. 6, we illustrate the proposed statistic using real data. Finally, in Sect. 7, we conclude this study.

2 Two-sample linear rank statistic

Let $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$ denote two random samples of sizes m and n from populations with cumulative distribution functions F_1 and F_2, respectively. The total sample size is denoted by $N = m + n$. Let V_i, $i = 1, \ldots, N$ be 1 if the i^{th} smallest of N observations is from Y; otherwise, it is 0. A one-sided location-scale test is appropriate when the direction of change is predicted and the variability increases with an increase in means. Subsequently, we are interested in testing the following hypothesis:

$$H_0 : F_1(x) = F_2(x),$$

against

$$H_1 : F_2(x) = F_1\left(\frac{x - \theta_1}{\theta_2}\right), \quad \theta_1 > 0 \text{ or } \theta_2 > 1 \text{ or both.}$$

The general form of the two-sample linear rank statistic with score function $a(i)$ is defined as follows:

$$\text{LRS} = \sum_{i=1}^{N} a(i) V_i.$$

The best-known nonparametric statistic for testing the location parameter is the Wilcoxon rank-sum statistic W (Gibbons and Chakraborti 2021), with score function $a(i) = i$. Additionally, among the most famous nonparametric scale statistics is the Ansari–Bradley statistic AB (Gibbons and Chakraborti 2021), with score function $a(i) = (N + 1)/2 - |i - (N + 1)/2|$. When the locations differ, the use of the Ansari–Bradley statistic is strongly discouraged because it may display bizarre behavior as in the case when the maximum number of X elements is less than the minimum number of Y elements. For example, if $m = 6$ and $n = 5$ for all possible samples, the Ansari–Bradley statistic is invariably 21, irrespective of the scale values.

Practically speaking, even when the underlying distributions are continuous, rounding frequently results in ties. For example, observations may be rounded to the first or second decimal point. Thus, we do not always assume continuous distributions but consider discrete distributions. Let $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(N)}$ denote the pooled sample's order statistics. Assumedly, the pooled sample decomposes into d groups of equal observations, the i^{th} group ($i = 1, \ldots, d$) containing t_i observations; that is,

$$Z_{(1)} = \cdots = Z_{(t_1)} < Z_{(t_1+1)} = \cdots = Z_{(t_1+t_2)} < \cdots < Z_{(t_1+\cdots+t_{d-1}+1)} = \cdots = Z_{(t_1+\cdots+t_d)}.$$

Additionally, we define $S_i = \sum_{h=1}^{i} t_h$ and $S_0 = 0$. The number of values associated with X and Y in the i^{th} class is denoted by c_{1i} and c_{2i}, respectively, implying that $c_{1i} + c_{2i} = t_i$ for each $i = 1, \ldots, d$. Thereafter, a tie-adjusted linear rank statistic based on the mid-rank method for score $a(\cdot)$ is defined as

$$\mathrm{LRS} = \sum_{i=1}^{d} c_{2i}\Psi(i), \quad \text{where} \quad \Psi(i) = a\left(\frac{S_i + S_{i-1} + 1}{2}\right).$$

Subsequently, the expected value and variance of LRS under H_0 are given by

$$\mathrm{E[LRS]} = \frac{n}{N}\sum_{i=1}^{d} t_i\Psi(i),$$

$$\mathrm{V[LRS]} = \frac{mn}{N^2(N-1)}\left\{N\sum_{i=1}^{d} t_i\Psi(i)^2 - \left(\sum_{i=1}^{d} t_i\Psi(i)\right)^2\right\},$$

see e.g., Yamaguchi and Murakami (2023). Further, let $\mathrm{LRS}^{\dagger}$ be another tie-adjusted linear rank statistic as follows:

$$\mathrm{LRS}^{\dagger} = \sum_{i=1}^{d} c_{1i}\Psi^{\dagger}(i) \quad \text{where} \quad \Psi^{\dagger}(i) = a^{\dagger}\left(\frac{S_i + S_{i-1} + 1}{2}\right).$$

Thereafter, the covariance of LRS and $\mathrm{LRS}^{\dagger}$ is given by

$$\mathrm{cov}(\mathrm{LRS}, \mathrm{LRS}^{\dagger}) = \frac{mn}{N^2(N-1)}\left\{N\sum_{i=1}^{d} t_i\Psi(i)\Psi^{\dagger}(i) - \sum_{i=1}^{d} t_i\Psi(i)\sum_{i=1}^{d} t_i\Psi^{\dagger}(i)\right\},$$

see e.g., Yamaguchi and Murakami (2023). We refer to Gibbons and Chakraborti (2021) for the formulas for E[LRS], V[LRS], and $\mathrm{cov}(\mathrm{LRS}, \mathrm{LRS}^{\dagger})$ in the absence of ties. However, notably, $d = N$ is possible and can be considered a special case. In this study, we focused on the midrank method. When using the score based on the average score method, another tie-adjusted linear rank statistic is defined as

$$\mathrm{LRS} = \sum_{i=1}^{d} c_{2i}\Psi(i), \quad \text{where} \quad \Psi(i) = \frac{1}{t_i}\sum_{j=S_{i-1}+1}^{S_i} a(j).$$

To evaluate E[LRS], V[LRS], and $\mathrm{cov}(\mathrm{LRS}, \mathrm{LRS}^{\dagger})$ in the presence of ties for certain statistics, we offer the function of R code in Github (https://github.com/h-murakami-stat/One-Sided-Lepage-test/blob/main/Github_20240926_One-Sided-Lepage-Type-Test.R) for some specific scores.

3 One-sided location-scale statistic

3.1 One-sided Lepage and Lepage-type statistics

The statistic for the location-scale test proposed in Lepage (1971) is

$$\mathrm{LEP} = \left(\frac{\mathrm{W} - \mathrm{E[W]}}{\sqrt{\mathrm{V[W]}}}\right)^2 + \left(\frac{\mathrm{AB} - \mathrm{E[AB]}}{\sqrt{\mathrm{V[AB]}}}\right)^2,$$

where E[·] and V[·] are the expected value and variance of W or AB under the null hypothesis for the continuous distribution. Further details can be found in Gibbons and Chakraborti (2021). Notably, the odd translation invariant statistic $a(i) + a(N + 1 - i) =$ "constant" is uncorrelated with the even translation statistic $a^{\dagger}(i) = a^{\dagger}(N + 1 - i)$ under the null hypothesis for continuous distributions. This sufficient condition is known as the Randles and Hogg condition (Randles and Hogg 1971). As W and AB are the odd translation invariant statistic and even translation statistic, respectively, W and AB are uncorrelated under the assumption of continuous distributions under the null hypothesis. However, LEP is unsuitable for one-sided alternatives. Therefore, Neuhäuser (2001) proposed the one-sided Lepage statistic for continuous distributions, as follows:

$$\mathrm{LEP}_1 = \frac{w_1(\mathrm{W} - \mathrm{E[W]}) - w_2(\mathrm{AB} - \mathrm{E[AB]})}{\sqrt{w_1^2\mathrm{V[W]} + w_2^2\mathrm{V[AB]}}}, \quad w_1 = \frac{|\mathrm{STW}|}{|\mathrm{STW}| + |\mathrm{STAB}|}, \quad w_2 = 1 - w_1,$$

where STW and STAB are the standardized statistics for W and AB, respectively. Notably, LEP_1 is asymptotically distributed according to a normal distribution with zero mean and unit variance; see Lepage (1971). However, in practice, ties occur frequently, for example, by rounding to the first decimal point. Thus, we must consider ties and discrete distributions. In this case, W and AB are not uncorrelated, as noted by Rublík (2007). Thereafter, we suggest the one-sided Lepage statistic, namely, OLS_1, in the presence of ties, as follows:

