Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Adaptive Multiple Comparisons With the Best

Haoyu Chen[1,2,3] | Werner Brannath[4] | Andreas Futschik[3]

[1]Vetmeduni Vienna, Wien, Austria | [2]Vienna Graduate School of Population Genetics, Vienna, Austria | [3]Johannes Kepler University Linz, Linz, Austria | [4]Kompetenzzentrum fur Klinische Studien, Universität Bremen, Bremen, Germany

Correspondence: Andreas Futschik (andreas.futschik@jku.at)

Received: 30 August 2023 | **Revised:** 19 April 2024 | **Accepted:** 29 April 2024

Funding: This project was supported by the Austrian Science Fund (FWF; DK W1225-B20).

Keywords: adaptive subset selection | evolve and resequence | Gupta's rule | multiple comparison | multiple decision | *R*-values | Schweder–Spjøtvol estimator

ABSTRACT

Subset selection methods aim to choose a nonempty subset of populations including a best population with some prespecified probability. An example application involves location parameters that quantify yields in agriculture to select the best wheat variety. This is quite different from variable selection problems, for instance, in regression.

Unfortunately, subset selection methods can become very conservative when the parameter configuration is not least favorable. This will lead to a selection of many non-best populations, making the set of selected populations less informative. To solve this issue, we propose less conservative adaptive approaches based on estimating the number of best populations. We also discuss variants of our adaptive approaches that are applicable when the sample sizes and/or variances differ between populations. Using simulations, we show that our methods yield a desirable performance. As an illustration of potential gains, we apply them to two real datasets, one on the yield of wheat varieties and the other obtained via genome sequencing of repeated samples.

1 | Introduction

Choosing the best population(s) concerning an unknown parameter is of interest in a wide range of subject areas. In agriculture, for instance, farmers and corresponding researchers are often interested in choosing a subset of the most productive wheat brands from a broad variety of choices. In biological sciences, identifying a subset of genomic candidate positions in evolve and resequence experiments that contain the true target of selection can be interesting for researchers. It is also a problem in system designs such as inventory systems (Hsu and Nelson 1988) as well as clinical trials where some treatment arms need to be selected for further investigation.

Subset selection (Chang and Huang 2001) rules and multiple decision procedures (Bechhofer 1954) are methods to solve this problem. The proposed methods can, however, be quite conservative if the parameters of interest are not close to the so-called least favorable configuration (LFC). The LFC is the parameter configuration that minimizes the probability that a truly best population will be selected given some chosen procedure. Calibrating a method under the LFC can lead to the selection of many non-best populations. In this paper, we propose adaptive methods that make such methods less conservative.

Consider a set of $k > 1$ independent populations $\{\pi_1, \pi_2, \dots, \pi_k\}$ with corresponding sample data $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_k)$, for which we aim to find a subset of best. Depending on the situation, the distribution of the sample data and its parameters may vary as well as the definition of the best parameter values.

In this paper, we will focus on the scenario where the populations are normally distributed $\boldsymbol{X}_i \sim N(\theta_i, \sigma_i^2)$, and define "best" as the population(s) with the largest location parameter θ. Let $\boldsymbol{\Theta}$ be the

parameter space of θ, and let $\boldsymbol{\Theta}_i \in \boldsymbol{\Theta}$ be the subset of parameters such that population π_i is among the best populations. Typically, there exists at least one best population, hence $\boldsymbol{\Theta} = \cup_{i=1}^{k} \boldsymbol{\Theta}_i$. Some selection rule based on continuously distributed test statistics $(T_1, T_2, \dots, T_k)$ computed using sample data $(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_k)$ from each of the k populations is used. Assuming sample sizes and variances are equal between populations, Gupta's rule (Gupta and Panchapakesan 1972) is a common choice, with the test statistic under the unknown and equal variance case is

$$T_{\mathrm{G1}_i} := \frac{\sqrt{n}}{s}\left(\max_j \bar{X}_j - \bar{X}_i\right), \tag{1}$$

where s is the pooled sample standard deviation. A population is then determined as one of the best if its test statistic T_i is smaller than some predetermined critical value c_i. The critical value is selected such that the probability of correct selection (PCS) is controlled:

$$\mathrm{PCS}(\theta) := \mathbf{P}_\theta(T_i \le c_i) \ge 1 - \alpha, \quad \forall \theta \in \boldsymbol{\Theta}_i, \tag{2}$$

where α is the level of significance, which is analogous to the idea of the level of significance in suitably chosen hypothesis tests. The matching hypotheses to be tested can be found in Supplement Section S.2 (the prefix "S." refers to sections/figures in the Supporting Information). Unfortunately, the exact computation of the PCS requires the knowledge of the true underlying parameters. Therefore a so-called LFC that minimizes $\mathrm{PCS}(\theta)$ in θ needs to be identified in practice to ensure the control of PCS under all possible parameter values. Let $\boldsymbol{\Theta}_{\mathrm{LFC}}$ be the parameter set satisfying the LFC, and $F_\theta^{(i)}$ is the cumulative distribution function of T_i under parameter θ. Then we require that

$$F_{\theta_i}^{(i)}(x) \ge F_{\theta_{\mathrm{LFC}}}^{(i)}(x), \ \forall x \in \mathbb{R} \tag{3}$$

for any parameter vector $\theta_i \in \boldsymbol{\Theta}_i$ and $\theta_{\mathrm{LFC}} \in \boldsymbol{\Theta}_{\mathrm{LFC}}$ for all choices of i.

When using Gupta's rule, the LFC occurs when all populations are best, i.e. $\theta_1 = \theta_2 = \cdots = \theta_k$. Under the LFC, the distribution function of T_{G1_i} is

$$F_{\theta_{\mathrm{LFC}}}^{\mathrm{G1}}(x) = \int_0^\infty \int_{-\infty}^\infty \Phi^{k-1}(xu + y)\phi(y)g(u)dydu \tag{4}$$

where Φ is the distribution function of a standard normal, ϕ its density, and $g(u)$ is the density function of $\chi_{k(n-1)}/\sqrt{k(n-1)}$. Here χ_z is a chi distribution with z degrees of freedom, which denotes the distribution of the square root of a chi-square random variable. When the variance is known, the distribution function can be obtained by setting the degrees of freedom to infinite (see Supplement Section S.1).

Based on this, Hsu (1984) proposed a measure called R values that is analogous to p values as evidence of "Rejecting" a population:

$$R_i = 1 - F_{\theta_{\mathrm{LFC}}}^{(i)}(T_i).$$

A population π_i is then selected as best if $R_i \ge \alpha$. When the populations have an equal sample size and variance, Gupta's rule is analogous to Dunett's test in multiple-hypothesis testing. Gupta's test statistic can be seen as a transformation of the maximum test statistic in Dunnett's test. More details on this can be found in Supplement Section S.2.

Although this would be possible, usually there is no further multiple-testing correction applied to the set of R values. Applying for instance a familywise-error correction such as Bonferroni–Holm on the R_i, would additionally control the probability of selecting all best populations in case there are multiple best populations. However, it would make the procedure also much more conservative, especially with many best populations.

