The Annals of Applied Statistics
2022, Vol. 16, No. 2, 936–958
https://doi.org/10.1214/21-AOAS1526

PERMUTATION TESTS UNDER A ROTATING SAMPLING PLAN WITH CLUSTERED DATA

BY JIAHUA CHEN[1,a], YUKUN LIU[2,d], CAROLYN G. TAYLOR[1,b] AND JAMES V. ZIDEK[1,c]

[1]*Department of Statistics, University of British Columbia,* [a]*jhchen@stat.ubc.ca,* [b]*cgtaylor@stat.ubc.ca,* [c]*jim@stat.ubc.ca*
[2]*School of Statistics, East China Normal University,* [d]*ykliu@sfs.ecnu.edu.cn*

The distribution of lumber strength of any grade may evolve, for example, due to climate change, forest fire, changes in processing methods, and other factors. So, in North America the forest products industry monitors the evolution of their means, percentiles, or other parameters to ensure the wood products meet the industrial standard. For administrative convenience and informativeness, one may adopt a rotating sampling plan by sampling 36 mills in the initial occasion and having six of them replaced in each successive occasion for the next five occasions. The strength data on a specified number, commonly 10 pieces of lumbers from each sampled mills, are obtained. Under such rotating plans the observations on pieces from the same mill are correlated, and the observations on samples from the same mill taken on different occasions are also correlated. Ignoring these correlations may lead to invalid inference procedures. Yet accommodating a cluster structure in parametric models is difficult and entails a high level of misspecification risk. In this paper we explore symmetry in the clustered data collected via a rotating sampling plan to develop a permutation scheme for testing various hypotheses of interest. We also introduce a semiparametric density ratio model to link the distributions of the response variable over time. The combination retains the validity of the inference methods while extracting maximum information from the sampling plan. A simulation study indicates that the proposed permutation tests firmly control the type I error whether or not the data are clustered. The use of the density ratio model improves the power of the tests. We also apply the proposed tests to data from the motivating application. The proposed permutation tests effectively address many real-world issues with trust worth inference conclusions.

1. Introduction.

The general theory described in this paper was developed to solve an inferential problem that arises in a long term monitoring program in the forest products industry of North America. We begin by describing in the next subsection how that problem arises. We then describe the general theory, whose potential domain of application goes well beyond that specific domain.

1.1. *Monitoring the breaking strength of lumber.*

The industrialization of lumber manufacturing began more than a century ago with the establishment of standards for lumber. These were defined primarily by strength, reducing characteristics such as knots. The impact of each such characteristic was published in what has become ASTM D245 (ASTM (2006)). Lumber could then be graded largely on the basis of these characteristics and, thanks to the standards document, their strength metrics predicted. Suitably adjusted metrics become the grade's design value DV, a value that would be exceeded with high probability when the lumber is put into service.

Wood has become an increasingly important natural resource because, unlike other resources, for example, petroleum, its production is sustainable. Numerous products are now

Received February 2020; revised August 2021.

Key words and phrases. Density ratio model, empirical likelihood, exchangeability, multiple sample, percentila, rank test.

made from wood including biofuel as well as building materials. Wood's versatility and sustainability have led to its gradually replacing these other resources, especially in this era of climate change. In fact, lumber manufacturing has become a major global industry. The wood used to make lumber comes not only from forests but also from plantations. Trees are cut down and trimmed to get logs that are transported to mills where they are sawn into lumber in an optimal way. The pieces of lumber are then graded, primarily according to their strength. Thus, at any given time the population of lumber is subdivided into subpopulations represented by grades. And these subpopulations will change dynamically over time, as the lumber is produced and consumed.

These subpopulations may be thought of as samples from an essentially infinite conceptual super-population, defined by the processes by which the logs are harvested and lumber manufactured. It is these super-populations that are of concern in long-term monitoring programs, especially because of climate change and its potential impact on trees, for example, through insect infestations and major forest fire. That concern can be expressed though the probability distributions of metrics that express that impact on the properties of lumber strength, such as in pound per square inch psi.

Lumber possesses many types of strength: notably under stretching (ultimate tensile strength or UTS), compression, or bending (modulus of rupture or MOR). Its stiffness (modulus of elasticity or MOE), which is related to all these other characteristics, is, unlike MOR, not measured by destructive testing. That DV is a specified quantile of the strength distribution, commonly a median or the fifth percentile. Thus, the grade of a piece of lumber for engineering applications depends on its intended use. The top grade is both strong and expensive. The development of the modern grading system has been a triumph of structural engineering since it has standardized lumber properties. Thus, wood, a heterogeneous material, unlike say aluminum, can be used with the assurance that the lumber made from it has a low probability of failure when used for its intended purpose, where "probability" would refer to this super-population.

The importance of lumber has led to the need to monitor those metrics over time. The first such long-term monitoring program was established in 1994 in the southeastern United States. Cross-sectional samples were taken annually using a stratified-by-region sampling plan. The number of mills in each region was determined, and the primary sampling units (PSUs) within a region were chosen by simple random sampling. One or two bundles, that is, secondary sampling units (SSUs), of about 300 pieces each were selected. From each, a "lot" with 10 pieces was chosen in a prescribed way, and their mechanical strengths were measured.

Canada also established a long-term monitoring program. Planning for a pilot program began in 2005; a preliminary analysis showed a substantial variation between mills, within mills, and between lots. The goal at the time was to measure temporal trends in the elasticity of lumber or formally its modulus of elasticity MOE. Due to its efficiency in estimating trends, the *rotating sampling plan* was selected for a specified grade of lumber, with a six-year rotation. We sample 36 mills initially and replace six of them on each successive occasion. This plan has the benefit of limiting the mill response burden, makes a consistent random mill effect over time (six years) plausible, and refreshes the sample to maintain some degree of cross-sectional validity.

This led to new challenges: the statistical theory needed to assess trends in MOR under a rotating sampling plan did not exist, although there are many recent and relevant publications on the rotating sampling plan, such as Karna and Nath (2015), Nijman, Verbeek and van Soest (1991), Park, Choi and Kim (2007). The Forest Products Stochastic Modelling Group (FPSMG), based at the University of British Columbia, was, therefore, established. It was cofunded by FPInnovations, a nonprofit industrial research lab, and the Natural Sciences and Engineering Research Council of Canada. The FPSMG, which has involved engineers and wood scientists at FPInnovations working in collaboration with statistical faculty and students, is in its eleventh year at the time of writing. It has made numerous contributions to the theory and practice of strength measurement and monitoring for forest products; see, for example, Zidek and Lum (2018), Cai, Chen and Zidek (2017), Chen et al. (2021), and Chen and Liu (2013).

Meanwhile, for reasons beyond the scope of this paper, a separate North American long-term monitoring program has been specified in a revision of an American Society for Testing and Materials (ASTM) standards document (D1990). It assumes a cross-sectional sample once every five years and specifies, among other things, that a Wilcoxon test be used to assess change in the fifth percentile of the MOR. The document ignores both the PSU and SSU cluster effects induced by their random effects. In a companion article (Chen et al. (2021)) to this one, an alternative method has been proposed for use in the new ASTM monitoring plan. Chen et al. (2021) is based on a cross-sectional clustered data without longitudinal effect and focuses on parameter estimation rather than hypothesis testing.

1.2. *Summary and outline of the paper.* This paper targets the hypothesis testing problem on trends in strength percentiles for rotating sample designs, where samples are taken every year or every specified period, while it is also applicable to detecting trends in other population parameters. Its genesis lies in the need for a method that can handle the cluster effect across time and space. In Section 2 we describe the rotating sampling plan and its implied random effects. Some existing and potential data analysis approaches will be given in this section. Section 3 presents a general foundation for inference, based on permutation tests; the necessary theorems and their proofs appear here. Applying this methodology requires the analyst to choose a test statistic with which to apply the permutation strategy, and a number of possibilities are presented in Section 4. A particularly novel choice is based on the so-called density ratio model (DRM). Section 5 presents some populations for which the theory can be applied. Section 6 presents simulation results. We find the type I errors of all permutation tests are well controlled while their asymptotic versions (if existing) have inflated type I errors if derived for independent and identically distributed (IID) observations. The use of the DRM improves the power of the tests and leads to an efficiency gain, in general, with a sensible choice of the basis function. Finally we devote Section 7 to the real-world application that leads to the work described in this paper. We describe the sampling plan employed as well as some summary statistics of the real data. We apply all viable tests to this data set. The results show our permutation tests effectively prevent many potentially damaging clustering effects. Our methods lead to trust worth inference conclusions. The paper ends with a discussion in Section 8 and an Appendix that gives technical details for the numerical strategy employed in the simulation experiments.

2. Rotating sampling plan, random effects, and assumptions. Consider a grand population made of a finite number of PSUs, and the composition of the PSUs remains stable over time. Each PSU is made of a practically infinite number of SSUs. The SSUs may be thought of as samples from a conceptual super-population that evolves over time and space. We are interested in monitoring the trends in the distribution of metrics on SSUs formed by the sample from the super-population at different times. In the motivating application the PSUs are lumber producing mills, and SSUs are pieces of lumber. A rotational sampling plan draws a number of PSUs initially and replaces a subset on each successive occasion. For instance, one may sample 36 mills initially and have six of them replaced on each successive occasion. One further samples 10 pieces of lumber from each sampled mills and obtains the strength data. We monitor the trend in the distribution of metrics on SSUs over time.

2.1. *Literature review and assumptions.* While rotating sampling plans are administratively convenient and informative, they lead to challenges in developing inferential tools, due to their induced cluster effects across both time and space. Moreover, ignoring the cluster structure leads to inflated type I errors of some established tests (Datta and Satten (2008), Verrill, Kretschmann and Evans (2015)). In fact, as far as we know, no exact methods exist for monitoring the change of population quantiles, based on rotating sampling plans for this application. As suggested by a referee, some approaches might be adopted for this purpose (Berg, Cecere and Ghosh (2014), Francisco and Fuller (1991)).

This paper suggests that permutation tests are well suited to meet these challenges. The permutation is a general approach that plays an active role in modern statistical practice (Hemerik and Goeman (2018), Pesarin and Salmaso (2010), Hemerik, Solari and Goeman (2019)). To enhance statistical efficiency, we further recommend the semiparametric density ratio model or DRM (Anderson (1979), Qin and Zhang (1997)) to link the multiple distributions of the response variable at different time points derived from a rotating sampling plan. We use the empirical likelihood (Owen (2001), EL) to construct test statistics.