$$\begin{aligned}
\mathrm{OLS}_1 &= \frac{w_3(\mathrm{W} - \mathrm{E[W]}) - w_4(\mathrm{AB} - \mathrm{E[AB]})}{\sqrt{w_3^2\mathrm{V[W]} + w_4^2\mathrm{V[AB]} - 2w_3w_4\mathrm{cov(W, AB)}}},\\
w_3 &= \frac{\mathrm{STP}_1}{\mathrm{STP}_1 + \mathrm{STP}_2},\\
w_4 &= 1 - w_3,\\
\mathrm{STP}_1 &= 1 - \text{p-value of STW for upper tail},\\
\mathrm{STP}_2 &= 1 - \text{p-value of STAB for lower tail},
\end{aligned}$$

where W and AB in OLS_1 are based on tie-adjusted scores. For a detailed expression of E[·] and V[·] of W and AB for the discrete distribution; see Hollander et al. (2013). Furthermore, cov(W, AB) can be easily derived using the code in Github (https://github.com/h-murakami-stat/One-Sided-Lepage-test/blob/main/Github_20240926_One-Sided-Lepage-Type-Test.R).

The Mood statistic (Gibbons and Chakraborti 2021) is a nonparametric two-sample statistic used to test variances. The asymptotic relative efficiencies of the Ansari–Bradley and Mood statistics to the F statistic under the assumption of a normal distribution are 0.609 and 0.76, respectively. Although the asymptotic relative efficiency of the Ansari–Bradley statistic is lower than that of the Mood statistic, the former has

widespread applications; see Lahmiri (2023) and Omer et al. (2023). However, the power of the Mood statistic is higher than that of the Ansari–Bradley statistic for various distributions. Therefore, Murakami and Neuhäuser (2024) considered another one-sided location-scale statistic as follows:

$$\begin{aligned}
\mathrm{OLS}_2 &= \frac{w_5(\mathrm{W} - \mathrm{E}[\mathrm{W}]) + w_6(\mathrm{M} - \mathrm{E}[\mathrm{M}])}{\sqrt{w_5^2 \mathrm{V}[\mathrm{W}] + w_6^2 \mathrm{V}[\mathrm{M}] + 2w_5 w_6 \mathrm{cov}(\mathrm{W}, \mathrm{M})}}, \\
w_5 &= \frac{\mathrm{STP}_1}{\mathrm{STP}_1 + \mathrm{STP}_3}, \\
w_6 &= 1 - w_5, \\
\mathrm{STP}_3 &= 1 - \text{p-value of STM for upper tail},
\end{aligned}$$

where STM is the standardized statistic of M. Notably, M is the even translation statistic, and W and M are uncorrelated under the assumption of continuous distributions under the null hypothesis. As M also satisfies the asymptotic normality theorem of Chernoff and Savage (1958), OLS_2 is asymptotically distributed according to a normal distribution with zero mean and unit variance.

Remark 1 As p-values and test statistics are related one-to-one, we do not have to replace the weights w_1 and w_2 with w_3 and w_4 or w_5 and w_6 for OLS_1 and OLS_2. However, the correlation matrices of OLS_1 and OLS_2 become a singular matrix when $\mathrm{STAB} = \mathrm{STM} = 0$, that is, $\mathrm{M} = \mathrm{E}[\mathrm{M}]$ and $\mathrm{AB} = \mathrm{E}[\mathrm{AB}]$. Therefore, the weight based on the standardized test statistic is not useful for the maximum statistic discussed in the subsequent subsection.

3.2 Maximum statistic and adaptive procedure

In practical analysis, we must determine whether to use OLS_1 or OLS_2 before conducting the hypothesis test. This approach, a so-called adaptive procedure (Hogg et al. 1975), involves employing a selector to select the statistics to use.

3.2.1 Maximum statistic

A simple manner to solve the aforementioned problem is to use the larger of the two statistics as the statistic. Therefore, we propose a maximum statistic based on OLS_1 and OLS_2 as follows:

$$\mathrm{MAX} = \max(\mathrm{OLS}_1, \mathrm{OLS}_2).$$

We assume that $m, n \to \infty$, $m/N \in (0, 1)$. As OLS_1 and OLS_2 satisfy the asymptotic normality theorem, based on Proposition 5 in Kössler (2010), the limiting distribution of $\mathrm{OLS}_{\max}$ is given by

$$\mathbb{P}(\mathrm{MAX} < \ell) = \int_{-\infty}^{\ell}\int_{-\infty}^{\ell} \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\right) dx_1 dx_2,$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$
$$\rho = \frac{\mathrm{E}[\mathrm{OLS}_1\mathrm{OLS}_2] - \mathrm{E}[\mathrm{OLS}_1]\mathrm{E}[\mathrm{OLS}_2]}{\sqrt{\mathrm{V}[\mathrm{OLS}_1]\mathrm{V}[\mathrm{OLS}_2]}} = \mathrm{E}[\mathrm{OLS}_1\mathrm{OLS}_2].$$

3.2.2 Adaptive procedure

The selector statistic should be a measure of the heaviness of the distribution tail. Hogg et al. (1975, 2018) proposed

$$\hat{Q} = \frac{\hat{U}_{0.05} - \hat{L}_{0.05}}{\hat{U}_{0.5} - \hat{L}_{0.5}},$$

where $\hat{L}_\gamma$ and $\hat{U}_\gamma$ denote the averages of the smallest and largest γN order statistics, respectively. When $0.5N$ and $0.05N$ are not integers, the fractional items are used. If $\hat{Q}$ is large, the distribution seems heavy-tailed, whereas a small value indicates that the distribution is light-tailed. Hogg et al. (2018, pp. 623) indicated that the distribution tail seems heavy when $\hat{Q}$ is large (7 or more). Therefore, in this study, we use

$$\mathrm{ADP} = \begin{cases} \mathrm{OLS}_2 & \text{if } \hat{Q} \le 7 \\ \mathrm{OLS}_1 & \text{if } \hat{Q} > 7 \end{cases}.$$

4 Two stage designs

4.1 Two stage design of Bauer and Köhne (1994)

The procedure of Bauer and Köhne (1994) combines the p-values of the separate statistics of the two stages using Fisher's combination statistic. Let p_1 and p_2 be the p-values based on the data of the first and second stages, respectively. For Fisher's combination statistic, the null hypothesis can be rejected at the end of the trial if

$$p_1 p_2 \le c_\alpha = \exp\left[-\frac{1}{2}\chi_4^2(1-\alpha)\right],$$

where $\chi_4^2(1-\alpha)$ is the $(1-\alpha)$-quantile of the central χ^2 distribution with four degrees of freedom. For the case of a nominal significance level of $\alpha = 0.05$, we have $c_{0.05} = 0.0087$; see Bauer and Köhne (1994). In clinical trials, boundaries for early stopping after interim analysis should be incorporated for both ethical and

economic reasons. Let α_0 be the lower limit; therefore, the trial is terminated because of insufficient effects if $p_1 \geq \alpha_0$. A suitable boundary for early stopping without rejecting H_0 could be $\alpha_0 = 0.5$ (Bauer and Köhne 1994). Furthermore, early stopping with the rejection of H_0 is possible if the p-value of the first stage is sufficiently small, that is, $p_1 \leq \alpha_1$. For $\alpha = 0.05$ and $\alpha_0 = 0.5$, it follows $\alpha_1 = 0.0233$ (Bauer and Köhne 1994). Additionally, for $\alpha = 0.025$ with $\alpha_0 = 0.5$, we have $\alpha_1 = 0.0102$ and $c_{0.025} = 0.0038$. This procedure has been widely applied to testing problems other than location and location-scale; see Marozzi (2013a), who proposed and compared several procedures for the scale problem. For further details, refer to Wassmer and Brannath (2016).