To be informative, selected subsets should contain as few non-optimal populations as possible. Controlling the PCS under the LFC can however lead to conservative results (Hsu 1984), resulting in a much larger than desired PCS under the true parameter value and several non-best populations in the selected subset, especially if the true parameters are far away from the LFC.

A similar problem also exists in the field of multiple hypothesis testing, where methods such as the Bonferroni procedure are used to control the family-wise error rate (FWER). Such procedures can also become conservative unless all null hypotheses are true. To solve this issue, Langaas, Lindqvist, and Ferkingstad (2005) have proposed approximate procedures based on the estimate of true nulls, given that the number of hypotheses is large. Finner and Gontscharuk (2009) further derived theoretical properties of this plug-in estimate concerning the control of FWER. Alternative ways of true null estimations have also been proposed by Storey (2002) and Schweder and Spjøtvoll (1982), given that the marginal p values corresponding to the true nulls are uniformly distributed. Further, Hoang and Dickhaus (2020) have suggested the usage of randomized p values to remedy the above uniform assumption, given that the p values are independent.

Similar approaches have not been considered in the context of subset selection. Under the LFC, the marginal R values from the best populations are also uniformly distributed, like p values under the null hypothesis. The conservativeness of subset selection methods increases with population size and distance of parameters from LFC. Considering the connection between Gupta's rule and Dunett's test mentioned previously, we reduce this conservatism through a similar adaptive approach to multiple hypothesis testing. Given that the number of populations k is large, we propose an estimate $\hat{K}$ on the number of best populations. Note that this estimate can also be influenced by the population sizes and variances as well as the number of populations that are close to the best. Based on this estimate, adaptive R values are obtained to decide whether a population should be selected. An alternative estimate based on the randomization method of Hoang and Dickhaus (2020) is also explored.

We also investigate scenarios when the sample sizes differ between populations. This can cause issues in subset selection as it makes the distribution function of the test statistic under LFC no longer exactly computable. Gupta has suggested a few ways (Gupta and Huang 1974, 1976), both of which are uniformly more conservative. We introduce instead a simpler alternative approach based on Tukey's conjecture (Hayter 1984) that does not require a change in the distribution function while also being less conservative. When variances between populations are also

unequal, the Behrens–Fisher problem (Kim and Cohen 1998) will arise. We will use Welch's (Best and Rayner 1987) approximate solution to solve this, but other approaches such as the one by Patil (1969) also exist.

Besides the selection of populations, there has been additional work in the area. Such work has been summarized in Finner et al. (2021), where, for instance, the computation of confidence sets for the difference to the best population(s) in terms of the parameter value is discussed. Partition testing as well as approaches based on decision paths is also explained. These methods are able to provide additional insight into the ordering of populations in terms of their parameters.

Using simulation results, we show that our method has desirable performance under a wide range of parameter values. It also has a very broad field of applications. As examples, we apply it to find a subset of the most productive wheat variety using data from Kansas State University (2022) as well as a subset that includes the true selective target in evolve and resequencing studies (Barghi et al. 2019). In both cases, our proposed adaptive approach gives satisfying results. Our proposed methods are implemented in an R package (R Core Team 2021) available at https://github.com/xthchen/adass.

2 | Adaptive Selection Methods

Ideally, R values, R_i, should follow a uniform distribution if population π_i is best. Analogous to p values in hypothesis testing, the PCS would then be equal to the desired value:

$$\text{PCS}(\theta) = \mathbf{P}_\theta(R_i > \alpha) = 1 - \alpha, \quad 0 \le \alpha \le 1. \tag{5}$$

While this is true if $\theta \in \mathbf{\Theta}_{\text{LFC}}$, for other sets of parameter values, we obtain a higher than desired PCS, and thereby a too large set of best populations. Let $\theta \in \mathbf{\Theta} \cap \mathbf{\Theta}^c_{\text{LFC}}$ and π_i be the best population, we have

$$\begin{aligned}\text{PCS} = \mathbf{P}_\theta(R_i > \alpha) &= \mathbf{P}_\theta(1 - F_{\theta_{\text{LFC}}}(T_i) > \alpha) \\ &= \mathbf{P}_\theta\left(T_i < \left[F_{\theta_{\text{LFC}}}\right]^{-1}(1-\alpha)\right) \\ &= F^{(i)}_{\theta_i}\left(\left[F_{\theta_{\text{LFC}}}\right]^{-1}(1-\alpha)\right) \\ &> 1 - \alpha, \end{aligned} \tag{6}$$

where the last inequality is true due to the definition of the LFC in (3). Therefore if θ is very far away from the LFC, PCS can be considerably above $1 - \alpha$.

To obtain R values that are closer to the uniform distribution on the set $\mathbf{\Theta}_i$, we propose an alternative way of computing R values. Our approach uses the estimated number of best populations $\hat{K}$. We will denote $\hat{K}$ as the "effective number" of populations. The selection is then reduced to a smaller subset containing only $\hat{K}$ populations instead of k, resulting in a smaller multiplicity correction. Notice that the parameter space changes, when the number of populations decreases. Nevertheless, LFC still occurs when all populations are best, just with a smaller total number of populations. Hence using the equal and unknown variance case as an example, the distribution function of T_{G1_i} under LFC now becomes

$$F^{\text{G1}}_{\hat{\theta}_{\text{LFC}}}(x) = \int_0^\infty \int_{-\infty}^\infty \Phi^{\hat{K}-1}(xu + y)\phi(y)\hat{g}(u)dydu,$$

where $\hat{g}(u)$ is the density of $\chi_{\hat{K}(n-1)}/\sqrt{\hat{K}(n-1)}$. This will reduce conservativeness both due to the reduction in multiple testing as well as the true parameter becoming closer to its corresponding LFC after the removal of $k - \hat{K}$ presumably non-best populations. We call the R values produced using the above distribution function adaptive R values.

2.1 | Adaptive Selection

The Schweder–Spjøtvoll estimator (Schweder and Spjøtvoll 1982) has been commonly used in the field of multiple hypothesis testing to estimate the proportion of true nulls. In the context of subset selection, this is analogous to estimating the proportion of best populations β. The Schweder–Spjøtvoll estimator is computed as follows:

$$\hat{\beta} = \frac{1 - G(\lambda)}{1 - \lambda},$$

where $G(\cdot)$ is the empirical cumulative distribution function of the p values, $\lambda \in [0, 1)$ a tuning parameter, and $\hat{\beta}$ the estimated proportion of true nulls. When applying it in the context of subset selection and R values, we rewrite it to the following form:

$$\hat{\beta}_\lambda = \frac{|\{R_j > \lambda\}|}{k(1-\lambda)}, \tag{7}$$

where $|\cdot|$ is the cardinality of the set. We then set our estimate on the number of effective populations as

$$\hat{K}_\lambda = \max[k\hat{\beta}_\lambda, 2].$$

Here we are setting the estimate to be at least two in order to have at least one comparison in between.