We now go over some specifics and assumptions. Let $n = mN$ be the number of PSUs sampled from the population initially in a rotating sampling plan. Here, m is the number of PSUs replaced on each occasion, and N is the number of occasions in the rotating sampling plan. When $n = 36$ and $m = 6$, six PSUs are replaced in each of the next $N - 1 = 5$ occasions. Let r be the number of SSUs drawn from each sampled PSU.

For ease of presentation, the number of SSUs sampled in each selected PSU is assumed to be the same, although our approach works more generally. The super-population nature of the targeted application makes it natural to regard the sampling of PSUs be done with replacement. This assumption facilitates methodological developments, though it may lead to harmless conservative inference procedures. Hence, it is widely accepted in the survey context (Rao and Shao (1992)). We denote the data obtained on SSUs, sampled from a PSU, as $\mathbf{y}_{k,i} = (y_{k,i,1}, \dots, y_{k,i,r})^\tau$. To simplify the notation, we let $\mathbf{y}_k = \{\mathbf{y}_{k,i} : i \in s_k\}$ with $k = 0, 1, 2, \dots, K$ and s_k being the set of PSUs in the sample on occasion k. We assume that vector $\mathbf{y}_{k,i}$, for $i \in s_k$, have the same multivariate distribution, denoted by $F_k(\mathbf{y})$, for data in occasion k. Namely, this data set is representative of the super-population at occasion k. This simplification assumption may not be suitable when the rotating sampling plan contains some complex features. For instance, the inclusion probability of a PSU may depend on some covariates or on the response values of its SSUs, making the plan informative (Pfeffermann and Sverchkov (2009)). Ignoring these factors may yield to large biases and erroneous inference. Tackling these issues must be guided by the real-world application. We are getting better understanding of such issues and hope to tailor the proposed method to these plans in the future.

The observations within each $\mathbf{y}_{k,i}$ (for fixed k, i) are dependent because they are SSU values from the same PSU and obtained on the same occasion. This leads to within-population cluster/random effects. Data in $\mathbf{y}_{k_1,i}$ and $\mathbf{y}_{k_2,i}$, with $k_1 \neq k_2$ and fixed i, are collected from the same PSU on two occasions. Their connection through shared PSU likely leads to longitudinal cluster/random effects. In summary, the data collected by a rotating sampling plan have both longitudinal and cross-sectional clustering structures.

2.2. *Properties of the population and sampling plan.* Let s_k be the indices of the PSUs included in the kth sample ($k = 0, 1, \dots$). To fix ideas, we highlight the following properties of the population and data from rotating sampling plans:

1. Multiple samples are collected on several occasions from the same grand population via a rotating sampling plan, and the response values for the same unit may evolve.

2. Each cluster i forms a vector-valued time series of responses $\mathbf{y}_{k,i}$ over $k = 0, 1, \dots, K$. The time series formed by different clusters are mutually independent.

3. The joint distribution F_k of $\mathbf{y}_{k,i}$, which is common for all i, is exchangeable with marginal distribution G_k.

4. The marginal distributions of any single response G_k, $k = 0, 1, \dots, K$, satisfy the DRM to be specified in equation (5) with a known basis function $\mathbf{q}(\cdot)$. For expository simplicity the specific features of the DRM will be given later.

5. When $G_k = G_{k+1}$, the joint distributions of $\{\mathbf{y}_{t,i}, t = 0, 1, \dots, k-1, k, k+1, \dots, K\}$ and $\{\mathbf{y}_{t,i}, t = 0, 1, \dots, k-1, k+1, k, \dots, K\}$ are identical for any i.

The properties above, except No. 4, are not too technical and are plausible in the targeted applications. The DRM assumption in No. 4 is also reasonable: its validity mostly relies on the nonradical evolution of the population characteristics. Using this model leads to improved efficiency when it is approximately satisfied. The efficiency gain remains when this assumption is mildly violated, as we will show in the simulation section.

We note that G_k is the distribution of the response value of a single SSU randomly selected from the kth population. In this paper we propose a permutation test for hypotheses concerning functionals of G_k, based on multiple samples collected via the rotating sampling plan described above.

3. Permutation tests.

Let F be the data-generating distribution and R a test statistic designed to test a null hypothesis against a specific alternative hypothesis: H_0 and H_a. We assume that a larger R supports $F \in H_a$. To construct a test of size $\alpha \in (0, 1)$, we look for a constant c_α such that

$$\sup\{P(R > c_\alpha | F) : F \in H_0\} = \alpha.$$

Let the observed value of R be R_{obs}. The test rejects H_0 if $R_{\text{obs}} > c_\alpha$. One may equivalently compute a p-value

$$p = \sup\{P(R > R_{\text{obs}} | F) : F \in H_0\}$$

and reject H_0, when $p \le \alpha$, for that will imply $R_{\text{obs}} > c_\alpha$ and hence rejection by the Neymann-Pearson hypothesis-testing criterion.

Given the above, the ultimate task of developing a valid test is to find an effective statistic R and a way to compute the resulting p-value while bypassing the need to specify c_α explicitly. In the context of tests based on multiple samples from a rotating sampling plan, let

$$R_n = R_n(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_K)$$

be the test statistic of choice, with the subindex added to highlight its dependence on the sample size. Suppose the population distribution does not change from occasion 0 to occasion 1: namely, $G_0 = G_1$. Then, $(\mathbf{y}_{0i}, \mathbf{y}_{1i})$ and $(\mathbf{y}_{1i}, \mathbf{y}_{0i})$ have the same distribution for all $i \in s_0 \cap s_1$. Taking advantage of this symmetry, we design a permutation procedure as follows:

Step I. For each $j \in s_0 \cap s_1$, generate a random permutation (a, b) of $(0, 1)$, independent of all other random variables, such that

$$P\{(a, b) = (0, 1)\} = P\{(a, b) = (1, 0)\} = 0.5, \tag{1}$$

and let $(\mathbf{y}^*_{0,j}, \mathbf{y}^*_{1,j}) = (\mathbf{y}_{a,j}, \mathbf{y}_{b,j})$. Let $\mathbf{y}^*_{0,j} = \mathbf{y}_{0,j}$ and $\mathbf{y}^*_{1,j} = \mathbf{y}_{1,j}$ for $j \in s_0 - s_1$ and $j \in s_1 - s_0$, respectively.

Step I+. Let $|s_0 - s_1|$ be the number of units in $(s_0 - s_1)$. Draw $|s_0 - s_1|$ units from $(s_0 - s_1) \cup (s_1 - s_0)$, using simple random sampling without replacement. Denote the resulting clustered observations by $\mathbf{y}^*_{0i}$ $(i = 1, 2, \dots, |s_0 - s_1|)$ and the remaining clustered observations by $\mathbf{y}^*_{1i}$ $(i = 1, 2, \dots, |s_1 - s_0|)$.

Step II. Form a permuted multiple-sample $\{\mathbf{y}^*_{0,j}, j \in s_0\}$, $\{\mathbf{y}^*_{1,j}, j \in s_1\}$, and $\mathbf{y}^*_{ik} = \mathbf{y}_{ik}$ for $i \in s_k$ for $k = 2, \dots, K$.

We now present the proposed permutation test.

PERMUTATION TEST. For each permuted multiple-sample, compute the value of the test statistic

$$R^*_n = R_n(\mathbf{y}^*_0, \mathbf{y}^*_1, \dots, \mathbf{y}^*_K).$$

Generate permutation samples repeatedly and independently, say $M = 10{,}001$ times. Compute the permutation test p-value

$$p^* = \text{Proportion of } \{R^*_n > R_{\text{obs}}\}.$$

Reject the null hypothesis if $p^* < \alpha$ where α is the nominal level of the test.

In applications the practitioner conducts the test on a single data set, whereas in research projects analyses may be done with thousands of simulated data sets. Hence, it is computationally affordable to choose a large M in applications. The margin of error of p^* with the currently recommended M is about $(0.95 \times 0.05)^{0.5} \times 1.96/M^{0.5} \le 0.005$. Allowing M to be an odd number helps to avoid minor operational issues. In our simulation study we use a much smaller M to allow for a large number of simulation repetitions. Our reliance on the average performance of the tests, rather than on accurate approximations in each repetition, validates our choice of a smaller M.

THEOREM 3.1. *Let* $(\mathbf{y}^*_1, \mathbf{y}^*_2, \dots, \mathbf{y}^*_K)$ *be a permutation multiple-sample obtained via Steps I and II above. Assume that the null hypothesis* $G_0 = G_1$ *is true, and the model assumptions specified in the summary subsection hold. Then, we have the following results*:

(a) $R^*_n = R_n(\mathbf{y}^*_0, \mathbf{y}^*_1, \dots, \mathbf{y}^*_K)$ *has the same distribution as* $R_n(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_K)$.

(b) *Given* $\{\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_K\}$, R^*_n *has a discrete uniform distribution over all possible values in the range of* $R_n(\mathbf{y}^*_0, \mathbf{y}^*_1, \dots, \mathbf{y}^*_K)$.

PROOF. (a) When the null hypothesis holds, the joint distribution of $(\mathbf{y}_{0,i}, \mathbf{y}_{1,i}, \mathbf{y}_{2,i}, \dots, \mathbf{y}_{K,i})$ is the same as that of $(\mathbf{y}_{1,i}, \mathbf{y}_{0,i}, \mathbf{y}_{2,i}, \dots, \mathbf{y}_{K,i})$ for all i, including all $i \in (s_0 - s_1) \cup (s_1 - s_0)$. At the same time, $\mathbf{y}_{0,i}, \mathbf{y}_{1,i}, \mathbf{y}_{2,i}, \dots, \mathbf{y}_{K,i}$ with different i's are mutually independent. Therefore, the permutation Step I results in a new data set whose joint distribution remains the same as that of $\{\mathbf{y}_{k,i}, i \in s_k, k = 0, 1, \dots, K\}$. Therefore, $R^*_n = R_n(\mathbf{y}^*_0, \mathbf{y}^*_1, \dots, \mathbf{y}^*_K)$ has the same distribution as $R_n(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_K)$.

(b) The permutation prescribed in equation (1) ensures that every permutation outcome has an equal probability. Hence, R^*_n has a uniform distribution on these possible values. This argument ignores rare but possible ties among these values. In such cases we interpret the uniform distribution as a distribution with probabilities proportional to the cardinality of each distinct permutation outcome. □

REMARK 1. The observed value R_{obs} of R_n may be regarded as one random outcome of R^*_n. The conclusions in the above theorem hence ensure that the type I error of the permutation test equals the nominal level, excluding the round-off error.