4.2 Two stage design of Lehmacher and Wassmer (1999)

Lehmacher and Wassmer (1999) considered the weighted inverse normal combination function

$$C(p_1, p_2) = 1 - \Phi(\xi_1 \Phi^{-1}(1 - p_1) + \xi_2 \Phi^{-1}(1 - p_2)),$$

where ξ_1 and ξ_2 denote pre-specified positive weights such that $\xi_1^2 + \xi_2^2 = 1$ and Φ^{-1} denote the inverse of the standard normal cumulative distribution function $\Phi(\cdot)$. As $C(p_1, p_2)$ is uniformly distributed, we obtain a level α test when using the decision boundaries $\alpha_0 = 1$, $\alpha_1 = 0$, and $c_\alpha = \alpha$. Thereafter, we obtain α_1 such that:

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(p_1, p_2) \leq \alpha\}} dp_2 dp_1 = \alpha.$$

We can select various ξ_1 and ξ_2. However, we assume $\xi_1 = \xi_2 = 1/\sqrt{2}$. Like Sect. 4.1, let α_0 be the lower limit, such that the trial is terminated because of insufficient effects if $p_1 \geq \alpha_0$. A suitable boundary for early stopping without rejecting H_0 could be $\alpha_0 = 0.5$. Furthermore, early stopping with the rejection of H_0 is possible if the p-value of the first stage is sufficiently small, that is, $p_1 \leq \alpha_1$. Thereafter, for $\alpha = 0.05$ and $\alpha_0 = 0.5$, it follows $\alpha_1 = 0.0044$ and $c_{0.05} = 0.05$. Additionally, for $\alpha = 0.025$ with $\alpha_0 = 0.5$, we have $\alpha_1 = 0.0011$ and $c_{0.025} = 0.025$.

5 Numerical Results

5.1 Simulation settings

- Aims: A simulation study aims to investigate the validity of statistics for the one-sided alternative in two-stage designs. We focus on the 5% significance level. For results with the 2.5% significance level, see the supplemental material.
- Data-generating mechanisms:

  Simulation studies provide empirical results for specific scenarios. The performance of the proposed statistic was based on Monte-Carlo simulations performed using R. The simulated powers are obtained by conducting 100,000

Monte-Carlo simulations for each scenario. We use random numbers for the following distributions:

- N(μ, σ^2): normal distribution with mean μ and variance σ^2.
- U(a, b): uniform distribution with interval (a, b).
- C(θ_1, θ_2): Cauchy distribution with location parameter θ_1 and scale parameter θ_2.
- EXP(λ): exponential distribution with rate parameter λ.
- Chisq(ν): chi square distribution with ν degrees of freedom.
- Gum(θ_1, θ_2): Gumbel distribution with location parameter θ_1 and scale parameter θ_2.
- RN(μ, σ^2): rounded normal distribution with mean μ and variance σ^2. Herein, we round the second decimal place.
- NB(η, Prob): negative binomial distribution with number of success η and success probability Prob.
- Pois(λ): Poisson distribution with shape parameter λ.

- Target of analysis:

  Let m_1 and m_2 (n_1 and n_2) denote the sample sizes of X's (Y's) at the first and second stages, respectively. For no interim analysis, the sample sizes are $m_1 + m_2$ ($n_1 + n_2$) for X's (Y's) sample. In Sect. 5.2, we present the performance of the type-I error using an asymptotic distribution for $m_1 = m_2 = n_1 = n_2 = 15$ and 30. In Sect. 5.3, we investigate the power performance when $(m_1, n_1, m_2, n_2) = (25, 25, 25, 25)$ and (35, 35, 15, 15).
- Methods:

  We use the following one-sided statistics with a setting similar to Neuhäuser (2001), using asymptotic critical values for all test statistics:

  - NO_IA: an unweighted OLS_1, OLS_2, ADP or MAX without an interim analysis.
  - FIA_UN: an unweighted OLS_1, OLS_2, ADP or MAX for both stages and combination per Bauer and Köhne (1994).
  - FIA_WE: an unweighted OLS_1, OLS_2, ADP or MAX for the first stage and a weighted OLS_1, OLS_2, ADP or MAX for the second stage and combination per Bauer and Köhne (1994).
  - SIA_UN: an unweighted OLS_1, OLS_2, ADP or MAX for both stages and combination per Lehmacher and Wassmer (1999).
  - SIA_WE: an unweighted OLS_1, OLS_2, ADP or MAX for the first stage and a weighted OLS_1, OLS_2, ADP or MAX for the second stage and combination per Lehmacher and Wassmer (1999).

- Performance measure:

  This describes the numerical quantity used to assess the performance of various statistics. Herein, we focused on the type-I error rate and power of test statistics. Test statistics should be insensitive to changes in the distributions under the null hypothesis. Power is a measure of the ability of the hypothesis-testing setup to detect a particular effect if it is truly present. Therefore, the power performance is used to compare different statistical testing procedures;

that is, a test statistic with a higher power is more efficient than that with a smaller power.

We approximate the standard error of the simulated power with the central limit theorem as follows:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{100000}},$$

where $\hat{p}$ denotes estimated power. The maximum error is obtained for $\hat{p} = 0.5$; thus, the maximum standard error is 0.00158 (for 100,000 simulation runs).

5.2 Type-I error rate

This section presents the type-I error performance using the asymptotic distribution in Fig. 1.

Figure 1 indicates that although we carry out asymptotic tests, the type-I error of all statistics is close to the significance level for small sample sizes. Therefore, the results reveal that we can safely apply the proposed statistic in the scenarios considered in the simulation study. Very small sample sizes are hardly ever used in studies with an interim analysis.

5.3 Power comparison

In this section, we investigate the power performances of the test statistics for various distributions. We use normal and Cauchy distributions as examples of symmetric distributions. Additionally, we use exponential and chi square distributions with positive support and focus on Gumbel distribution with support $\mathbb{R}$ as examples of asymmetric distributions. For the discrete distribution, we use the Poisson distribution as an example. Figures 2, 3, 4, 5, 6 and 7 show the power of various statistics for normal, Cauchy, exponential, chi square, Gumbel, and Poisson distributions, respectively. Simulation results depict that the power(s) of (i) FIA_WE and SIA_WE are similar to or more powerful than that of FIA_UN and SIA_UN, (ii) SIA_WE is similar to or higher than that of FIA_WE in the scenarios considered in the simulation study. Then, we describe the simulation results for the test statistics based on SIA_WE. For the result of comparison of all test statistics, see the supplemental material.

Figures 2, 3, and 6, that is location-scale family of distribution, suggest that the power of (i) OLS_2 is greater than that of OLS_1 for the shifted pure location parameter, (ii) OLS_1 is greater than that of OLS_2 for the changed pure scale parameter and location-scale parameter, (iii) MAX is similar to that of the maximum of $\{\mathrm{OLS}_1, \mathrm{OLS}_2\}$ for all cases, and (iv) MAX is similar to or higher than that of ADP for all cases.

Figures 4 and 5, that is asymmetric distribution with positive support, indicate that the power of (i) OLS_2 is greater than that of OLS_1 for the shifted pure location, changed pure scale, and difference of location-scale parameters and (ii) MAX is similar to that of the maximum of $\{\mathrm{OLS}_1, \mathrm{OLS}_2\}$ and ADP for all cases.