In Storey (2002), a slight modification of the previously shown estimator has been suggested which we adopt here for the population selection task. Given the same tuning parameter λ, we can rewrite the modified estimator as

$$\hat{\beta}_{\lambda^{\text{mod}}} = \frac{|\{R_j > \lambda\}| + 1}{k(1-\lambda)}. \tag{8}$$

Here the additional +1 in the numerator artificially inflates the estimated effective number of populations. This will give a more conservative result, but the impact will be small when the number of populations is large. The approach can therefore help to control PCS under LFC, especially when the number of populations is small. The theoretical properties of our proposed estimator can be found in Supplement Section S.3, which shows that our estimator is asymptotically unbiased and consistent given that some assumptions are met.

In theory, the adaptive selection (AS) method can be applied iteratively to reduce the effective number further. However, from

our simulations, this approach can become too anticonservative under the LFC, with limited improvement compared to the AS method when far from the LFC.

2.2 | Randomized Adaptive Selection

The Schweder–Spjøtvoll estimator is originally derived under the assumption that the p values from the true nulls are uniformly distributed. In the case of subset selection, as shown in (6), this is only true when we are under LFC and the joint distribution of the test statistics is continuous. Lemma 1 in Supplement Section S.3 shows that when we are not under the LFC, the estimator will have a positive mean bias, resulting in a more conservative outcome.

To solve this issue, Hoang and Dickhaus (2020) proposed an approach of transforming p values through randomization such that the distribution of p values from the true nulls becomes closer to uniform. We apply this to R values in a similar spirit:

$$R_j^{\text{rand}} = U_j \mathbf{1}\{R_j > c\} + \frac{R_j}{c}\mathbf{1}\{R_j \le c\}, \tag{9}$$

where U_j is drawn from a standard uniform distribution $U[0,1]$ and $\mathbf{1}\{\cdot\}$ is the indicator function. The choice of parameter $c \in [0,1]$ influences the distribution of R_j^{rand} greatly. Extreme parameter choices of $c = 0$ and $c = 1$ lead to $R^{\text{rand}} = U_j$ and $R_j^{\text{rand}} = R_j$, respectively. Since neither choice is optimal, a trade-off point needs to be found such that the R values from the best populations are close to uniformly distributed, while maintaining small R values for the non-best. When using R^{rand} instead of R, our estimator for β becomes

$$\hat{\beta}_r = \frac{|\{R_j^{\text{rand}} > \lambda\}|}{k(1-\lambda)} \tag{10}$$

with λ the same tuning parameter with the same restrictions as in (7). For some fixed λ, the additional tuning parameter c is chosen such that the mean bias of $\hat{\beta}_r$, $\mathbf{E}_\theta[\hat{\beta}_r] - \beta$, is minimized while remaining positive. We show in Supplement Section S.4 that the following choice of c provides an approximate solution of this criterion:

$$c = \underset{c}{\text{argmax}} \sum_{j=1}^{k} \left[\lambda \mathbf{1}(R_j > c) + \mathbf{1}\left(\frac{R_j}{c} \le \lambda\right)\mathbf{1}(R_j \le c)\right].$$

The estimated effective number of populations is then

$$\hat{K}_r = \max[k\hat{\beta}_r, 2].$$

2.3 | Tuning Parameter λ

It has been shown in Schweder and Spjøtvoll (1982) that the choice of λ has a large impact on the estimator of both methods mentioned above. A large λ will result in a large $\text{Var}(\hat{K})$, while smaller choices of λ increase bias. Under our subset selection setup, we found that larger choices of λ generally give more anticonservative results, resulting in uncontrolled PCS when the populations are closer to the LFC (see Section 4.1). To tune this parameter, a few existing methods have been summarized by Langaas, Lindqvist, and Ferkingstad (2005). We will outline one of the viable methods we employ below.

Since the choice of λ is a trade-off between bias and variance, we aim to minimize the mean square error (MSE). A choice of λ (Storey 2002) that minimizes the MSE of the estimate of K satisfies

$$\underset{\lambda}{\text{argmin}}\, \text{MSE}(\hat{K}(\lambda)) = \underset{\lambda}{\text{argmin}}\, \mathbf{E}\left[\left(\hat{K}(\lambda) - K\right)^2\right],$$

where K is the true number of best populations. This can be estimated using a bootstrap approach (Efron and Tibshirani 1993):

$$\widehat{\text{MSE}}(\hat{K}(\lambda)) = \frac{1}{S}\sum_{i=1}^{S}(\hat{K}^{*i}(\lambda) - K)^2,$$

where the S resampling estimates $\hat{K}^{*i}(\lambda)$ are produced through the bootstrapping of R values. The true number of best populations K is however unknown. Given that the assumptions in Supplement Section S.3 are met, Lemma 1 in the same section shows that

$$\mathbf{E}_\theta[\hat{K}(\lambda)] \ge \min_{\lambda' \in (0,1)} \hat{K}(\lambda') \ge K.$$

Hence, we can use $\min_{\lambda' \in (0,1)} \hat{K}(\lambda')$ as a plug-in estimate for K. The estimated MSE is computed as:

$$\widehat{\text{MSE}}(\hat{K}(\lambda)) = \frac{1}{S}\sum_{i=1}^{S}\left(\hat{K}^{*i}(\lambda) - \min_{\lambda' \in (0,1)} \hat{K}(\lambda')\right)^2. \tag{11}$$

This method of tuning λ can be used in both the AS and randomized adaptive selection (RAS) methods. The approach requires the assumptions outlined in Supplement Section S.3 to be met and therefore requires the number of populations to be large. In such a situation, the tuning process may become computationally intensive, especially when applied to the RAS method due to the need to tune the additional parameter c.

3 | Populations With Unequal Sample Sizes/Variances

Previously, we limited ourselves to scenarios where all populations have equal sample sizes and a common true variance. Modified subset selection methods have been proposed in the literature when this is not the case. (See, for instance, Chen, Dudewicz, and Lee 1976 or chapter 12 in Gupta and Panchapakesan 2002 for a summary of selection rules in such a situation. See also Supplement Sections S.6 and S.7 for alternative selection rules and multiple hypothesis testing approaches such as Dunnett's T3 test.)

It has been proven by Hayter (1984), that the Tukey–Kramer procedure is uniformly more conservative when the sample sizes or variances of the populations are unequal. More specifically, Hayter (1984) proved the following:

Theorem 1. *For independently distributed* $Y_i \sim N(0, \sigma_i^2)$, $0 \leq \sigma_i \leq \infty$ *and some fixed* $q > 0$, *the following probability is strictly minimized for all* i, j *when* $\sigma_i = \sigma_j$.

$$\mathbf{P}\left[|Y_j - Y_i| \leq q\sqrt{\sigma_j^2 + \sigma_i^2}, \quad \forall i \text{ and } j\right].$$

As an alternative to the unequal sample-size methods proposed above, we explore whether the above result on Tukey's conjecture carries over to subset selection. More specifically, we investigate via simulations whether the LFC in the equal sample size (equal variance) scenario (4) provides a conservative null distribution in the unequal variance case.