REMARK 2. The alternative hypothesis does not appear relevant in the proof or theorem statement, but it matters for the actual test. It determines the choice of the test statistic R_n. We choose the R_n that is the most sensitive to the departure of the distribution in the direction of H_a, rather than arbitrary departures from the null hypothesis. For this reason the stochastic size of R_n should increase when the data-generating distribution F is conceptually deep in H_a and far from H_0. In our target application, for example, if H_a states that the population mean of G_1 is larger than that of G_0, then an effective choice of R_n is the difference in the two sample means (namely $\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_0$). A larger difference in the population means leads to a stochastically larger difference in the sample means. If one chooses R_n to be the difference in the two sample variances, the resulting test may also suggest that H_0 (unequal mean) should be rejected, but for a wrong reason.

REMARK 3. Step I+ permutes the units in $(s_0 - s_1) \cup (s_1 - s_0)$. The conclusion in Theorem 3.1 breaks down when Step I+ is included: namely, R^* may have a slightly different distribution from R_n under H_0. However, under the null hypothesis the difference introduced by this extra step is minor. At the same time the units in $(s_0 - s_1) \cup (s_1 - s_0)$ contain crucial information when H_a is true. Hence, we recommend that Step I+ be included. Our simulation study shows that the type I errors are not affected.

REMARK 4. In applications, things may not go as planned. A few PSUs may drop out from the rotating sampling plan. New mills may open, and dormant mills may resume production. Some modification is needed: permute only units sampled on both occasions, and use Step I+ to handle the unmatched mills.

4. Statistics of choice in permutation tests.

In this section we propose some promising statistics R_n for the permutation test. The choice of R_n affects the statistical efficiency but not the validity of the test.

4.1. *Straightforward choices of test statistics.*

Let the null hypothesis be $H_0 : G_0 = G_1$ and the alternative be $H_a : \xi(G_0) > \xi(G_1)$ with $\xi(G)$ being the mean, the quantiles of G, or another population parameter.

Two immediate choices are the classical t and Wilcoxon rank-sum statistics with the cluster structure in the data ignored. The first one is

$$T = \frac{\bar{y}_0 - \bar{y}_1}{\sqrt{(1/n_0 + 1/n_1)s^2}}. \tag{2}$$

Here, n_0 and n_1 are the numbers of SSUs sampled in occasions 0 and 1, $\bar{y}_1$ and $\bar{y}_0$ are the sample means, and s^2 is the pooled sample variance ignoring the cluster structure. The second one is

$$W = \sum_{i \in s_1} \sum_{j \in s_0} \sum_{1 \le u,v \le r} \mathbb{1}(y_{0,i,u} > y_{1,j,v}), \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function and the summation is over all observations on occasions 0 and 1. The Wilcoxon statistic is usually normalized in order to use the central limit theorem, but this is unnecessary when the permutation approach is applied.

These two tests were originally designed to handle IID data. The t-test further requires that the data are from a normal distribution, and it detects the difference in the population means. The Wilcoxon rank-sum test is nonparametric and primarily used to detect a location shift in two distributions, although, in theory, it works only on the size of $P(X < Y)$. Such limitations are often overlooked in applications, yet the tests serve general purposes surprisingly well. However, this is not true for clustered data. For such data the tests have inflated sizes (higher type I errors) if the clustering and temporal dependence are ignored. The generalization of the Wilcoxon test to independent clusters can be found in Datta and Satten (2005, 2008), and Rosner, Glynn and Lee (2006). Their results are not applicable to clustered data with longitudinal random effects.

Let $\hat{G}_0$ and $\hat{G}_1$ be the distributions fitted by any reasonable method. We may use a straightforward statistic for the permutation test,

$$R_n(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_K) = \xi(\hat{G}_0) - \xi(\hat{G}_1). \tag{4}$$

Obvious choices for $\hat{G}_0$ and $\hat{G}_1$ are the empirical distributions ignoring the cluster structure, based on samples from G_0 and G_1. Another possibility will be given in the next section. We are most interested in this type of statistic for population percentiles.

4.2. *DRM-assisted choices.* Under rotating sampling plans the multiple-samples are collected from closely related distributions. They naturally share some intrinsic latent structure. Accounting for this structure leads to more efficient estimates of G_0 and G_1 and, therefore, more powerful permutation tests. We recommend the DRM introduced by Anderson (1979); we believe that it fits a broad range of situations. The DRM has been successfully used by many researchers, including Qin and Zhang (1997), Qin (1998), and Keziou and Leoni-Aubin (2008).

The DRM links the population distributions G_k, $k = 0, 1, 2, \dots, K$, by

$$dG_k(y) = \exp\{\boldsymbol{\theta}_k^\top \mathbf{q}(y)\}\, dG_0(y) \tag{5}$$

for some prespecified basis function $\mathbf{q}(y)$ and parameter $\boldsymbol{\theta}_k$. Note that $\boldsymbol{\theta}_0 = 0$ when G_0 is chosen as the base distribution. We require the first component of $\mathbf{q}(y)$ to be 1 to make the first component of $\boldsymbol{\theta}$ a normalization parameter. We use the EL of Owen (2001) as the platform for the inference. In the spirit of the EL, we require G_0 to have the form $G_0(y) = \sum_{k,i,u} p_{k,i,u} \mathbb{1}(y_{k,i,u} \le y)$. We construct the *composite log likelihood function*

$$\ell_n^{\mathrm{C}}(G_0, \dots, G_K) = \sum_{k,i,u} \log p_{k,i,u} + \sum_{k,i,u} \boldsymbol{\theta}_k^\tau \mathbf{q}(y_{k,i,u})$$

with the summation over all possible indices (k, i, u). The DRM assumption implies the constraints

$$\int \exp\{\boldsymbol{\theta}_k \mathbf{q}(y)\}\, d\mathrm{G}_0(y) = \sum_{k,i,u} p_{k,i,u} \exp\{\boldsymbol{\theta}_k^\tau \mathbf{q}(y_{k,i,u})\} = 1$$

for all $k = 0, 1, \dots, K$. The log-likelihood is "composite" because the observations involved are dependent; see Lindsay (1988) and Varin, Reid and Firth (2011) for an introduction to and a general discussion of the composite likelihood.

Given $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$, maximizing $\ell_n(G_0, \dots, G_K)$, with respect to G_0, leads to the profile log empirical likelihood function (in the same notation),

$$\begin{aligned} \ell_n^{\mathrm{C}}(\boldsymbol{\theta}) &= \ell_n^{\mathrm{C}}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \\ &= \sup\Big\{\ell_n^{\mathrm{C}}(G_0, \dots, G_K) : \sum_{k,i,u} p_{k,i,u} \exp\{\boldsymbol{\theta}_j^\tau \mathbf{q}(y_{k,i,u})\} = 1;\ j = 0, 1, \dots, K\Big\}. \end{aligned} \tag{6}$$

Suppose $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \dots, \hat{\boldsymbol{\theta}}_K$ are maximum EL estimators under DRM. The corresponding fitted distribution functions are

$$\check{G}_j(y) = \sum_{k,i,u} \hat{p}_{k,i,u} \exp(\hat{\boldsymbol{\theta}}_j^\tau \mathbf{q}(y_{k,i,u})) \mathbb{1}(y_{k,i,u} \le y). \tag{7}$$

They can then be used in (4) to form statistics for the permutation tests. We give some specific statistics next.

Detecting changes in quantiles under DRM. Let $\xi_\alpha(G)$ be the (100α)th percentile of G with H_0 and H_a being $\xi_\alpha(G_0)=\xi_\alpha(G_1)$ and $\xi_\alpha(G_0)>\xi_\alpha(G_1)$, respectively. The solution for the two-sided alternative follows the same principle.

Under the DRM assumption we give two choices. The first choice is to let

$$R_n=\xi_\alpha(\check{G}_0)-\xi_\alpha(\check{G}_1),$$

where $\check{G}_0$ and $\check{G}_1$ are the fitted distribution functions given in (7). Note that $\check{G}_1(y)$ is a discrete distribution assigning probability $\hat{p}_{k,i,u}\exp(\hat{\boldsymbol{\theta}}_1^\tau\mathbf{q}(y_{k,i,u})$ to $y_{k,i,u}$. Once $\hat{\boldsymbol{\theta}}_1$ is obtained, the rest of the calculation is very simple. How to compute $\hat{\boldsymbol{\theta}}_j$ will be explained later.

The second choice is the empirical likelihood ratio statistic with a computationally friendly alternative. We first pool the samples from G_0 and G_1 to obtain the (100α)th sample percentile: $\hat{\xi}_\alpha$. We then compute the profile constrained composite empirical likelihood

$$\begin{aligned}\ell_n^{\mathrm{CC}}(\boldsymbol{\theta})=\sup\Biggl\{&\sum_{k,i,u}\log p_{k,i,u}+\sum_{k,i,u}\boldsymbol{\theta}_k^\tau\mathbf{q}(y_{k,i,u}):\\ &\sum_{k,i,u}p_{k,i,u}\exp\{\boldsymbol{\theta}_s^\tau\mathbf{q}(y_{k,i,u)}\}=1\text{ for }s=0,1,\dots,K;\\ &\sum_{k,i,u}p_{k,i,u}\exp\{\boldsymbol{\theta}_s^\tau\mathbf{q}(y_{k,i,u)}\}\{\mathbb{1}(y_{k,i,u}\le\hat{\xi}_\alpha)-\alpha\}=0\text{ for }s=0,1\Biggr\}.\end{aligned}\tag{8}$$

The recommended statistic for a permutation test is then

$$R_{n,\xi}=\sup\ell_n^{\mathrm{C}}(\boldsymbol{\theta})-\sup\ell_n^{\mathrm{CC}}(\boldsymbol{\theta}).\tag{9}$$

Against two-sided alternative $(\xi_\alpha(G_0)\neq\xi_\alpha(G_1))$ as in many applications, we directly use $R_{n,\xi}$ in the permutation test. In our targeted application the alternative is one-sided. We use $\operatorname{sgn}(\xi_\alpha(G_0)-\xi_\alpha(G_1))R_{n,\xi}$ in the permutation test.

We use an R-function, called *multiroot*, from the R package *rootSolve* to solve the optimization problem. It solves equations formed by the Lagrange multiplier method for constrained maximization. With the corresponding derivative functions provided, this R-function works well. The details are given in the Appendix.