Fig. 1 Simulated type-I error for various distributions with $m_1 = m_2 = n_1 = n_2 = 15$ when $\alpha = 0.05$. Continued for $m_1 = m_2 = n_1 = n_2 = 30$

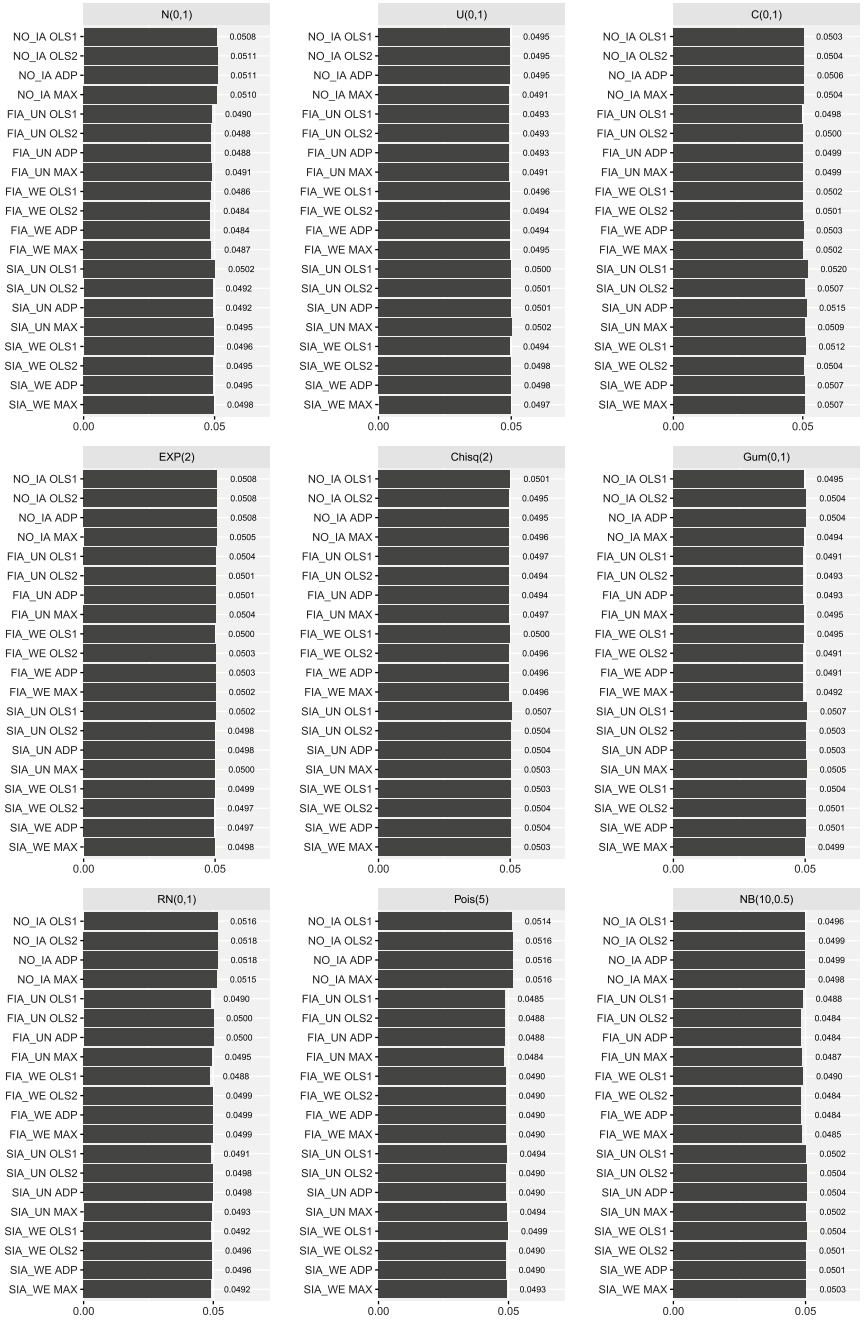

Fig. 1 (continued)

Figure 7 reveals that the power(s) of (i) OLS_2 is greater than that of OLS_1 for

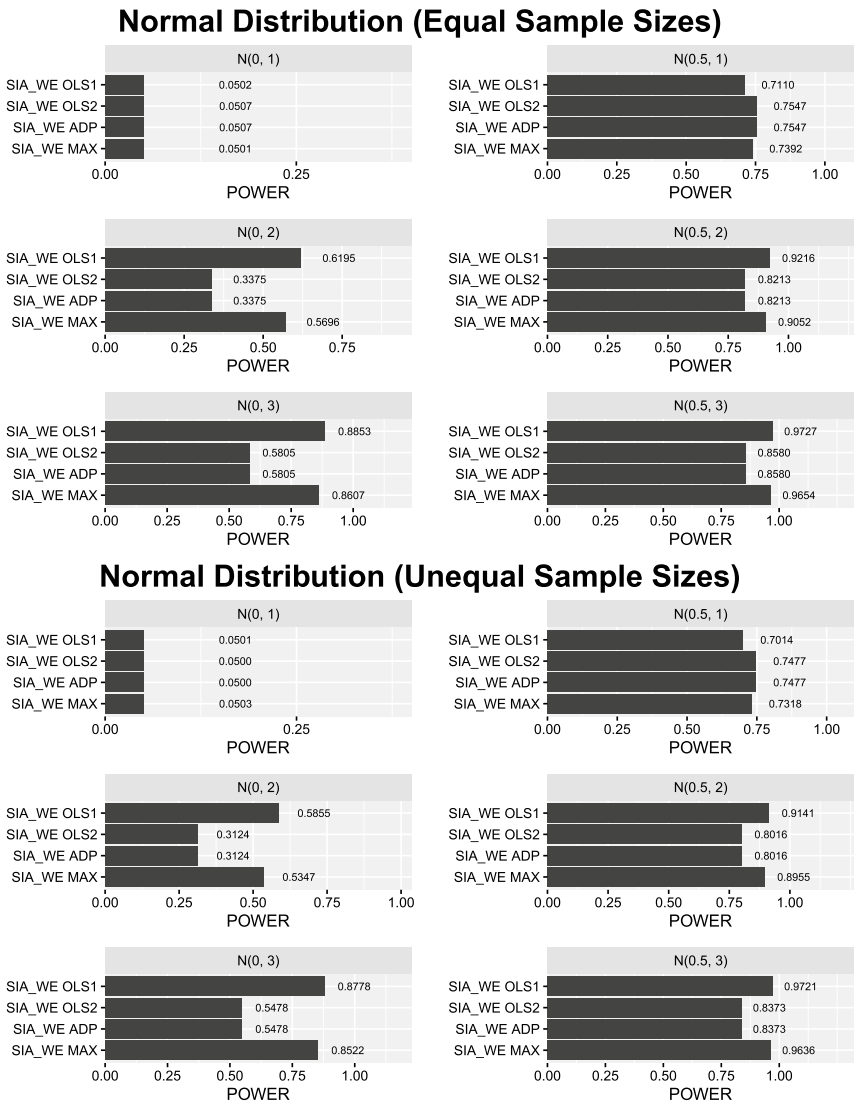

Fig. 2 Simulated type-I error and power for N(0, 1) vs. N(μ_2, σ_2^2) when $\alpha = 0.05$

all cases except Case 5, (ii) ADP and MAX are similar to that of the maximum of $\{\text{OLS}_1, \text{OLS}_2\}$, and (iii) determining the critical winner between ADP and MAX is critical.

Furthermore, Neuhäuser (2001) investigated the power of χ^2 distributions with different degrees of freedom. We compare the power of the difference of the (i) rate parameter for the exponential distribution, (ii) degrees of freedom for χ^2 distribution, and (iii) rate parameter for Poisson distribution in Fig. 8.