For unequal sample sizes, we explore the following test statistics together with the null distribution (4):

$$T_{\mathrm{TK1}_i} := \frac{1}{s\sqrt{\frac{1}{2}\left(\frac{1}{n_j} + \frac{1}{n_i}\right)}}\left(\bar{X}_j - \bar{X}_i\right), \quad j = \underset{m}{\operatorname{argmax}}\, \bar{X}_m, \tag{12}$$

where n_i denoted the sample size of population i and the harmonic mean of the two population sample sizes is used. Note that this test statistic is identical to the equal sample size case shown in (1) if the sample sizes are equal ($n_i = n_j$). Our simulations in Section 4.4 suggest that the results by Hayter (1984) carry over to our situation.

An alternative test statistic we can employ when sample sizes are unequal is the following:

$$T_{\mathrm{TK2}_i} := \max_j \left(\frac{1}{s\sqrt{\frac{1}{2}\left(\frac{1}{n_j} + \frac{1}{n_i}\right)}}\left(\bar{X}_j - \bar{X}_i\right) \right). \tag{13}$$

The population(s) that maximizes this expression is not necessarily the one with the largest mean. This test statistic is less conservative than (12). According to Supplement Section S.12, there are minor violations in PCS under the LFC for the nonadaptive method. For the adaptive methods, we observe slight violations even when far from the LFC. We will therefore focus on (12) in the main paper.

For the scenario where both sample sizes and variances are different between populations, and where the sampled variance of population i is s_i^2, we use the following test statistic, while maintaining the same distribution function shown in (4):

$$T_{\mathrm{TK3}_i} := \frac{1}{\sqrt{\frac{1}{2}\left(\frac{s_{j'}^2}{n_{j'}} + \frac{s_i^2}{n_i}\right)}}\left(\bar{X}_{j'} - \bar{X}_i\right), \quad j' = \arg\max_j \bar{X}_j \tag{14}$$

Note that due to the Behrens–Fisher problem (Kim and Cohen 1998), the degrees of freedom need to be approximated. We employ the Welch's approximate solution (Best and Rayner 1987):

$$\nu_{i,j} = \frac{(s_j^2/n_j + s_i^2/n_i)^2}{s_j^4/n_j^2(n_j - 1) + s_i^4/n_i^2(n_i - 1)} \tag{15}$$

for this purpose. Similar to (13), we can use the maximum test statistic to reduce conservativeness. However, as mentioned previously, this can lead to an uncontrolled PCS even for the nonadaptive method.

To obtain asymptotic unbiasedness, we allow the effective number of populations to be larger than the total number of populations in Supplement Section S.3. When the sample sizes and/or variances are unequal between populations, this increase in population size causes problems with the degrees of freedom computation. For conservativeness, we propose to increase the degrees of freedom as if the additional populations have the smallest population size and largest variance among all populations. Our simulation in Supplement Figure S1 shows that the changes in degrees of freedom have little impact on the results.

4 | Simulation Examples

To illustrate the performance of our proposed adaptive method, we carry out a simulation study. We will employ the two adaptive methods suggested in Section 2, namely AS and RAS, and compare them to the nonadaptive selection (NAS) case. Unless otherwise mentioned, we simulate 100,000 iterations per scenario. We assume 100 populations, with the best populations having a mean of 0 and the non-best populations having a mean of −2. The aim is to select a subset of normally distributed populations that have the largest true location parameter. We will only consider scenarios with a PCS of higher than 0.5 since a lower value is usually not desirable. In Sections 4.1–4.3, we consider the case of a known common variance. Scenarios with unknown variances, as well as scenarios with unequal sample size and/or variance, can be found in Section 4.4.

4.1 | Probability of Correct Selection

We obtain the distribution of PCS for best populations by plotting the desired level of significance α against $\mathbf{P}(R_i \leq \alpha | \theta \in \mathbf{\Theta}_i)$. Ideally, as mentioned in Section 2, the R values from the best populations should be uniformly distributed, corresponding to the diagonal black line in the plots shown below. If the distribution function is below the diagonal line, we are too conservative, obtaining a higher-than-desired PCS.

Two scenarios are considered in Figure 1. The left panel shows the case where we are far from the LFC, with only 10% of the population being best. The right panel shows the cases where we are under the LFC. In the left panel, we see that all adaptive methods reduce conservativeness considerably compared to the basic subset selection method. With the same tuning parameter, randomizing the R values will further reduce conservativeness slightly. In the right panel, the basic subset selection method is exact, with both variants of adaptive methods having some

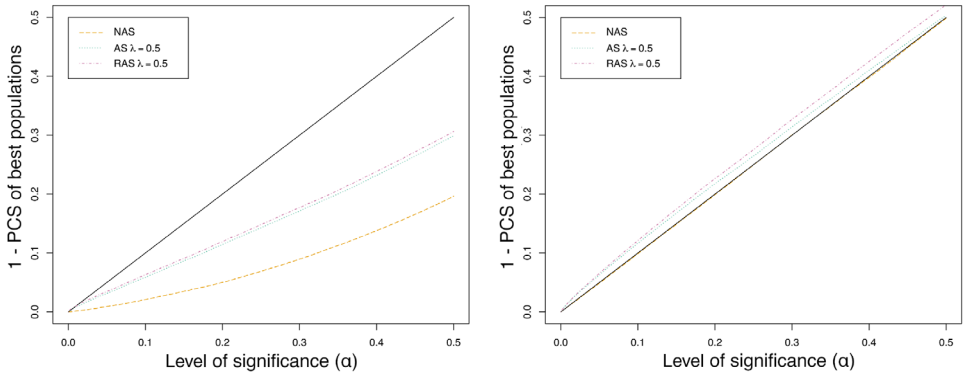

FIGURE 1 | Results showing 1 - PCS of best populations against the level of significance α. Both scenarios contain 100 populations each of sample size 1 and known unit variance. In the left panel, 10 populations have a mean of 0, with the other having a mean of −2. In the right panel, all populations have a mean of 0.

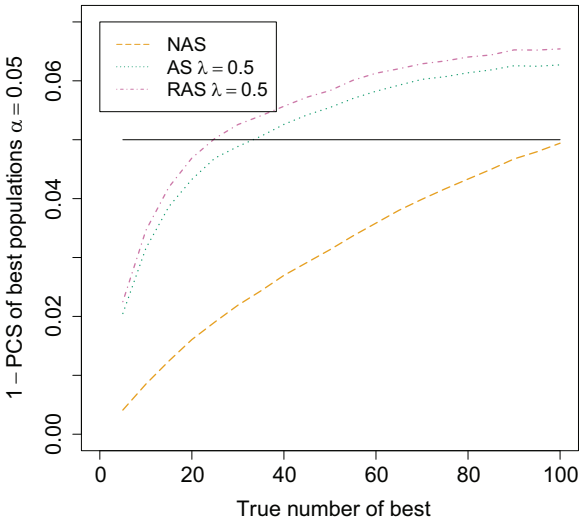

FIGURE 2 | Results showing 1 - PCS of best populations against the true number of best with $\alpha = 0.05$. The scenario contains 100 populations each of sample size 1 and known unit variance. We set the best populations to have a mean of 0, while the non-best have a mean of −2. Here the level of significance is fixed to 0.05, while we change the percentage of best populations.

minor inflation in the PCS, with the inflation being slightly more prominent for the RAS method.