Detecting changes in the mean under DRM. Suppose we wish to test for $\mu_0=\mu_1$. Once the fitted $\check{G}_j(y)$ is obtained, as given in (7), we can compute their means and construct a test statistic based on their differences. However, when $\mathbf{q}(y)=(1,y,\dots)$, the mean of $\check{G}_k(y)$ equals the sample mean of the data from G_k. Hence, this test reduces to t-test in spirit, except for normalization constant. We, therefore, do not consider this test.

At the same time the DRM based likelihood ratio test is viable. To reduce computational burden, we modify the statistic slightly. We first pool the samples from G_0 and G_1 to obtain the pooled sample mean $\hat{\mu}_{01}$. We then compute the profile constrained composite empirical likelihood

$$\begin{aligned}\ell_n^{\mathrm{CC}}(\boldsymbol{\theta})=\sup\Biggl\{&\sum_{k,i,u}\log p_{k,i,u}+\sum_{k,i,u}\boldsymbol{\theta}_k^\tau\mathbf{q}(y_{k,i,u}):\\ &\sum_{k,i,u}p_{k,i,u}\exp\{\boldsymbol{\theta}_s^\tau\mathbf{q}(y_{k,i,u)}\}=1\text{ for }s=0,1,\dots,K;\\ &\sum_{k,i,u}p_{k,i,u}\exp\{\boldsymbol{\theta}_s^\tau\mathbf{q}(y_{k,i,u)}\}\{y_{k,i,u}-\hat{\mu}_{01}\}=0\text{ for }s=0,1\Biggr\}.\end{aligned}$$

The recommended statistic for a permutation test is then

$$R_{n,\mu}=\sup\ell_n^{\mathrm{C}}(\boldsymbol{\theta})-\sup\ell_n^{\mathrm{CC}}(\boldsymbol{\theta})\tag{10}$$

with $\ell_n^{\mathrm{C}}(\boldsymbol{\theta})$ given in (6). We use its signed version when the alternative is one-sided.

5. Populations satisfying model assumptions. To support the proposed permutation test under the DRM with clustered data and as preparation for a meaningful simulation study and real data case study, we set up several examples.

EXAMPLE (Normal data). Let $\epsilon_{k,i,u}$, for $k = 0, 1, \ldots; i = 1, 2, \ldots; u = 1, 2, \ldots$ be IID standard normal random variables. Let η_i, $i = 1, 2, \ldots$ be IID standard normal random variables and $\eta_{k,i}$, $k = 0, 1, \ldots K$, $i = 1, 2, \ldots$ another set of IID standard normal random variables, where these are mutually independent of $\epsilon_{k,i,u}$. Let

$$y_{k,i,u} = \mu_k + \sigma_{k,1}\eta_i + \sigma_{k,2}\eta_{k,i} + \sigma_{k,3}\epsilon_{k,i,u}$$

for some nonrandom constants μ_k and $\sigma_{k,j}$, $j = 1, 2, 3$.

Based on this construction, the random variables $y_{k,i,u}$, $u = 1, 2, \ldots$ with fixed k, i are not independent but are identically and normally distributed. Their joint distribution is exchangeable within the cluster indexed by (k, i). Furthermore, observations on the units in the same cluster taken on different occasions, for example, y_{k_1,i,u_1} and y_{k_2,i,u_2}, $k_1 \neq k_2$, are correlated through the shared random effect η_i. Given k, the random variables $y_{k,i,u}$ over $i = 1, 2, \ldots$ and $u = 1, \ldots, r$ have identical marginal distributions. We denote this distribution by G_k. It is easy to verify that $G_0(y), G_1(y), \ldots, G_K(y)$ satisfy the DRM conditions with the basis function $\mathbf{q}(y) = (1, y, y^2)^\tau$.

In this model, μ_k is the nonrandom effect specific to the population on occasion k. The random effect η_i is specific to the ith cluster and shared over different occasions through the moderator $\sigma_{k,1}$. The random effect $\eta_{k,i}$ is specific to cluster i and independent over different occasions. The response value of the uth unit in the ith cluster on occasion k is given by $y_{k,i,u}$.

EXAMPLE (Gamma data). A one-parameter Gamma distribution has a degree of freedom parameter γ with density function

$$g^*(y; \gamma) = y^{\gamma-1} \exp\{-y\}\mathbb{1}(y \geq 0)/\Gamma(\gamma),$$

where $\Gamma(\gamma)$ is the well-known Gamma function.

Let $\mathbf{x}$ be a vector and a and b two real numbers. We denote the vector comprised of $ax_i + b$ as $a\mathbf{x}+b$. With this convention we create a complex cluster structure through the operation for $k = 0, 1, \ldots, K$ and $j = 1, 2, \ldots$: $\mathbf{y}_{k,j} = \lambda_k(\epsilon_j + \epsilon_{k,j} + \mathbf{x}_{k,j})$. The elements of the stochastic models for $\mathbf{y}_{k,j}$ are specified as follows:

1. The ϵ_j are independent with distribution $g^*(y; \gamma_1)$. Given cluster j, its value remains the same for all k, so this term leads to a longitudinal random effect.
2. The $\epsilon_{k,j}$ are independent with distribution $g^*(y; \gamma_2)$. It is shared by the entries in cluster j on occasion k, and this design leads to the cross-sectional random effect.
3. $\mathbf{x}_{k,j}$ is a vector of independent random variables with distribution $g^*(y; \eta_k)$ where η_k is the degrees of freedom of occasion k. They contribute most of the variations in the response vector $\mathbf{y}$. The difference in η_k leads to changes in the marginal distribution.
4. λ_k introduces additional scale fluctuations over the occasions.

The marginal distributions of $y_{k,i,u}$ are k-specific and denoted by G_k. Because of the independence between $x_{k,j,u}$, $\epsilon_{k,i}$, and ϵ_j, and the property of the Gamma distribution, G_k is also a Gamma distribution with rate parameter λ_k and degrees of freedom $\gamma_1 + \gamma_2 + \eta_k$. Gamma distributions satisfy the DRM specified in (5) with $\mathbf{q}(y) = (1, y, \log(y))^\tau$.

In summary, by generating multiple samples from this model we obtain $\{\mathbf{y}_{k,i}, i \in s_k\}_{k=0}^K$ with both longitudinal and cross-sectional random effects, as described in Section 2.

EXAMPLE (General data). Consider a population made of a large number of realized values of a random sample from a super-population F. Denote these as $x_{k,i,u}$ with (k, i, u) carrying no structural information at the moment. Let $\mathbf{y}_{k,i} = \varphi(x_{k,i,1}, \ldots, x_{k,i,r}; \epsilon_{k,i}, \epsilon_i)$, where:

1. $\varphi(\cdot; \epsilon_{k,i}, \epsilon_i)$ is an r-dimensional vector-valued function, symmetric in $x_{k,i,1}, \ldots, x_{k,i,r}$;
2. ϵ_i: $i = 1, 2, \ldots$ are IID;
3. $\epsilon_{k,i}$ are independent for different (k, i), and they are identically distributed given k.

In this general setting the multiple samples $\{\mathbf{y}_{k,i}, i \in s_k\}_{k=0}^{K}$ have the cross-sectional and longitudinal random effects described in Section 2. In addition, when $G_k = G_{k+1}$ for some k, exchanging $\mathbf{y}_{k,i}$ and $\mathbf{y}_{k+1,i}$ for any subset of i in $s_k \cap s_{k+1}$ does not change the joint distribution of the multiple sample. At the same time the population distributions clearly share some general properties. A DRM with an appropriately rich basis function $\mathbf{q}(y)$, such as $\mathbf{q}(y) = (1, y, y^2, \log y)$ when y takes positive values, will be a good approximation for the population distributions $G_0, G_1, \ldots, G_K$.

6. Simulation experiments.

In this section we present simulation results to illustrate the effectiveness and necessity of the proposed permutation test. We consider the problem of testing for changes in the mean and for changes in quantiles.

6.1. *Data with normal distributions.*

We generate data from the normal model, as described in the last section. The specific model parameters are chosen as follows:

1. The number of occasions/populations is $K + 1 = 5$.
2. The number of units per cluster is either $r = 5$ or $r = 10$.
3. The standard deviations are either $(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 2)$ or $(1, 2, 3)$.
4. The population means vector is one of: $(8, 8, \ldots)$, $(8, 7.6, \ldots)$, and $(8, 7.2, \ldots)$ with unspecified means randomly generated on each repetition as $8 + 0.5N(0, 1)$.
5. The number of clusters (primary units) in each sample is either $n = 36$ or $n = 48$. The rotating sampling plan replaces $m = 6$, or $m = 8$ clusters on each occasion.

The above choices lead to $2 \times 2 \times 4 \times 2 = 32$ distinct settings. Compounded with the permutations, this leads to a computation heavy analysis. We must reduce the computational burden. Since the overall sample size increases either with more clusters or with larger cluster sizes, in the simulation we avoid the option of increasing both sample size and cluster size. These settings cover a broad range of qualitatively different situations:

- With different values of (σ_2, σ_3), we learn the performance of these tests for both relatively weak and strong cross-sectional cluster effects.
- We learn if DRM-based methods benefit from their ability to borrow strength, compared with methods that use only information in the samples from the populations of interest.
- With different cluster sizes or numbers of clusters, we learn about the consistency of these tests. That is, the power increases to 1 when the sample size goes to infinity.

We consider the problem of testing whether the strength distribution has a smaller mean/percentile in the second occasion. When the means vector is set to $(8, 8, \ldots)$, the first two distributions are identical. The rejection rate of a test in this case reflects its size. The rejection rate of any effective test should be higher when the population means vector changes to $(8, 7.6, \ldots)$ or $(8, 7.2, \ldots)$, where the alternative hypothesis holds.

In the simulations we set the nominal level to 5%, the number of repetitions to 1000, and the number of permutations to 501. We recommend a much larger number of permutations in applications. In the simulations the rejection rates are averages of 1000 repetitions. The precision of the individual p-values has little impact on evaluating the overall performance of the permutation test.