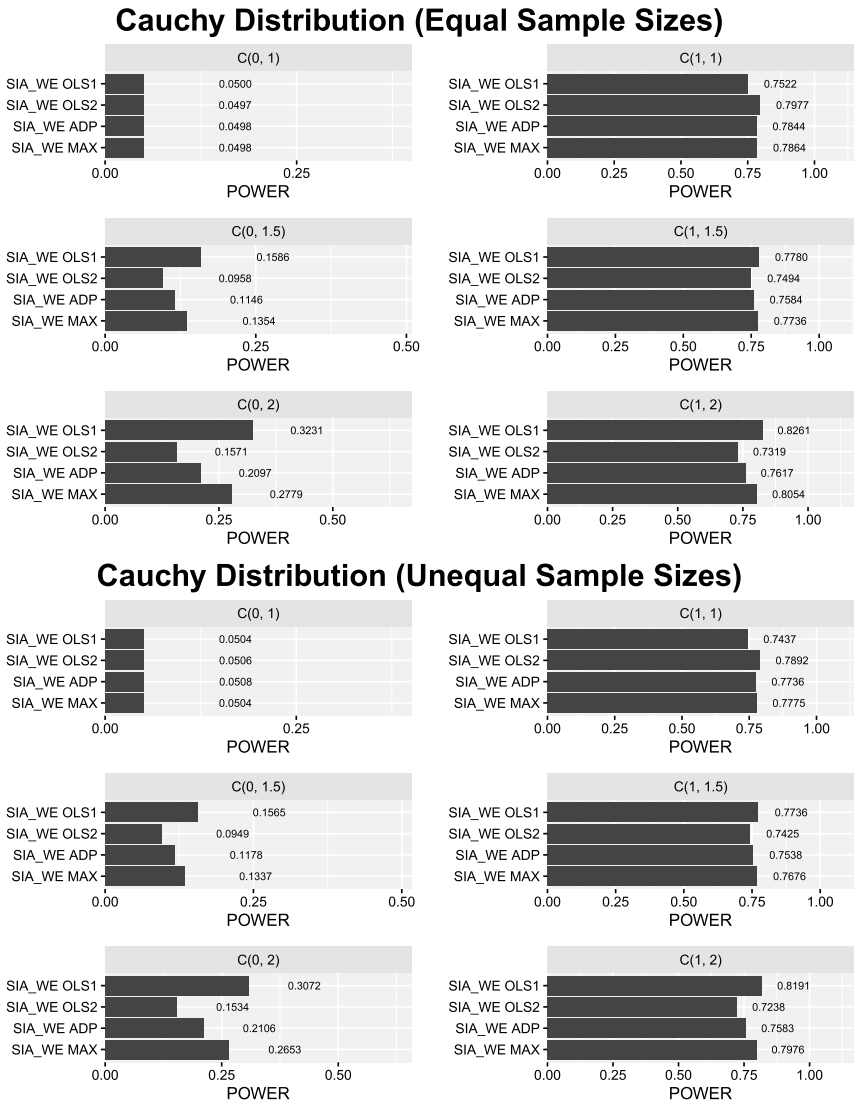

Fig. 3 Simulated type-I error and power for C(0, 1) vs. C(θ_1, θ_2) when $\alpha = 0.05$

According to Figure 8, the power of (i) SIA_WE is higher than that of FIA_WE, (ii) SIA_WE is similar to or more powerful than that of SIA_UN, (iii) ADP is similar to that of MAX, that is, the differences between these statistics are minimal.

Consequently, we recommend using MAX for the one-sided location-scale testing problem in the two-stage design in the scenarios considered in the simulation study. Similar patterns were obtained at a significance level of 2.5% (see the supplemental material).

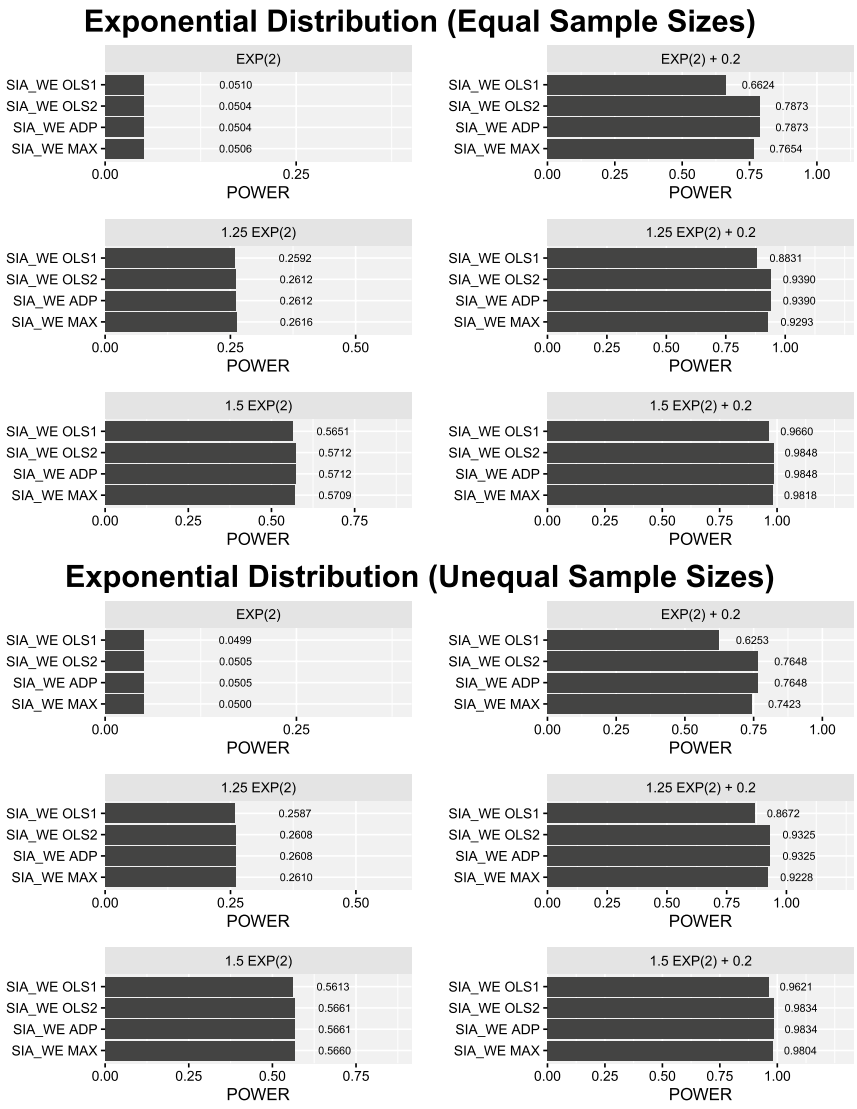

Fig. 4 Simulated type-I error and power for EXP(2) vs. $\theta_2 * \text{EXP}(2) + \theta_1$ when $\alpha = 0.05$

6 Example: Familial adenomatous polyposis

In this section, the tests are described using a real-life dataset. Familial Adenomatous Polyposis is an inherited condition caused by mutations in the Adenomatous Polyposis Coli gene, which causes early and frequent formation of precancerous polyps in the colon at a young age; it invariably results in the development of colon cancer at a young age. As an endpoint, we use the number of colonic polyps, which include two treatments, sulindac and placebo, and consider only men. To illustrate,