The number of populations is rather large in Figure 1. In Supplement Section S.8, we show a scenario where there are only 10 populations with the other parameters remaining identical. In this setting, adaptive methods perform significantly worse.

To further explore the behavior of adaptive methods, we show in Figure 2 the effect of changing the percentage of best populations. Here we fixed the level of significance α to 0.05, shown as a black horizontal line, and varied the percentage of best between 5% and 100%. Note that 100% best would denote the LFC, and any value considerably above the black line indicates inflation in PCS. As shown in the figure, all methods show reduced conservativeness when approaching LFC. The nonadaptive method controls PCS uniformly as expected, while all adaptive methods show inflation between 0.01 and 0.02 when close to the LFC.

4.2 | Power and Subset Size

One would expect less conservative methods to have higher power when far from the LFC. Figure 3 confirms this for our proposed adaptive methods. Furthermore, the power of the RAS method is slightly above that of the AS method, in line with its smaller PCS shown in Figures 1 and 2. Here, we borrow the phrase "power" from multiple hypothesis testing and define it as the probability of an R value of a given non-best population being smaller than some chosen level of significance α.

In Supplement Section S.9, we illustrate by how much this gain in power reduces the size of the selected subset by reducing the number of included non-best populations.

4.3 | Tuning Parameter λ

In Sections 4.1 and 4.2, we considered two variants of our proposed adaptive method and showed their potential to significantly reduce conservativeness and increase power. However, the extent of this improvement for both methods is highly dependent on the tuning parameter λ. We show the effect of changing the tuning parameter λ in Supplement Section S.10.

So far, we used a uniform choice of $\lambda = 0.5$ for noniterative and $\lambda = 0.2$ for iterative methods. In our simulations, this choice provided a good trade-off between conservativeness reduction and PCS control. However, the optimal choice of λ depends on quantities such as the number of populations and the proportion of best. It can therefore be beneficial to apply the automatic tuning method outlined in Section 2.3 despite the increase in computational cost.

In Figure 4, we compare the performance of automatic tuning against a simple choice of $\lambda = 0.5$ when using the RAS method. When we are far from the LFC (left panel), the automatic tuning method is less conservative. On the other hand, the PCS is not

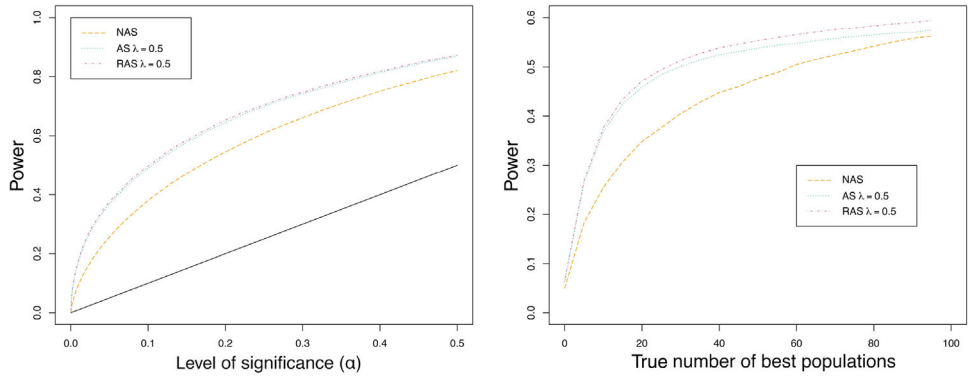

FIGURE 3 | Power analysis. Both scenarios contain 100 populations each of sample size 1 and known unit variance. In the left panel, 10 populations have a mean of 0, with the other having a mean of −2. In the right panel, the level of significance is fixed at 0.05 while we change the percentage of best populations.

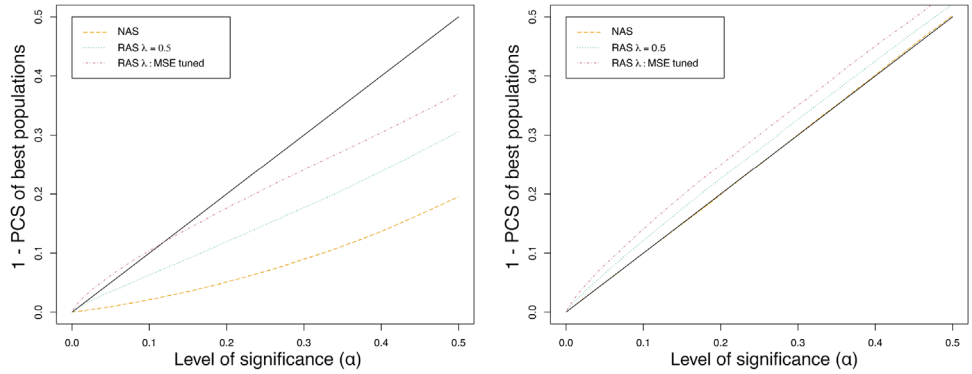

FIGURE 4 | Comparing automatic λ tuning versus $\lambda = 0.5$. Both scenarios contain 100 populations each of sample size 1 and known unit variance. In the left panel, 10 populations have a mean of 0 with the other having a mean of −2. In the right panel, all populations have a mean of 0. The legend "MSE tuned" denotes automatic tuning as described in Section 2.3.

as tightly controlled under the LFC (right panel). Supplement Section S.11 provides results for the AS method.

4.4 | Unequal Sample Size and/or Variance

In this section, we consider scenarios with unequal sample sizes and variances. We also compare our proposed test statistics (12)–(14) with older proposals by Gupta and Huang (1976) outlined in Supplement Section S.6.

In Figure 5, we randomly choose the sample size for each population between 5 and 10 and assume a known variance of one. Among the nonadaptive methods, our proposed approach (12) performs better than Gupta's modification. The distribution of the R values from both adaptive methods is even closer to uniform, regardless of whether we are far or close to the LFC.

For a scenario with unequal variances, we additionally draw the variance of each population from $U[0.5, 1.5]$. As shown in Figure 6, we are more conservative when far from LFC compared to the known variance case. The difference between adaptive and nonadaptive scenarios remains large, however. The PCS is controlled even under the LFC.

5 | Real-World Applications

In this section, we apply our proposed methods to real-world datasets. Our first example deals with finding the most productive variety of wheat in terms of yield. The second is from population genetics. There are other potential applications such as multiarm clinical trials, where our proposed methods might be used to determine which arm to drop during intermediate stages.

5.1 | Application in Choosing Wheat Variety

It is often challenging for farmers to choose a crop variety with maximal yield and profit. Given some control in variables such as environmental conditions and soil type, crop yields from the same variety can often be seen as normally distributed. After adjusting for other covariates, subset selection can be used to eliminate inferior variants. Compared to classical subset selection methods, our proposed adaptive rules can be expected to be less conservative.