TABLE 1
Rejection rate of tests for equal means with clustered normal data. Notice that the asymptotic tests, unlike the permutation tests, have inflated type I errors; see the line corresponding to a means vector of (8.0, 8.0)

	$(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 2)$					$(\sigma_1, \sigma_2, \sigma_3) = (1, 2, 3)$				
	Asymptotic		Permutation			Asymptotic		Permutation		
(μ_0, μ_1)	T	W	T	W	R_μ	T	W	T	W	R_μ
					$K + 1 = 5, r = 5, n = 36$					
(8.0, 8.0)	9.6	8.8	4.3	4.0	4.2	13.1	12.5	5.6	5.6	5.7
(8.0, 7.6)	48.6	46.1	32.0	30.9	32.0	32.4	32.1	15.7	14.8	15.6
(8.0, 7.2)	86.5	85.4	74.9	74.1	74.9	61.8	60.4	37.9	36.3	37.4
					$K + 1 = 5, r = 10, n = 36$					
(8.0, 8.0)	14.7	13.9	4.4	4.3	4.4	21.1	21.1	4.6	4.3	4.7
(8.0, 7.6)	60.5	59.7	36.5	35.8	36.4	45.3	45.0	19.3	19.9	19.3
(8.0, 7.2)	95.0	94.3	82.1	81.9	82.0	72.9	71.9	42.2	40.9	42.2
					$K + 1 = 5, r = 5, n = 48$					
(8.0, 8.0)	8.8	8.2	4.1	4.8	4.0	12.2	12.2	4.9	4.9	4.9
(8.0, 7.6)	54.8	54.4	37.7	38.0	37.6	39.7	39.0	21.3	20.8	21.2
(8.0, 7.2)	93.2	92.0	86.3	85.3	86.3	67.0	65.9	48.6	47.4	48.5

6.1.1. *Population mean.* We include three asymptotic tests: tests based on T and W, as in (2) and (3), with rejection decisions made based on their limiting distributions ignoring cluster structure. We also include three permutation tests, based on T, W, and R_μ, defined in (10) with rejection decisions made based on the p-value evaluated by the proposed permutation approach. For convenience, we refer to the former as asymptotic tests and the latter as permutation tests. Here, H_0 claims that the first two populations have equal means and the test is one-sided, so H_a claims that the second mean is smaller. Table 1 gives the simulation results; the rejection rates are in the Asymptotic and Permutation columns.

The setting with $(\mu_0, \mu_1) = (8.0, 8.0)$ lies on the boundary of the null hypothesis. When $r = 5$ and $n = 36$, two asymptotic tests have much higher than the nominal level rejection rates. The lowest one is 8.2%, and the highest is 21.1% at 5% level. The inflated type I errors of asymptotic tests, based on T and W, are due to ignoring the clustering structure. These rates are much closer to the nominal 5% for the permutation tests. The worst case is a null rejection rate of 5.7%, which is still in the range of the simulation error, given the 1000 repetitions. The comments on type I errors remain whether the cluster size is increased to $r = 10$ or the number of clusters is increased to $n = 48$.

In each of the block of $r = 5$ and $n = 36$, block of $r = 10$ and $n = 36$, or block of $r = 5$ and $n = 48$, the rejection rates increase when (μ_0, μ_1) change from $(8.0, 8.0)$ to $(8.0, 7.6)$ and to $(8.0, 7.2)$. These additional results are as expected and give general support to the validity of our simulation experiments. We also observe that their permutation tests have comparable powers in all cases. This observation extends to the next two simulation experiments. Considering the W test is nonparametric, this is rather surprising.

The general message is that ignoring the cluster structure leads to inflated type I errors for asymptotic tests. Our permutation procedure is an effective way to handle the clustering induced by the rotating sampling plan. The differences between three test statistics are negligible. As this remains the same in other simulation experiments, we recommend the permutation t-test for its computational advantage.

6.1.2. *Percentiles.* Percentiles are of particular interest in many applications, but neither the t-test nor the w-test are designed to detect their changes. In this section we examine permutation tests based on the following statistics introduced earlier:

$$R_{\mathrm{EM}} = \xi_\alpha(\hat{G}_0) - \xi_\alpha(\hat{G}_1),$$
$$R_{\mathrm{EL}} = \xi_\alpha(\check{G}_0) - \xi_\alpha(\check{G}_1),$$

and $R_{n,\xi}$ defined in (9).

Note that R_{EM} is based on the empirical distributions $\hat{G}_0$ and $\hat{G}_1$, R_{EL} is based on the fitted distributions $\check{G}_1$ and $\check{G}_0$ under the DRM with the basis function vector $\mathbf{q}(x) = (1, x, x^2)^\tau$ since we know that the marginal distributions are normal. No corresponding asymptotic theory is available for these statistics, but the permutation-based methods do not rely on asymptotic theory. We denote these tests by R_{EM}, R_{EL}, and $R_{n,\xi}$ in Table 2.

We consider two null hypotheses: the first two distributions have the same fifth percentile or the same 50th percentile. The alternative hypotheses claim that the second distribution has lower percentiles. The rejection rates of these tests are presented in Table 2.

Based on Table 2, we notice that all permutation tests have well-controlled type I errors. The DRM based permutation test has superior power. One will notice that this observation remains true in the next two simulation experiments. Due to computational simplicity with similar power properties, we recommend R_{EL} in applications.

A comparison of the results in the left and right halves of the table shows that the sizes of the permutation tests are not affected by the strength of the random effects. When the random effects are strong, the data contain less information. Hence, the powers on the right half of the table are generally lower. The powers increase when the cluster size increases or the number of clusters increases.

Finally, it is clear that change in a lower percentile is harder to detect than change in the median under a nonparametric model assumption. This explains the power differences for testing the changes in the fifth and 50th percentiles. The simulation results are consistent

TABLE 2

Rejection rate of tests for equal percentiles with clustered normal data. Notice that all tests have accurate type I errors; see the line corresponding to a means vector of (8.0, 8.0)

	$(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 2)$						$(\sigma_1, \sigma_2, \sigma_3) = (1, 2, 3)$					
	5th percentile			50th percentile			5th percentile			50th percentile		
(μ_0, μ_1)	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$
						$K + 1 = 5, r = 5, n = 36$						
(8.0, 8.0)	5.9	4.0	4.4	4.1	4.3	4.3	4.4	4.1	4.2	5.0	5.6	5.6
(8.0, 7.6)	16.9	21.3	20.8	28.5	32.2	31.9	10.3	10.7	11.1	13.9	15.7	15.3
(8.0, 7.2)	39.7	49.0	47.6	65.3	75.2	74.5	20.3	21.4	21.0	31.2	38.2	37.6
						$K + 1 = 5, r = 10, n = 36$						
(8.0, 8.0)	4.8	5.3	5.0	4.5	4.5	4.5	4.6	4.8	4.7	5.0	4.6	4.7
(8.0, 7.6)	19.9	23.6	23.7	29.9	36.3	35.9	15.6	15.1	15.0	19.2	19.6	19.9
(8.0, 7.2)	48.6	57.2	56.7	77.4	81.7	81.7	27.5	30.9	29.9	38.0	42.4	42.3
						$K + 1 = 5, r = 5, n = 48$						
(8.0, 8.0)	4.3	4.2	4.4	4.9	4.1	3.9	5.7	6.2	6.0	4.4	4.7	4.7
(8.0, 7.6)	21.6	25.5	25.0	33.8	37.9	38.0	11.7	13.9	13.7	19.5	20.8	20.4
(8.0, 7.2)	51.8	60.6	59.6	76.9	86.4	86.3	26.8	30.7	30.0	41.3	48.7	48.1

with this intuition, and they also serve as sanity check. One may also conclude from the results in Tables 1 and 2 that detecting changes in the median (50th percentile) is harder than detecting changes in the mean.

6.2. *Data with Gamma distributions*. In this section we generate data from the Gamma model with the parameters chosen as follows:

1. The degrees of freedom vector $(\eta_0, \eta_1, \dots) = (8.0, 8.0, \dots)$, $(8.0, 7.6, \dots)$, or $(8.0, 7.2, \dots)$ with unspecified entries randomly generated in each repetition from $8 + 0.5N(0, 1)$.
2. The degrees of freedom vector is $(\gamma_1, \gamma_2) = (2.0, 1.5)$ or $(2.0, 3.0)$.
3. The scale parameter is $(1.0, 1.0, \dots)$ with unspecified entries randomly generated in each repetition as $1 + 0.2U$, U being a uniform [0, 1] random variable.

Similar considerations apply to this case. The above settings enable us to examine the performance of these tests in a broad range of situations. We use $\mathbf{q}(x) = (1, \log x, x)$ under the DRM assumption. The other specifications are the same as those in the section on normal data.

6.2.1. *Population means*. We now mimic the simulation conducted with the normal data. The null hypothesis is that the first two populations have the same mean, and the alternative is that the second population has a smaller mean. The qualitative findings from the results in Table 3 generally mirror those in Table 1. We again see the inflated type I errors of the asymptotic tests and well-controlled type I errors for the permutation tests. The powers of the permutation tests increase with increased cluster size or increased number of clusters. The powers of the permutation test in the left half are higher than the powers in the right half of the table. It hints the data are more informative when γ values are smaller. We observe that the DRM based permutation tests have comparable powers. Our point is on the success of permutation tests in various situations, and this point remains clear in this simulation.

TABLE 3

Rejection rate of tests for equal means clustered Gamma data. Notice that the asymptotic tests, unlike the permutation tests, have inflated type I errors; see the line corresponding to a means vector of (8.0, 8.0)

	$(\gamma_1, \gamma_2) = (2, 1.5)$					$(\gamma_1, \gamma_2) = (2, 3.0)$				
	Asymptotic		Permutation			Asymptotic		Permutation		
(η_0, η_1)	T	W	T	W	R_μ	T	W	T	W	R_μ
					$K + 1 = 5, r = 5, n = 36$					
(8.0, 8.0)	8.9	8.5	4.4	4.4	4.5	16.1	15.1	4.5	4.5	4.5
(8.0, 7.6)	34.5	33.2	22.8	22.3	22.7	46.1	45.9	19.4	21.4	19.3
(8.0, 7.2)	68.8	68.4	53.7	56.2	53.6	77.9	78.3	49.7	49.7	49.4
					$K + 1 = 5, r = 10, n = 36$					
(8.0, 8.0)	12.6	13.4	4.3	5.0	4.2	19.1	19.3	5.6	5.7	5.6
(8.0, 7.6)	51.9	50.3	29.4	29.6	29.5	46.9	46.7	19.2	20.4	19.2
(8.0, 7.2)	86.1	85.9	68.0	67.6	67.9	76.5	76.7	46.2	47.1	46.2
					$K + 1 = 5, r = 5, n = 48$					
(8.0, 8.0)	14.1	13.2	5.2	5.8	5.3	11.3	11.7	5.3	4.7	5.3
(8.0, 7.6)	53.6	53.9	31.5	31.6	31.6	38.0	39.0	21.5	22.4	21.4
(8.0, 7.2)	92.0	92.3	76.8	77.6	76.6	71.1	71.0	53.4	53.9	53.4