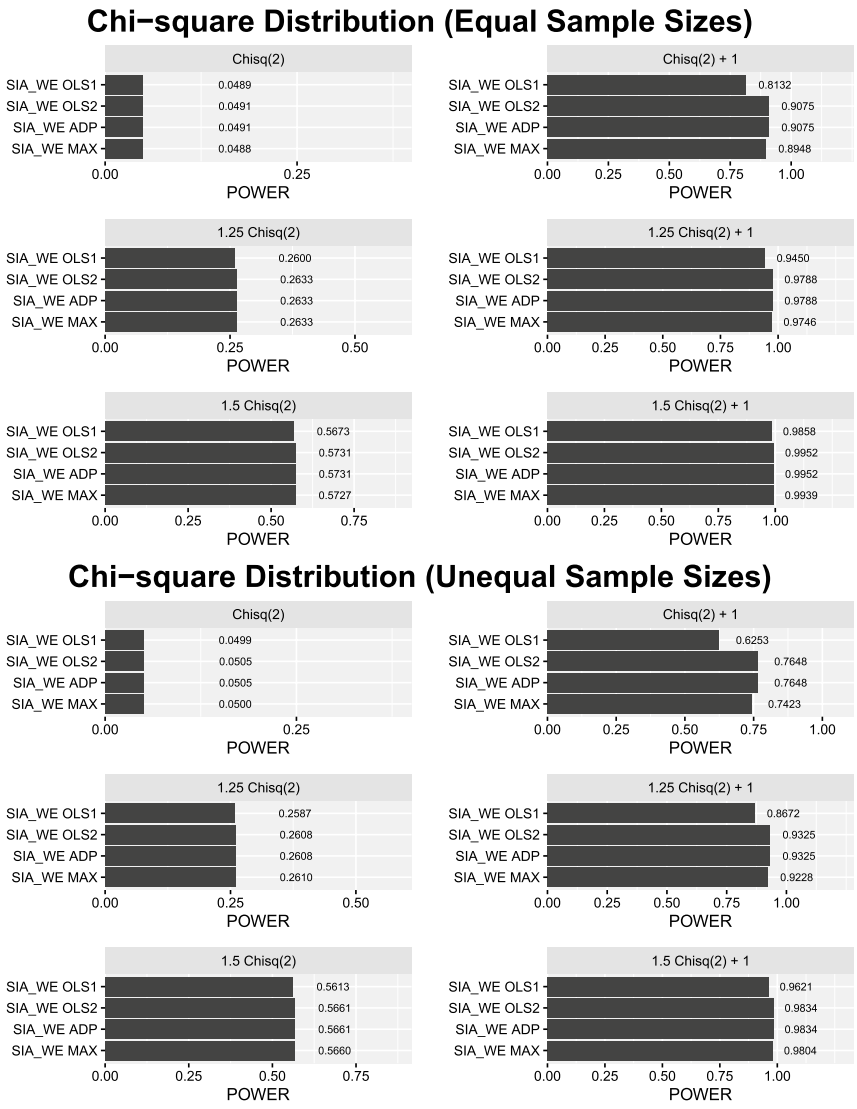

Fig. 5 Simulated type-I error and power for Chisq(2) vs. $\theta_2 * \text{Chisq}(2) + \theta_1$ when $\alpha = 0.05$

we consider 3 and 12 months as the two stages of the trial. This data set is publicly available in the R package "medicaldata" and named "polyps". The sizes of the first and second samples in the first stage are $(m_1, n_1) = (6, 7)$. Likewise, we obtain $(m_2, n_2) = (5, 7)$ in the second stage.

None of the investigated tests reject the null hypothesis at the 5% significance level (Table 1). However, when applying the 10% significance level, the tests lead to different decisions, that is, only SIA_UN and SIA_WE with OLS_2, ADP and MAX have p-values less than 10%. Moreover, the p-values based on

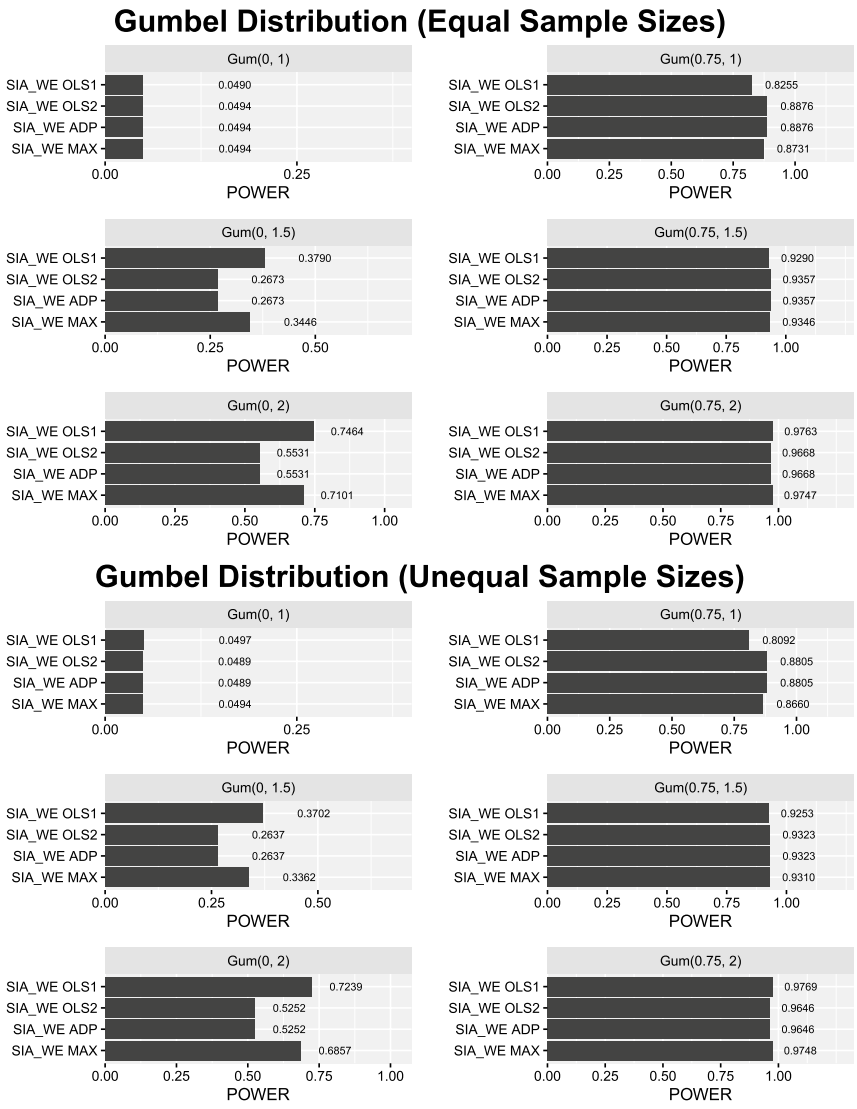

Fig. 6 Simulated type-I error and power for Gum(0, 1) vs. Gum(θ_1, θ_2) when $\alpha = 0.05$

SIA_WE are smaller compared to SIA_UN for OLS_2, ADP and MAX, respectively. It is consistent with the results of the simulation study, that the tests based on the weighted inverse normal combination function exhibit smaller p-values than those based on Fisher's combination function.