Here we will apply our method to the winter wheat varieties, using datasets provided by the Kansas State University Agricultural Experiment Station and Cooperative Extension Service (Kansas State University 2022). Annually, yields of wheat from

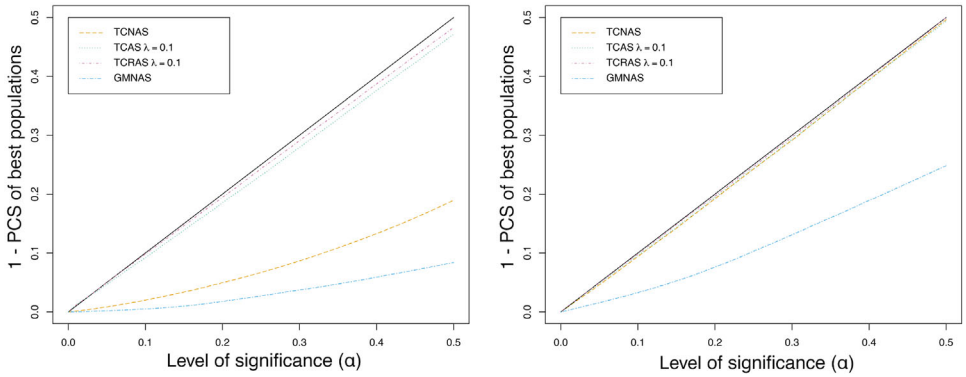

FIGURE 5 | Results showing 1 - PCS of best populations against the level of significance, when sample sizes are unequal. Both scenarios contain 100 populations each with a sample size randomly chosen between 5 and 10 and known unit variance. In the left panel, 10 populations have a mean of 0 with the other having a mean of −2. In the right panel, all populations have a mean of 0. The prefix "TC" denotes the method outlined in (12) of Section 3, and "GM" represents the method outlined in Supplement Section S.6.

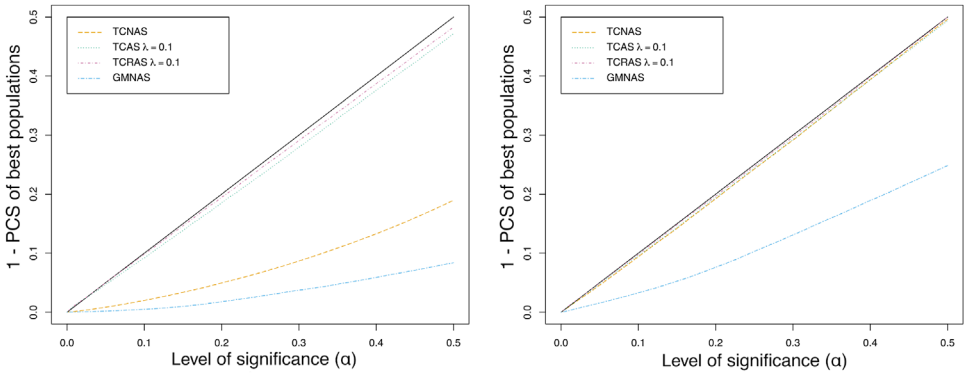

FIGURE 6 | Results showing 1 - PCS of best populations against the level of significance, when both sample size and variance are unequal. Both scenarios contain 100 populations each with a sample size randomly chosen between 5 and 10, and variance drawn from $U[0.5, 1.5]$. In the left panel, 10 populations have a mean of 0 with the other having a mean of −2. In the right panel, all populations have a mean of 0. The prefix "TC" denotes the method outlined in (12) of Section 3, and "GM" represents the method outlined in Supplement Section S.6.

different varieties planted at various farms and fields across the state of Kansas are measured in terms of the number of bushels per acre. To control for types of irrigation and season, we will focus our approach on dryland winter wheat only.

Yields reported by different farms for the same variety can then be seen as observations from the same population. However, since Kansas is a large state, environmental factors can still vary and affect the productivity of wheat. Additionally, farms may also employ different strategies when growing wheat, making yields from different farms less comparable. Since multiple wheat varieties are present at all farms (with the minimum number of varieties being 13), we subtract from each of the yields the mean farm yield. Further, since many varieties have been cultivated at different numbers of farms, this is an unequal sample size scenario.

We use the 2022 data to apply our methods. To reduce the effect of any potential outliers, we only use wheat varieties planted at more than six farms. After this filtering, 47 wheat varieties remain that were planted across 15 farms. Assuming equal variance across different varieties, we obtained a subset of best containing 11 varieties when aiming for a PCS of 95%. Using the AS method, we were able to further reduce the subset size to 7. When the equal variance assumption is dropped, we obtain slightly larger subsets, but again a smaller one with the AS method (12 vs. 10). A reason for dropping the equal variance assumption is that the land area each variety of wheat occupies at each farm is unknown. Since yield is measured in terms of the number of bushels per acre, wheat varieties with larger land areas will have a smaller yield variance. As there are multiple observations from different farms per variety, these differences may cancel out to some extent, however.

Details on the varieties that have been selected as potentially best by the different methods can be found in Table 1. This suggests that our proposed adaptive method provides less conservative subsets when applied to crop yields.

In theory, it is also possible to choose one farm and use data from different years as samples for each wheat variety instead. This reduces the impact of location and farming strategies of different farms, but the impact of weather varying across years needs to be taken care of. Also, many varieties of wheat are not consistently planted on the same farm across multiple years, possibly due to the introduction of new varieties or low productivity.

TABLE 1 | Wheat varieties that are best under different approaches.

Equal variety	Equal variety AS	Unequal variety	Unequal variety AS
KIVARI_AX	KIVARI_AX	KIVARI_AX	KIVARI_AX
WB4401	WB4401	WB4401	WB4401
AP18_AX	AP18_AX	AP18_AX	AP18_AX
GUARDIAN	GUARDIAN	GUARDIAN	GUARDIAN
CP7266AX	CP7266AX	CP7266AX	CP7266AX
STRAD_CL_PLUS	STRAD_CL_PLUS	STRAD_CL_PLUS	STRAD_CL_PLUS
AP_EVERROCK	AP_EVERROCK	AP_EVERROCK	AP_EVERROCK
CRESCENT_AX		CRESCENT_AX	CRESCENT_AX
LCS_RUNNER		LCS_RUNNER	LCS_RUNNER
BOB_DOLE		BOB_DOLE	BOB_DOLE
LCS_VALIANT		LCS_VALIANT	
		WB4595	

5.2 | Application in Evolve and Resequence Experiments

Evolve and resequence experiments (Turner et al. 2011) are used to study adaptation over time under a controlled environment. Genetic variants that are beneficial under such an environment tend to become more abundant over time. Besides such a systematic change, there are also random fluctuations in allele frequencies that depend on the population size and are known as "genetic drift." Using high-throughput whole genome sequencing techniques, researchers can obtain allele frequency estimates at chosen time points during the experiment from multiple independently evolving laboratory populations. Potential additional sources of variation are sampling and sequencing noise. We ignore them here for simplicity.