TABLE 4

Rejection rate of tests for equal percentiles with clustered Gamma data. Notice that all tests have accurate type I errors; see the line corresponding to a means vector of (8.0, 8.0)

	$(d_2, d_3) = (2, 1.5)$						$(d_2, d_3) = (2, 3.0)$					
	5th percentile			50th percentile			5th percentile			50th percentile		
(η_0, η_1)	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$
						$K+1=5, r=5, n=36$						
(8.0, 8.0)	5.6	4.1	4.1	5.1	4.1	4.1	4.7	4.0	3.8	4.5	4.4	4.2
(8.0, 7.6)	14.4	16.5	16.4	20.2	22.2	22.3	14.4	15.7	15.3	18.0	21.5	21.4
(8.0, 7.2)	30.6	37.2	36.1	47.0	56.8	56.4	31.6	35.9	35.4	46.3	50.9	51.0
						$K+1=5, r=10, n=36$						
(8.0, 8.0)	4.3	5.0	5.2	4.4	4.7	4.6	5.2	5.9	6.0	4.8	5.7	5.6
(8.0, 7.6)	22.1	18.6	20.0	27.6	27.1	30.3	17.0	12.9	15.1	17.8	19.5	19.6
(8.0, 7.2)	44.8	51.6	51.3	61.7	68.6	68.8	30.2	32.9	32.4	44.0	47.5	47.6
						$K+1=5, r=5, n=48$						
(8.0, 8.0)	6.6	4.6	4.5	5.6	5.2	5.2	5.0	5.3	4.8	4.7	5.2	5.2
(8.0, 7.6)	20.9	24.7	24.1	27.4	33.0	32.9	14.1	17.2	16.9	20.5	21.9	21.7
(8.0, 7.2)	50.0	59.8	59.7	70.1	78.8	78.9	29.7	36.6	35.8	47.7	54.2	54.4

6.2.2. *Percentiles.* We now move to testing hypotheses specifying equal fifth and 50th percentiles for the first two populations in the multiple samples. The alternative hypotheses are one-sided: $\xi(G_0) > \xi(G_1)$. The simulation results are in Table 4. The type I errors of the permutation tests are well controlled.

In all cases the powers of these tests increase when either the cluster size or the number of clusters increases. The differences between the left and right halves do not give much more information, and both support the general claims in this paper. We observe that the DRM-based permutation tests have superior powers. The support to the permutation tests and the use of DRM remain strong.

6.3. *Data from no-name distributions.* In many applications, historical data sets of the same nature are available. This paper recommends DRM to extract latent information from multiple samples to enhance efficiency. When applying the DRM, we must choose a basis function. In simulations we usually generate data from classical distributions, and so an appropriate basis function is readily available. In a parallel research project on a data-adaptive choice of the basis function, we have found that the performance is enhanced under the DRM assumption with $\mathbf{q}(y) = (1, \log|y|, y, y^2)$. We demonstrate this point in this section: the permutation tests still have well-controlled type I errors, and there can be efficiency gains under the DRM assumption.

We generate clustered data with all the features under a rotational sampling plan. We ensure that the population distributions share some latent features, but simple basis functions are not available. Nevertheless, we complete the simulation, as in the last two sections, for EM, EL, and ELR with $\mathbf{q}(y) = (1, \log|y|, y, y^2)$ when the DRM is assumed. Specifically, the data are generated as follows:

1. Form a finite population $\mathcal{P} = \{x_1, x_2, \ldots\}$ having a considerable size, based on data from a real-world application.

2. Randomly generate ϵ_j, $\epsilon_{k,j}$ from the standard uniform distribution. Let $b_{k,j}(x) = \exp\{\sigma_1 \epsilon_j + \sigma_{k,2} \epsilon_{k,j} \log x\}$ for some positive constants σ_1 and $\sigma_{k,2}$.

Sample r values, $(\lambda_{k,j,1},\dots,\lambda_{k,j,r})$, from a Gamma distribution with 20 degrees of freedom and scale parameter 0.05. Randomly draw r values x from $\mathcal{P}$ with probability proportional to $b_{k,j}(x)$ to form a cluster $\mathbf{y}_{k,j}=(\lambda_{k,j,1}x_{k,j,1},\dots,\lambda_{k,j,r}x_{k,j,r})$.

3. Form multiple samples from a rotational sampling plan as

$$\{\mathbf{y}_{0,j}:1\le j\le mN\};\{\mathbf{y}_{1,j}:1+m\le j\le mN+m\};\cdots\cdots.$$

Here are some explanations. Step 1 creates a grand population, and Step 2 uses a biased sample technique to mimic the evolution of the strength distribution over occasions $k=0,1,\dots,K$. The random numbers ϵ_j and $\epsilon_{k,j}$ induce longitudinal and cross-sectional random effects; we use σ_1 and $\sigma_{k,2}$ to adjust the strength of these effects. The λ values are introduced to avoid identical observed values in multiple samples. Since these values have mean 1 and small variance, this does not change the expected value from the case where $\lambda=1$ and is not random.

We simulated data with both $\sigma_1=2$ and $\sigma_1=4$ to examine the influence of the strength of the longitudinal random effects. We successively set $(\sigma_{0,2},\sigma_{1,2},\dots)=(6.0,6.0,\dots)$, $(6.0,4.5,\dots)$, and $(6.0,3.0,\dots)$ with unspecified entries generated in each repetition from $3+2U$, where U is a uniform [0, 1] random variable.

The base case has $K+1=5$, cluster size $r=5$, and number of clusters $n=36$. We then repeated the simulation with an increased cluster size $r=10$ in one setting and with an increased number of clusters $n=48$ in another setting.

The finite population $\mathcal{P}$ in this simulation is formed from data collected by students running experiments in a lab located at FPInnovations, Vancouver (Cai et al. (2016)). It is made up of 825 observed values of the MOR of a specific type of wood product. The sample mean of this data set is 6.57 (1000 psi), and the sample variance is 2.82. The rest of the simulation settings are the same as before. Simulation results are given in Tables 5 and 6.

The results mostly resemble the results in two other simulations. We do notice the type I errors of the permutation tests when $\sigma_1=2$, $r=10$ and $n=36$ are inflated. We take comfort that they are still within margin of error for the simulation $(1.96\times\sqrt{0.05\times0.95/1000})$.

TABLE 5
Rejection rate of tests for equal means with clustered no-name data. Notice that the asymptotic tests, unlike the permutation tests, have inflated type I errors; see the line corresponding to a means vector of (6.0, 6.0)

	$\sigma_1=2$					$\sigma_1=4$				
	Asymptotic		Permutation			Asymptotic		Permutation		
$(\sigma_{0,2},\sigma_{1,2})$	T	W	T	W	R_μ	T	W	T	W	R_μ
	$K+1=5, r=5, n=36$									
(6.0, 6.0)	17.5	17.6	5.2	5.1	5.4	17.5	17.9	4.3	5.0	4.4
(6.0, 4.5)	34.5	32.9	12.3	12.5	12.2	32.7	31.7	11.0	10.2	10.3
(6.0, 3.0)	54.7	55.2	25.8	26.4	25.6	52.0	51.1	23.7	23.7	23.6
	$K+1=5, r=10, n=36$									
(6.0, 6.0)	26.0	26.1	6.7	7.0	6.7	26.0	25.8	4.4	4.9	3.9
(6.0, 4.5)	45.4	45.1	12.7	12.5	12.3	42.5	42.5	12.9	12.9	12.6
(6.0, 3.0)	65.1	63.7	26.1	25.1	26.1	61.4	60.6	24.3	23.2	24.2
	$K+1=5, r=5, n=48$									
(6.0, 6.0)	19.9	19.2	5.1	4.7	5.0	19.6	20.7	4.8	5.3	4.8
(6.0, 4.5)	40.8	40.2	16.6	15.6	16.5	35.1	34.7	13.2	12.9	13.0
(6.0, 3.0)	62.1	62.1	33.2	32.7	33.1	56.6	56.3	31.0	29.7	30.9

TABLE 6
Rejection rate of tests for equal percentiles with clustered no-name data. Notice that all tests have accurate type I errors; *see the line corresponding to a means vector of* (6.0, 6.0)

	$\sigma_1 = 2$						$\sigma_1 = 4$					
	5th percentile			50th percentile			5th percentile			50th percentile		
$(\sigma_{0,2}, \sigma_{1,2})$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$
$K + 1 = 5, r = 5, n = 36$												
(6.0, 6.0)	5.2	4.8	4.9	5.1	5.3	5.3	4.6	5.4	5.4	4.8	4.9	4.5
(6.0, 4.5)	10.1	10.0	9.6	12.2	12.4	12.3	9.3	8.7	8.3	10.3	10.8	10.8
(6.0, 3.0)	18.8	19.5	19.6	25.0	26.3	26.3	17.3	16.8	16.8	22.3	21.1	21.6
$K + 1 = 5, r = 10, n = 36$												
(6.0, 6.0)	6.3	5.4	5.5	7.0	6.5	6.5	5.7	5.3	5.1	4.4	5.1	5.0
(6.0, 4.5)	10.4	9.9	9.5	12.5	12.6	12.5	11.1	10.2	10.3	11.8	11.9	11.6
(6.0, 3.0)	19.3	21.9	21.1	24.7	24.5	24.5	20.3	19.4	19.3	22.6	23.5	23.5
$K + 1 = 5, r = 5, n = 48$												
(6.0, 6.0)	5.6	5.8	5.4	4.2	4.6	4.5	5.2	4.5	4.6	5.8	5.6	5.7
(6.0, 4.5)	11.6	11.8	11.6	14.2	14.5	14.5	11.8	11.6	11.6	11.9	13.3	13.1
(6.0, 3.0)	20.1	22.1	21.5	30.6	31.2	30.9	18.9	20.8	20.2	25.9	28.7	28.4

There could be some other causes. We believe that this is not so important an issue and do not investigate further here. In all other cases, for mean or for percentiles, the type I errors of permutation tests fluctuates around the nominal level in a small range.

The powers of all the tests increase with the cluster size and with the number of clusters, though not as markedly. Moreover, even when the data are from distributions that do not fully conform to the DRM specifications, the inferences made under the DRM assumptions remain valid; the power comparison to EM remains favorable, although not decisively so. Thus, we recommend the use of the DRM without reservation.