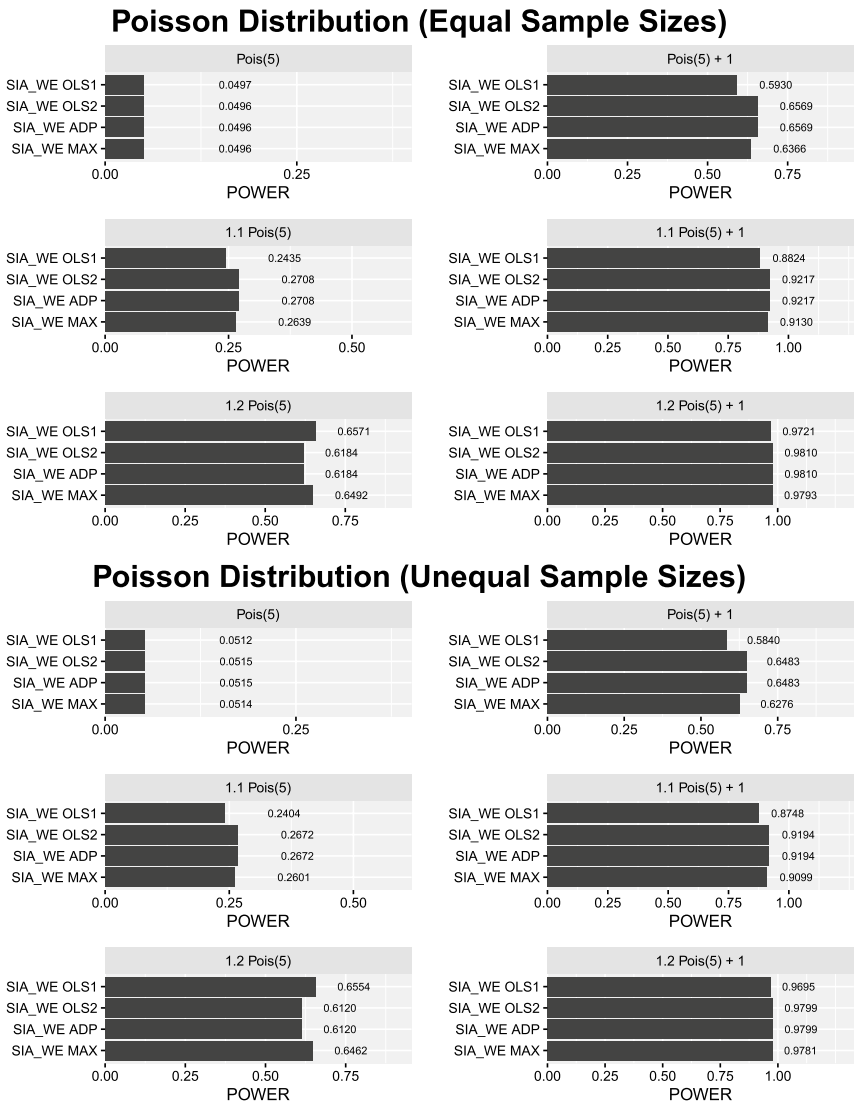

Fig. 7 Simulated type-I error and power for Pois(5) vs. $\theta_2 * \text{Pois}(5) + \theta_1$ when $\alpha = 0.05$

7 Concluding remarks

Adaptive designs allow for preplanned opportunities to modify the course of a trial based on accruing information. According to Bauer and Köhne (1994) and Lehmacher and Wassmer (1999), adaptive designs are highly flexible tools, and various modifications are possible after the interim analysis. An approach wherein the data from the first stage are utilized to weight a location-scale test was proposed by Neuhäuser (2001). However, situations exist wherein the test

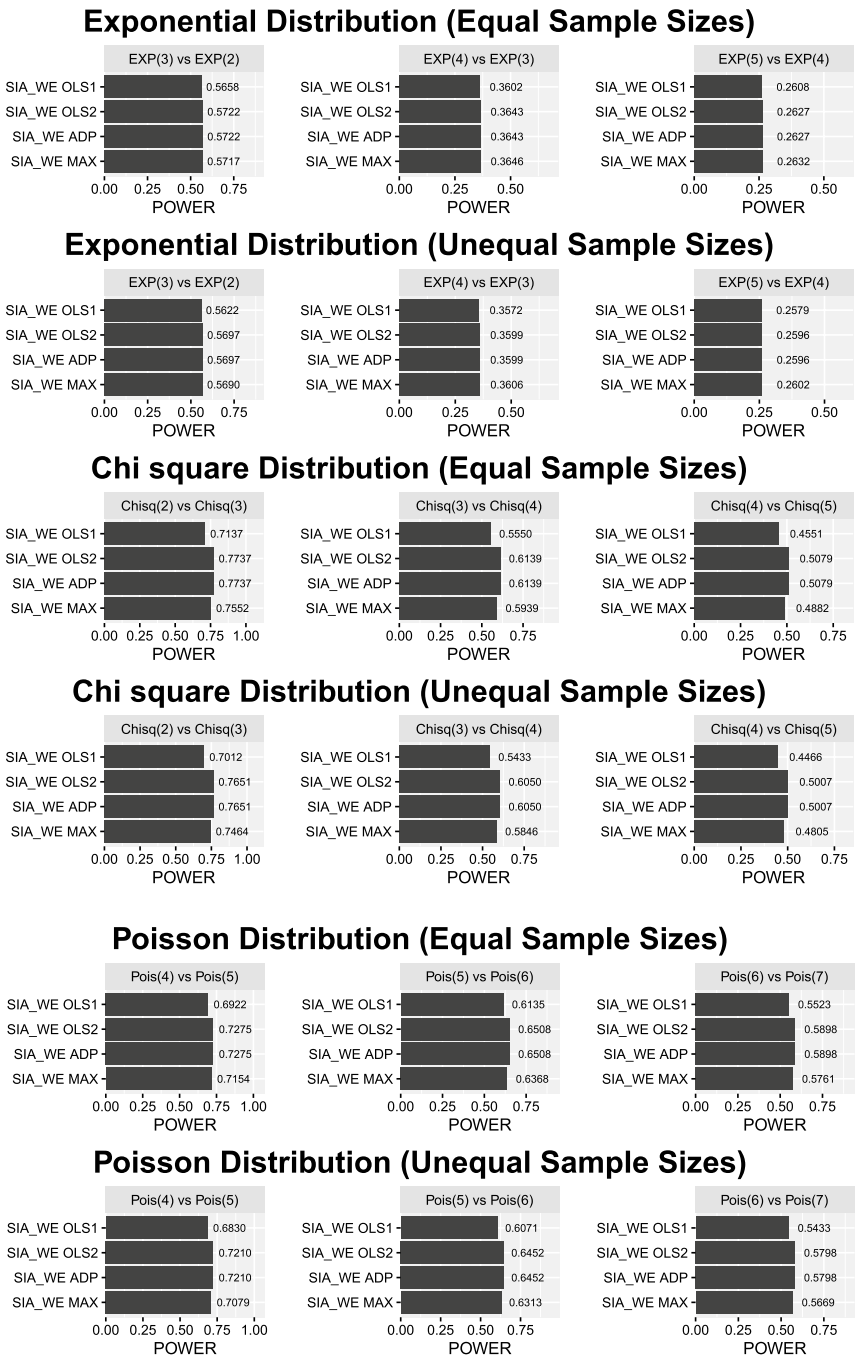

Fig. 8 Simulated power for (i) EXP(λ_1) vs EXP(λ_2) and (ii) Chisq(ν_1) vs Chisq(ν_2) when $\alpha = 0.05$. Continued for (iii) Pois(λ_1) vs Pois(λ_2) when $\alpha = 0.05$

Table 1 Number of colonic polyps. Test statistics and p-values of the considered tests

		OLS_1	OLS_2	ADP	MAX
NO_IA	Statistics	1.0108	1.1768	1.1768	1.1768
	P-values	0.1561	0.1196	0.1196	0.1377
FIA_UN	Statistics	0.0316	0.0254	0.0254	0.0278
	P-values	0.1408	0.1187	0.1187	0.1274
FIA_WE	Statistics	0.0342	0.0249	0.0249	0.0266
	P-values	0.1496	0.1169	0.1169	0.1231
SIA_UN	Statistics	0.1088	0.0823	0.0823	0.0904
	P-values	0.1088	0.0823	0.0823	0.0904
SIA_WE	Statistics	0.1148	0.0808	0.0808	0.0874
	P-values	0.1148	0.0808	0.0808	0.0874

with the weightings, $w_1 = |\text{STW}|/(|\text{STW}| + |\text{STAB}|)$ and $w_2 = 1 - w_1$, is less powerful than the unweighted test, that is $w_1 = w_2 = 1$, where STW and STAB are the standardized Wilcoxon rank-sum statistic and the standardized Ansari–Bradley statistic, respectively. In the present study, a new weighting is proposed that has, per our simulations, similar or greater power than the unweighted test in all scenarios considered in the simulation study. The adaptive test with the new weighting based on the weighted inverse normal combination function proposed by Lehmacher and Wassmer (1999) exhibits improved and stable power compared with the new weighting statistic based on Fisher's combination function. Indubitably, as the power is increased, one advantage of the proposed test could be the reduction in the average sample size required to obtain a pre-specified power.