Without the presence of selection, the true allele frequencies in subsequent generations are modeled for a given single nucleotide polymorphism (SNP) via binomial sampling from the previous generation (Neuhauser 2004):

$$f_i^{t+1} \sim \text{Binomial}(N_e, f_i^t)/N_e$$

with N_e being the effective population size. Under selection with selection coefficient s, binomial sampling is carried out under

$$f_i^{t,*} = \frac{(1+s)f_i^t}{1+sf_i^t} \tag{16}$$

instead of f_i^t. For small s and f_i^t, we have $f_i^{t,*}/f_i^t \approx 1+s$, that is, a relative increase of roughly s. Frequency changes at a genetic position that are larger than expected from this random process and additional noise, are seen as evidence of selection. The adapted chi-square or Cochran-Mantel-Haenszel (CMH) test (Spitzer, Pelizzola, and Futschik 2020) is often used to identify such signals. Due to their spatial vicinity, SNPs close to a position with a causal effect may also be affected by selection, an effect called genetic hitchhiking (Barton 2000). Depending on the distance from the selected locus, such positions often also lead to significant test results, although they do not provide adaptive benefits themselves.

One strategy to search for the causal SNP is to select a subset of SNPs that contains the one i with the highest expected change in allele frequency with a prespecified PCS. For this purpose, we observe the allele frequency change $X_{ij} = f_{ij}^{(T)} - f_{ij}^{(0)}$, between times 0 and T for SNP i from experimental replicate j.

Since, for sufficiently large populations, the temporal changes in allele frequency are approximately normally distributed with variance depending on the initial allele frequency, this may be phrased in terms of a Gaussian subset selection problem with equal sample sizes and unequal variances.

To explore the performance of our method in such a setup, we simulate data following an experimental design described in Barghi et al. (2019). More specifically, we simulate m replicate populations consisting of N_{hap} haplotypes from a window in the genome. We then randomly choose one selected SNP with a beneficial allele that occurs in h_{sel} haplotypes. We assume that the beneficial allele has a selective advantage of $s = 0.05$. The frequency dynamic is then simulated for 60 generations, with an effective population size of 500. The adapted CMH test is then performed with a 0.05 level of significance, and the Benjamini and Hochberg multiple testing correction is used. SNPs that provide significant p values are then chosen as "populations" on which subset selection is applied. For each population, a sample of m observations is available due to m replicate populations. For the adaptive approach, we employ the AS method with $\lambda = 0.5$. We set the level of significance to 0.05 to aim for a PCS of 95%.

Figure 7 provides simulated sizes of selected subsets assuming $m = 3, N_{\text{hap}} = 10$, and $h_{sel} = 2$. It turns out that the use of subset selection leads to a substantial reduction in the number of candidates for beneficial SNPs, with an additional reduction when using the AS method. In this example, all methods always included the selected target.

It should be mentioned that Gaussian subset selection works less well (in terms of PCS control) for beneficial SNPs with high initial frequencies, at least if the number of replicate populations is small. This is because the causal SNP may in this case not have the largest expected change in allele frequency (since f_i^t increases by

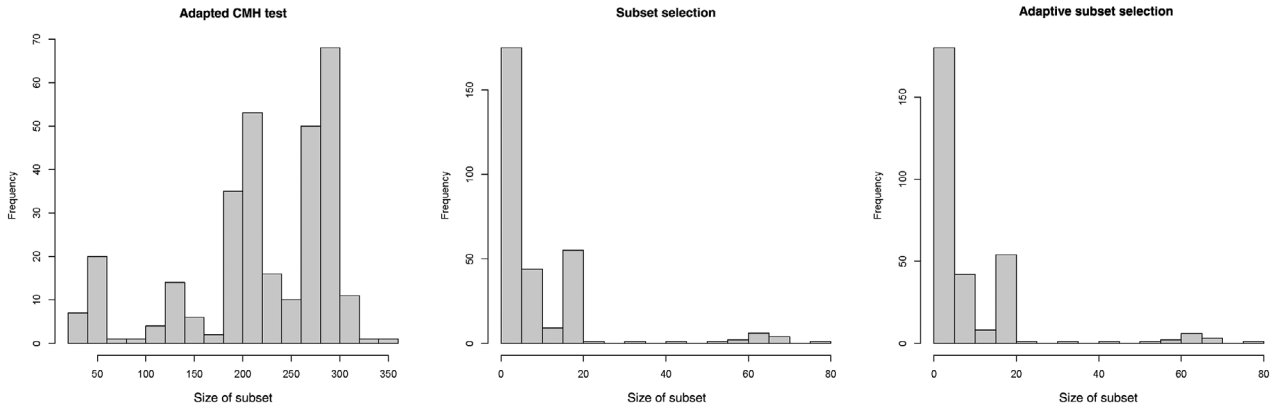

FIGURE 7 | Results showing the distribution of the subset size given a 0.05 level of significance. Here, we employ the parameter choice of $m = 3, N_{\text{hap}} = 10$, and $h_{sel} = 2$. The left panel shows the direct result from the adapted CMH test, the middle panel from basic subset selection, and the right panel using the AS method.

much less than a factor $1 + s$ in Equation 16 for large f_i^t) as well as the normal approximation working less well in the binomial tails.

5.3 | Potential Application in Clinical Trials

Clinical trials are an essential component of the medical research process that helps to determine the safety and efficacy of new drugs or treatments. Traditional trials often only consist of a treatment arm and a control arm to address a single question. More novel methods such as basket and umbrella trials expand on this design, having multiple treatment arms or addressing multiple questions at the same time. This allows multiple hypotheses to be tested simultaneously, offering enhanced efficiency and reduced cost (Mills et al. 2019).

Compared to umbrella trials, platform trials offer further ability to drop or add treatment arms after some interim analysis (Jaki 2015) and have been extensively used recently (Alexander et al. 2018; Fountzilas et al. 2022). The trial starts with multiple treatment arms at the first stage, followed by an interim analysis. The existing treatment arms are compared in terms of performance, with the worst performing arms dropped, and potentially some new arms added. These interim analyses can be repeated multiple times during the trial. They may be carried out using a subset selection rule. Indeed, one may use subset selection to drop non-best populations for the next part of the analysis. If the number of treatment arms is large, adaptive methods can also provide further advantages in terms of power.

An alternative application is on basket trials, where two arms are tested on multiple diseases. Here the aim is to identify the disease on which the treatment performs most effectively. Such a disease can then be defined as the best in a subset selection problem. Adaptive methods might be especially useful if such trials have many populations compared to platform trials.

6 | Conclusion

In this paper, we proposed adaptive methods based on Gupta's approach (Gupta and Panchapakesan 1972) that are significantly less conservative under scenarios that differ noticeably from the LFC. Our methods perform especially well when the number of populations is large.

The RAS method has slightly higher power compared to the AS method, but this comes at the cost of a higher PCS close to the LFC scenarios. The RAS method also performs better than the AS method when both use the automatic tuning approach for the tuning parameter λ, but at the price of a higher computational cost.