6.4. *The histogram of typical generated data in simulations.* The simulation examples were created based on our experience with the forestry data, though not from a rotating sampling plan. We include a plot of histograms of two typical samples of size 1800 from the normal and Gamma distributions, employed in the simulation, and the histogram of the MOR data set of size 825; see Figure 1. A larger sample size than these in the simulation experiment

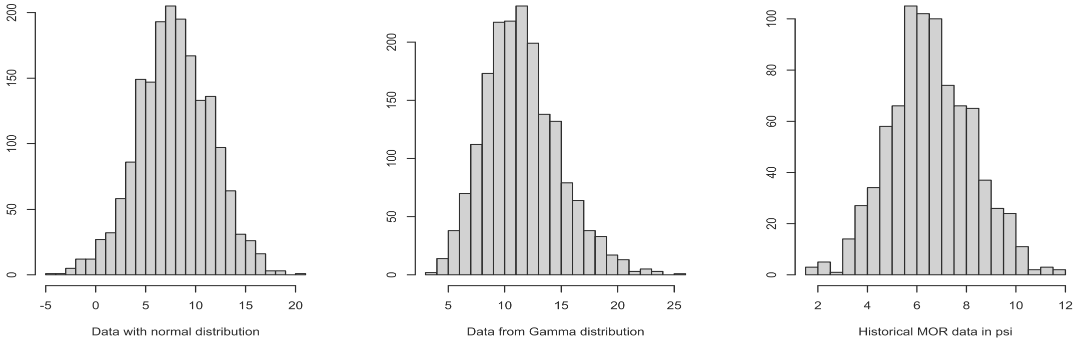

FIG. 1. *Histograms of the simulated and historical data sets. The first two plots are simulated data, as in Sections* 6.1 *and* 6.2. *The third plot are data used in Section* 6.3.

is chosen to depict the distributions more precisely. We notice that three data sets have different scales but similar shapes. Hence, the conclusions derived from the simulation experiments are useful references for real-world applications.

7. A real-world application. The National Lumber Grading Association of Canada (http://nlga.org/en/) carries out a rotating sampling plan for monitoring the quality of lumber for several different species groups. The samples are selected from across Canada to represent the global lumber population. Initially, there were three sampling periods per year with the year starting and ending in the fall. The goal was, in particular, to look for seasonal patterns as well as trends over time. Data were collected on the mechanical strength of two lots of five pieces each from each mill to capture possible effects due to time-of-day. The samples provided came from 16 different sampling occasions over eight years that started in the fall of 2010 and ended in the winter of 2018. There were three sampling periods (F, W and S representing Fall, Winter and Summer, respectively) from 24 mills per year for the first four years. Since no seasonal effects were found, there was one sampling period from 30 mills per year for the last four years (2015–2018). Each year four mills were removed, and four new ones were added. The result was a less costly plan, compared to taking cross sectional samples that test 360 pieces per year from 36 mills.

The histograms of the modulus of rupture (MOR) from the Ramp-up (Ru) and years 2 and 3 of regular monitoring are given in Figure 2. The sample means, the fifth and 50th percentiles are in Table 7 for the nine period samples within the three years.

To illustrate the application of proposed tests, we analyze the data collected on nine occasions during the Ramp-up and years 2 and 3 of regular monitoring. We test whether there is a significant change between each pair of years in the mean and in the fifth and 50th percentiles. We carried out $M = 10001$ permutations of the matching mills between the years. The margin of error, due to random permutation, is below 0.4% if the true type I error is around 5%. We included asymptotic T, W tests, permutation R_μ tests for population means, and R_{EM}, R_{EL} and R_ξ for the fifth and 50th percentiles in this analysis. We used $\mathbf{q}(y) = (1, \log|y|, y, y^2)$ in the EL-DRM approaches. The p-values of these tests for the three comparisons are given in Table 8.

The results in Table 8 indicate that the Ramp-up (Ru) strength distribution is significantly higher than those of years 2 and 3 in all three respects. The difference between years 2 and 3 is on the boundary of the significance. We believe there is some start-up effect, though this issue is beyond the scope of this paper.

Regarding the change in population means, the mean strength of the Ru population is clearly higher. Yet the extremely low p-values of the asymptotic T and W tests seem to

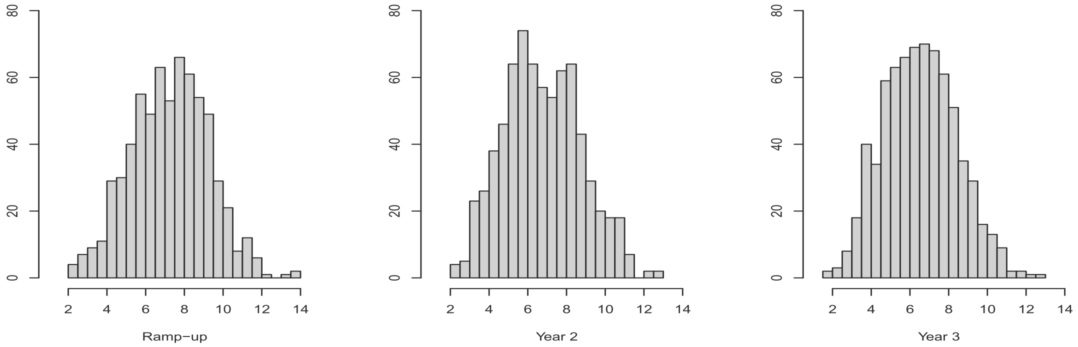

FIG. 2. *MOR for the Ramp-up* (*Ru*), *years* 2 *and* 3 *monitoring samples in* psi. *Data were derived from data collected in a rotating sampling plan by the National Lumber Grading Association of Canada.*

TABLE 7

Period sample summary statistics for MOR in psi *collected in a rotating sampling plan by the National Lumber Grading Association of Canada*

Period	mean	variance	5th percentile	50th percentile
Ru_F	7.07	3.59	4.15	6.99
Ru_S	7.38	4.21	4.14	7.42
Ru_W	7.32	4.20	3.62	7.49
yr2_F	6.73	3.97	3.86	6.49
yr2_S	6.59	4.14	3.40	6.43
yr2_W	6.98	3.77	3.93	6.98
yr3_F	6.59	3.34	3.82	6.50
yr3_S	6.68	3.85	3.47	6.69
yr3_W	6.31	3.48	3.32	6.28

have exaggerated the significance. The proposed permutation R_μ test also rejects the equal mean hypothesis decisively yet with a much more sensible level of significance. The mean difference between years 2 and 3 is noticeable but mild, calling a p-value of around 5%. The proposed permutation R_{EM} has a matching p-value.

Regarding the change in the fifth and 50th percentiles between Ru and the other two populations, all three permutation tests reject the equality assumption. The results for testing the differences between year 2 and year 3 are interesting. The p-values of the fully nonparametric permutation test R_{EM} are 0.31 and 0.16. It fails to build a case against the null hypotheses. Neither R_{EL} nor $R_{n,\xi}$ have a strong enough case at 5% nominal level. However, a closer examination leads to some interesting details. Both R_{EL} and $R_{n,\xi}$ tests build on a semiparametric DRM assumption. They are known to be more powerful when the DRM is suitable, especially for $R_{n,\xi}$ which is of likelihood ratio type. Evidently, both p-values of $R_{n,\xi}$ tests are close to the 5% nominal level for both fifth and 50th percentiles, and the p-value of R_{EL} for the 50th percentile is close to 5%.

In conclusion, the results in Table 8 are in complete agreement with our theory: the asymptotic tests are to be avoided because they likely produce unrealistic small p-values, the proposed permutation tests are reliable, and the DRM can enhance the power of the permutation tests.

8. Summary and discussion.

This paper develops novel permutation tests for comparing/monitoring changes in mean, percentiles, or other parameters of an evolving population based on data from a rotating sampling plan. The simulation and theoretical analyses presented in this paper show that the methods proposed in this paper are valid and effective. The approaches ignoring cluster effects often fail to control the type I error. The use of DRM improves the power of the tests.

TABLE 8

The p-values of the tests for all three comparisons between Ramp-up (*Ru*), *years* 2 *and* 3 *of the National Lumber Grading Association of Canada data*

Parameter	Mean			5th percentile			50th percentile		
Test	T	W	R_μ	R_{EM}	R_{EL}	$R_{n,\xi}$	R_{EM}	R_{EL}	$R_{n,\xi}$
Ru vs yr2	2.8e−06	1.1e−06	0.0015	0.038	0.010	0.010	0.0006	0.0015	0.0015
Ru vs yr3	3.7e−12	4.0e−12	0.0006	0.021	0.021	0.020	0.0005	0.0007	0.0008
yr2 vs yr3	1.1e−02	2.2e−02	0.0460	0.310	0.340	0.055	0.1600	0.0550	0.0560

The rotating sampling plan could be implemented so that mills are sampled with the inclusion probability proportional to the volume of their produce. Some mills may drop out and new mills may enter the population. The cluster sizes may fluctuate slightly between occasions. Real world applications are more complex than we depicted. The permutation tests investigated in this paper provide a starting point in these applications. We anticipate the proposed approaches can be adapted in many applications nevertheless. Our results can be tailored to real-world applications that do not fully fit into the current frame.

An anonymous referee drew our attention to some approaches in the literature that we can amend to handle the cluster data from rotating sampling plan. We studied a modified t-statistic, following the idea in Berg, Cecere and Ghosh (2014), and the linearization approach of Francisco and Fuller (1991) for quantiles. Both statistics are asymptotically normal, therefore, permitting tests based on their limiting distributions. However, in our simulation studies their asymptotic tests were repeatedly found to have inflated type I errors. We do not go after the cause as it is not the focus of this paper, but this line of thinking holds some potential.

APPENDIX: COMPUTATIONAL ISSUES

The numerical implementation of most of our proposed permutation tests is straightforward. The implementation of the ELR is conceptually simple but involves some tedious steps. To compute R_n defined following (8), we must solve the optimization problem $\sup_{\boldsymbol{\theta}} \ell_n^{\mathrm{CC}}(\boldsymbol{\theta})$. The constraints in the definition of $\ell_n^{\mathrm{CC}}(\boldsymbol{\theta})$ can be rewritten as

$$\sum_{k,i,u} p_{k,i,u}\big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\} - 1\big] = 0, \quad s = 0, \ldots, K;$$

$$\sum_{k,i,u} p_{k,i,u}\big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\big] = 0, \quad s = 0, 1.$$

Given $\boldsymbol{\theta}$, there always exists a $\mathbf{p}$, the vector formed by $\{p_{k,i,u}\}$, that solves the above equation system, provided vector $\mathbf{0}$ is an interior point of the convex hull of

$$\big\{\big(\big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\} - 1\big]_{s=0}^K, \big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\big]_{s=0}^1\big) : \\ k = 0, \ldots, K, i = 1, \ldots, n; u = 1, 2, \ldots, r\big\}.$$

The convex hull condition is universal and well known in the literature of empirical likelihood (Owen (2001)). If this convex hull does not contain $\mathbf{0}$, the adjusted empirical likelihood approach of Chen, Variyath and Abraham (2008) or the selfconcordance empirical likelihood of Owen (2013) may be used. In the current application we can show that, as long as $\hat{\xi}_\alpha$ does not fall outside either interval $(\min_{i,u} y_{s,i,u}, \max_{i,u} y_{s,i,u})$ for $s = 0, 1$, there always exist some $\boldsymbol{\theta}$ values such that the convex hull condition is satisfied. In applications, if the joint sample percentile $\hat{\xi}_\alpha$ has a nonextreme α value outside of one of these intervals, it is a strong indication that the population has significantly changed in some direction. It is not urgent to look into such rare possibilities.