We investigated the type-I error and power of the considered test statistics using Monte Carlo simulations for various distributions. Although the sample sizes were small to moderate, the investigated significance level of 5% was not breached. Additionally, the power was studied for continuous and discrete distributions. Noteworthily, discrete distributions were not investigated in Neuhäuser (2001). The simulation results indicated that the proposed statistic with the new weighting is a strong competitor to the existing one-sided Lepage statistics. Finally, we illustrated the tests using real-life data available in R.

In this study, we investigated statistical tests. Generally, estimation is also a significant concern in clinical trials. However, estimation in adaptive designs is more challenging. Point estimates computed using methods developed for classical fixed sample sizes may be biased. For the estimation and confidence intervals for adaptive designs, refer to Wassmer and Brannath (2016).

Acknowledgements The authors wish to express sincere thanks to the editor and an anonymous referee, whose comments improved the quality of the manuscript.

Funding Open Access funding provided by Tokyo University of Science. All authors have not received any funding for the research in this manuscript.

Data availability Familial adenomatous polyposis data set is publicly available in the R package "medicaldata" (Higgins 2022) and named "polyps".

Declarations

Conflict of interest All authors have declared no conflict of interest.

References

Bauer P, Köhne K (1994) Evaluation of experiments with adaptive interim analyses. Biometrics 50:1029–1041

Brunner E, Bathke AC, Konietschke F (2018) Rank and pseudo-rank procedures for independent observations in factorial designs. Springer, Cham, Switzerland

Büning H, Thadewald T (2000) An adaptive two-sample location-scale test of Lepage type for symmetric distributions. J Stat Comput Simul 65:287–310

Chernoff H, Savage I (1958) Asymptotic normality and efficiency of certain nonparametric test statistics. Ann Math Stat 29:972–994

Cucconi O (1968) Un nuovo test non parametrico per il confronto tra due gruppi campionari. Giornale degli Economisti e Annali di Economia 27:225–248

Dimairo M, Pallmann P, Wason J, Todd S, Jaki T, Julious SA, Mander AP, Weir CJ, Koenig F, Walton MK, et al. (2020) The adaptive designs consort extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. BMJ 369

Gibbons J, Chakraborti S (2021) Nonparametric Statistical Inference, 6th edn. CRC Press, Boca Raton, Florida

Higgins P (2022) Package 'medicaldata'. URL https://cran.r-project.org/web/packages/medicaldata/medicaldata.pdf, R package version 3.1

Hogg R, Fisher D, Randles R (1975) A two-sample adaptive distribution-free test. J Am Stat Assoc 70:656–661

Hogg R, Mckean J, Craig A (2018) Introduction to mathematical statistics, 8th edn. Pearson, New York

Hollander M, Wolfe D, Chicken E (2013) Nonparametric statistical methods, 3rd edn. John Wiley & Sons, Hoboken, New Jersey

Kössler W (2006) Asymptotic power and efficiency of lepage-type tests for the treatment of combined location-scale alternatives. Technical Report Humboldt-Universität zu Berlin

Kössler W (2010) Max-type rank tests, U-tests, and adaptive tests for the two-sample location problem—an asymptotic power study. Comput Stat Data Anal 54:2053–2065

Kössler W, Mukherjee A (2020) Distribution-free simultaneous tests for location-scale and Lehmann alternative in two-sample problem. Biom J 62:99–123

Lahmiri S (2023) A nonlinear analysis of cardiovascular diseases using multi-scale analysis and generalized hurst exponent. Healthcare Anal 3(100):142

Lehmacher W, Wassmer G (1999) Adaptive sample size calculations in group sequential trials. Biometrics 55:1286–1290

Lepage Y (1971) A combination of Wilcoxon's and Ansari-Bradley's statistics. Biometrika 58:213–217

Lepage Y (1975) Asymptotically optimum rank tests for contiguous location and scale alternatives. Commun Stat-Theory Methods 4:671–687

Marozzi M (2013) Adaptive choice of scale tests in flexible two-stage designs with applications in experimental ecology and clinical trials. J Appl Stat 40:747–762
Marozzi M (2013) Nonparametric simultaneous tests for location and scale testing: a comparison of several methods. Commun Stat - Simul Comput 42:1298–1317
Mukherjee A, Marozzi M (2019) A class of percentile modified lepage-type tests. Metrika 82:657–689
Mukherjee A, Kössler W, Murakami H (2021) Two new distribution-free two-sample tests for versatile alternative. Statistics 55:1123–1153
Murakami H (2007) Lepage type statistic based on the modified Baumgartner statistic. Comput Stat Data Anal 51:5061–5067
Murakami H (2016) A moment generating function of a combination of linear rank tests and its asymptotic efficiency. Test 25:674–691
Murakami H, Neuhäuser M (2024) A two-sample nonparametric test for one-sided location-scale alternative. J Appl Stat 1–29, https://doi.org/10.1080/02664763.2024.2392119
Neuhäuser M (2000) An exact two-sample test based on the Baumgartner-Weiss-Schindler statistic and a modification of Lepage's test. Commun Stat Theory Methods 29:67–78
Neuhäuser M (2001) An adaptive location-scale test. Biom J 43:809–819
Neuhäuser M (2012) Nonparametric statistical tests: a computational approach. CRC Press, Boca Raton, Florida
Nishino T, Murakami H (2019) The generalized Cucconi test statistic for the two-sample problem. J Korean Stat Soc 48:593–612
Omer DB, Las L, Ulanovsky N (2023) Contextual and pure time coding for self and other in the hippocampus. Nat Neurosci 26:285–294
Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR et al (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 16:1–15
Pettitt A (1976) A two-sample anderson-darling rank statistic. Biometrika 63:161–168
Randles R, Hogg R (1971) Certain uncorrelated and independent rank statistics. J Am Stat Assoc 66:569–574
Rublík F (2007) On the asymptotic efficiency of the multisample location-scale rank tests and their adjustment for ties. Kybernetika 43:279–306
Senn S (2016) Mastering variation: variance components and personalised medicine. Stat Med 35:966–977
Smit C, Swart N, Stoker D (1987) Weighted combinations of rank tests for location-scale alternatives. Commun Stat Theory Methods 16:3535–3553
Wassmer G, Brannath W (2016) Group sequential and confirmatory adaptive designs in clinical trials, vol 301. Springer, Cham, Switzerland
Yamaguchi H, Murakami H (2023) The multi-aspect tests in the presence of ties. Comput Stat Data Anal 180(107):680
Zar J (2010) Biostatistical analysis-5th, international edn. Pearson Education, Upper Saddle River, New Jersey