Moreover, we apply our adaptive methods also to scenarios when sample sizes and/or variances are unequal. Our proposed approach based on Tukey's conjecture is considerably less conservative than older approaches by Gupta and Huang, with no additional inflation in PCS.

Our results on real data also demonstrate the potential of our approach to solving practical problems. Our application to two completely different areas, one in wheat productivity, and the other in identifying targets of selection in evolve and resequence experiments, shows the wide range applicability of subset selection and our proposed adaptive method. Further applications in other areas such as system designs and clinical trials seem promising, and we hope that our examples can serve as an incentive to further explore applications in these areas.

Even though our work focuses mainly on modifications of Gupta's test statistic, adaptive subset selection could also be developed for other rules, further distributions, and definitions of best. Chang and Huang (2001) for example proposed an alternative rule that can be applied in the context of problems involving heteroscedasticity but does not guarantee PCS control. In Nagel (1970), methods are derived for a wide range of population distributions including Gamma, Binomial, and Poisson.

Acknowledgments

We are grateful to Professor Wen-Tao Huang who provided input to an earlier version of this work. This project was supported by the Austrian Science Fund (FWF; DK W1225-B20).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

We implemented our proposed approach in an R R Core Team (2021) package available at https://github.com/xthchen/adass. The datasets used are from Kansas State University (2022) and Barghi et al. (2019).

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "Reproducible Research" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

References

Alexander, B. M., S. Ba, M. S. Berger, et al. 2018. "Adaptive Global Innovative Learning Environment for Glioblastoma: GBM AGILE." *Clinical Cancer Research* 24, no. 4: 737–743.

Barghi, N., R. Tobler, V. Nolte, et al. 2019. "Genetic Redundancy Fuels Polygenic Adaptation in Drosophila." *PLOS Biology* 17, no. 2: 1–31.

Barton, N. H. 2000. "Genetic Hitchhiking." *Philosophical Transactions: Biological Sciences* 355, no. 1403: 1553–1562.

Bechhofer, R. E. 1954. "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations With Known Variances." *The Annals of Mathematical Statistics* 25, no. 1: 16–39.

Best, D. J., and J. C. W. Rayner. 1987. "Welch's Approximate Solution for the Behrens-Fisher Problem." *Technometrics* 29, no. 2: 205–210.

Chang, Y.-P., and W.-T. Huang. 2001. "Generalized Subset Selection Procedures Under Heteroscedasticity." *Journal of Statistical Planning and Inference* 98, no. 1: 239–258.

Chen, H. J., E. J. Dudewicz, and Y. J. Lee. 1976. "Subset Selection Procedures for Normal Means Under Unequal Sample Sizes." *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002)* 38, no. 3: 249–255.

Efron, B., and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Finner, H., and V. Gontscharuk. 2009. "Controlling the Familywise Error Rate With Plug-In Estimator for the Proportion of True Null Hypotheses." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71, no. 5: 1031–1048.

Finner, H., S.-Y. Tang, X. Cui, and J. C. Hsu. 2021. "Partitioning for Confidence Sets, Confident Directions, and Decision Paths." In *Handbook of Multiple Comparisons*, edited by X. Cui, T. Dickhaus, Y. Ding, and J. C. Hsu, chapter 4. New York: Chapman and Hall/CRC.

Fountzilas, E., A. Tsimberidou, H. Vo, and R. Kurzrock. 2022. "Clinical Trial Design in the Era of Precision Medicine." *Genome Medicine* 14: 101.

Gupta, S. S., and S. Panchapakesan. 1972. "On Multiple Decision (Subset Selection) Procedures." *Journal of Mathematical and Physical Sciences* 6: 1–71.

Gupta, S. S., and D.-Y. Huang. 1976. "Subset Selection Procedures for the Means and Variances of Normal Populations: Unequal Sample Sizes Case." *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002)* 38, no. 2: 112–128.

Gupta, S. S., and W.-T. Huang. 1974. "A Note on Selecting A Subset of Normal Populations With Unequal Sample Sizes." *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* 36, no. 4: 389–396.

Gupta, S. S., and S. Panchapakesan. 2002. *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. Philadelphia, PA: SIAM.

Hayter, A. J. 1984. "A Proof of the Conjecture That the Tukey-Kramer Multiple Comparisons Procedure is Conservative." *The Annals of Statistics* 12, no. 1: 61–75.

Hoang, A.-T., and T. Dickhaus. 2020. "On the Usage of Randomized p-Values in the Schweder–Spjøtvoll Estimator." *Annals of the Institute of Statistical Mathematics* 74: 289–319.

Hsu, J., and B. Nelson. 1988. "Optimization Over A Finite Number of System Designs With One-Stage Sampling and Multiple Comparisons With the Best." In *1988 Winter Simulation Conference Proceedings*, 451–457. Piscataway, NJ: IEEE.

Hsu, J. C. 1984. "Ranking and Selection and Multiple Comparison With the Best." In *Design of Experiments: Ranking and Selection*, edited by T. J. Santner and A. C. Tamhane, 23–33. New York: Marcel Dekker.

Jaki, T. 2015. "Multi-Arm Clinical Trials With Treatment Selection: What Can be Gained and at What Price?" *Clinical Investigation* 5: 393–399.

Kansas State University. 2022. "2022 Kansas Performance Tests With Winter Wheat Varieties." Contribution Number 23-025-S. Manhattan, KS: Kansas Agricultural Experiment Station.

Kim, S.-H., and A. S. Cohen. 1998. "On the Behrens-Fisher Problem: A Review." *Journal of Educational and Behavioral Statistics* 23, no. 4: 356–377.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad. 2005. "Estimating the Proportion of True Null Hypotheses, With Application to DNA Microarray Data." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, no. 4: 555–572.

Mills, E., J.-H. Park, E. Siden, et al. 2019. "Systematic Review of Basket Trials, Umbrella Trials, and Platform Trials: A Landscape Analysis of Master Protocols." *Trials* 20: 572.

Nagel, K. R. O. 1970. *On Subset Selection Rules With Certain Optimality Properties*. PhD thesis, Purdue University.

Neuhauser, C. 2004. *Mathematical Models in Population Genetics*, chapter 19. New York: John Wiley & Sons.

Patil, V. 1969. "Approximation to the Generalized Behrens-Fisher Distribution Involving Three Variates." *Biometrika* 56, no. 3: 687–689.

R Core, Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schweder, T., and E. Spjøtvoll. 1982. "Plots of p-Values to Evaluate Many Tests Simultaneously." *Biometrika* 69, no. 3: 493–502.

Spitzer, K., M. Pelizzola, and A. Futschik. 2020. "Modifying the Chi-Square and the CMH Test for Population Genetic Inference: Adapting to Overdispersion." *The Annals of Applied Statistics* 14, no. 1: 202–220.

Storey, J. D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64, no. 3: 479–498.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone. 2011. "Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in Drosophila Melanogaster." *PLOS Genetics* 7, no. 3: 1–10.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.