Once the existence of a solution is ensured, the optimization problem can be solved via the Lagrange multiplier method. Because of the nice properties of the DRM, we can find a simpler set of equations that can be solved by the function *multiroot* in the R-package *rootSolve* (Soetaert (2009), Soetaert and Herman (2009)). The details are as follows.

We first define a Lagrangian function,

$$g(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{p}) = \sum_{k,i,u} p_{k,i,u} + \sum_{k,i,u} \boldsymbol{\theta}_k^\tau \mathbf{q}(y_{k,i,u}) \\ - \sum_{s=0}^{K} t_s\Big[\sum_{k,i,u} p_{k,i,u} \exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u)}\} - 1\Big]$$

$$- \sum_{s=0}^{1} \lambda_s \sum_{k,i,u} p_{k,i,u}\big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\big]$$

with $\mathbf{t}$ and $\boldsymbol{\lambda}$ of length $K + 1$ and two vectors of Lagrange multipliers.

The maximum of $\ell_n^{\mathrm{CC}}(\boldsymbol{\theta})$ is attained at the $\boldsymbol{\theta}$ value that solves

$$\frac{g(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{p})}{\partial \mathbf{t}} = 0; \qquad \frac{g(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{p})}{\partial \boldsymbol{\lambda}} = 0; \qquad \frac{g(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{p})}{\partial \boldsymbol{\theta}} = 0; \qquad \frac{g(\mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{p})}{\partial \mathbf{p}} = 0$$

together with some values of $\mathbf{t}$, $\boldsymbol{\lambda}$, and $\mathbf{p}$.

Some algebra shows that the solution in $\mathbf{t}$ is given by $t_s = nr$, where nr is the total number of observations in each population in the rotating sampling plan, $s = 0, 1, \dots, K$, and the elements of $\mathbf{p}$ satisfy

$$\begin{aligned} p_{k,i,u}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \Bigg\{ & (nr) \sum_{s=0}^{K} \exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\} \\ & + nr(K+1) \sum_{s=0}^{1} \lambda_s \big[\exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\big] \Bigg\}^{-1}. \end{aligned}$$

Substituting the above expression into the Lagrangian equations, we obtain three sets of vector equations for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$,

$$\sum_{k,i,u} p_{k,i,u}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\} = 1;$$

$$\sum_{k,i,u} p_{k,i,u}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\{\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\} = 0;$$

$$\sum_{k,i,u} p_{k,i,u}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \mathbf{q}(y_{k,i,u}) \exp\{\boldsymbol{\theta}_s^\tau \mathbf{q}(y_{k,i,u})\}\big[1 + \lambda_s\{\mathbb{1}(y_{k,i,u} \le \hat{\xi}_\alpha) - \alpha\}\big] = (nr)^{-1} \sum_{i,u} \mathbf{q}(y_{k,i,u}),$$

with $s = 0, 1, \dots, K$ for the first equation, $s = 0, 1$ for the second equation, and $s = 0, 1, \dots, K$ again for the third equation. Note that the third equation takes vector values because $\mathbf{q}(\cdot)$ is vector-valued.

Furthermore, the derivatives of the above equations (more precisely, the related functions) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ can be found without technical difficulties. With this information provided to *multiroot*, solving for $\boldsymbol{\theta}$ in the simulation experiment was quite smooth. Of 1000 repetitions, the R-function failed to find the solution about 10 times when hypothesis testing for the fifth percentile in the third example, and it succeeded in all the other cases. Because this is a low failure rate, we did not try to determine the exact cause and instead dropped these cases from the final tally. We did, however, increase the number of repetitions in the simulation so that the number of successful repetitions in every setting was at least 1000.

Acknowledgments. We are indebted to Mr. Conroy Lum. We thank the Associate Editor and the referees for constructive comments that made the paper better focused.

Funding. The research was funded by FPInnovations and a Collaborative Research and Development Grant from the Natural Sciences and Research Council of Canada.

Dr. Liu's research was supported by the National Natural Science Foundation of China (11771144), the State Key Programs of the National Natural Science Foundation of China (71931004, 32030063), the development fund for Shanghai talents, and the 111 project (B14019).

REFERENCES

ANDERSON, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66** 17–26. MR0529143 https://doi.org/10.1093/biomet/66.1.17

ASTM (2006). Standard practice for establishing allowable properties for visually-graded dimension lumber. American Society for Testing and Materials, West Conshohocken, PA.

BERG, E., CECERE, W. and GHOSH, M. (2014). Small area estimation for county-level farmland cash rental rates. *Journal of Survey Statistics and Methodology* **2** 1–37.

CAI, S., CHEN, J. and ZIDEK, J. V. (2017). Hypothesis testing in the presence of multiple samples under density ratio models. *Statist. Sinica* **27** 761–783. MR3674695

CAI, Y., CAI, J., CHEN, J., GOLCHI, S., GUAN, M., KARIM, M. E., LIU, Y., TOMAL, J., XIONG, C. et al. (2016). An empirical experiment to assess the relationship between the tensile and bending strengths of lumber. The University of British Columbia, Department of Statistics, Technical Report # 276.

CHEN, J. and LIU, Y. (2013). Quantile and quantile-function estimations under density ratio model. *Ann. Statist.* **41** 1669–1692. MR3113825 https://doi.org/10.1214/13-AOS1129

CHEN, J., VARIYATH, A. M. and ABRAHAM, B. (2008). Adjusted empirical likelihood and its properties. *J. Comput. Graph. Statist.* **17** 426–443. MR2439967 https://doi.org/10.1198/106186008X321068

CHEN, J., LI, P., LIU, Y. and ZIDEK, J. V. (2021). Composite empirical likelihood for multisample clustered data. *J. Nonparametr. Stat.* **33** 60–81. MR4261898 https://doi.org/10.1080/10485252.2021.1914337

DATTA, S. and SATTEN, G. A. (2005). Rank-sum tests for clustered data. *J. Amer. Statist. Assoc.* **100** 908–915. MR2201018 https://doi.org/10.1198/016214504000001583

DATTA, S. and SATTEN, G. A. (2008). A signed-rank test for clustered data. *Biometrics* **64** 501–507, 667. MR2432420 https://doi.org/10.1111/j.1541-0420.2007.00923.x

FRANCISCO, C. A. and FULLER, W. A. (1991). Quantile estimation with a complex survey design. *Ann. Statist.* **19** 454–469. MR1091862 https://doi.org/10.1214/aos/1176347993

HEMERIK, J. and GOEMAN, J. (2018). Exact testing with random permutations. *TEST* **27** 811–825. MR3878362 https://doi.org/10.1007/s11749-017-0571-1

HEMERIK, J., SOLARI, A. and GOEMAN, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649. MR3992394 https://doi.org/10.1093/biomet/asz021

KARNA, J. P. and NATH, D. C. (2015). Rotationn sampling: Introduction and review of recent developments. *J. Assam Sci. Soc.* **56** 90–111.

KEZIOU, A. and LEONI-AUBIN, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *J. Statist. Plann. Inference* **138** 915–928. MR2384498 https://doi.org/10.1016/j.jspi.2007.02.009

LINDSAY, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (*Ithaca*, *NY*, 1987). *Contemp. Math.* **80** 221–239. Amer. Math. Soc., Providence, RI. MR0999014 https://doi.org/10.1090/conm/080/999014

NIJMAN, T., VERBEEK, M. and VAN SOEST, A. (1991). The efficiency of rotating-panel designs in an analysis-of-variance model. *J. Econometrics* **49** 373–399. MR1129125 https://doi.org/10.1016/0304-4076(91)90003-V

OWEN, A. (2001). *Empirical Likelihood*. CRC Press/CRC, New York.

OWEN, A. B. (2013). Self-concordance for empirical likelihood. *Canad. J. Statist.* **41** 387–397. MR3101590 https://doi.org/10.1002/cjs.11183

PARK, Y. S., CHOI, J. W. and KIM, K. W. (2007). A balanced multi-level rotation sampling design and its efficient composite estimators. *J. Statist. Plann. Inference* **137** 594–610. MR2298960 https://doi.org/10.1016/j.jspi.2005.12.007

PESARIN, F. and SALMASO, L. (2010). *Permutation Tests for Complex Data*: *Theory*, *Applications and Software*. John Wiley & Sons.

PFEFFERMANN, D. and SVERCHKOV, M. (2009). Inference under informative sampling. In *Handbook of Statistics* **29** 455–487. Elsevier.

QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619

QIN, J. and ZHANG, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84** 609–618. MR1603924 https://doi.org/10.1093/biomet/84.3.609

RAO, J. N. K. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79** 811–822. MR1209480 https://doi.org/10.1093/biomet/79.4.811

ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62** 185–192. MR2226572 https://doi.org/10.1111/j.1541-0420.2005.00389.x

SOETAERT, K. (2009). rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. R package 1.6.

SOETAERT, K. and HERMAN, P. M. J. (2009). *A Practical Guide to Ecological Modelling*: *Using R as a Simulation Platform*. Springer, New York. MR2492334 https://doi.org/10.1007/978-1-4020-8624-3

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

VERRILL, S., KRETSCHMANN, D. E. and EVANS, J. W. (2015). Simulations of strength property monitoring tests. Unpublished manuscript. Forest Products Laboratory, Madison, Wisconsin. Available at http://www1.fpl.fs.fed.us/monit.pdf.

ZIDEK, J. V. and LUM, C. (2018). Statistical challenges in assessing the engineering properties of forest products. *Annu. Rev. Stat. Appl.* **5** 237–267. MR3774747 https://doi.org/10.1146/annurev-statistics-041715-033633