Permutation tests under a rotating sampling plan
A problem in long term monitoring program

Jiahua Chen, Yukun Liu and James Zidek

Dept of Stat, UBC

Feb 19, 2020
Table of contents

1. Introduction

2. Previous and current contributions

3. Illustration
Problem description

- Lumber is manufactured from the trees of a forest.
- Trees are sawn in an optimal way to get pieces of lumber, classified into grades for engineering applications.
- Each grade has a published design value (DV) for each type of strength such as MOR and MOE.
Design values

- The DV is a specified quantile of the strength distribution, commonly the median or the fifth percentile.
- 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.
Monitoring of wood quality

• The quality of wood may vary due to many factors: changing climate, forest fires, insect infestations, and processing techniques.

• There is therefore a concern on whether these DVs are maintained over the years.

• This concern leads to the need of rigorous lumber-quality monitoring procedures.

• To address such concerns, we should collect data over years and test the hypothesis that the DVs are maintained.
Possible approaches

- The data may be collected by a cross-sectional plan, stratified-by-region with primary sampling units (PSUs) within a region, and secondary sampling units (SSUs) from the PSUs.
- This is the typical plan implemented in some parts of US: in the end, they collect 300 measures of module of elasticity (MOE) yearly.
- In statistical language, we wish to test if either one of the mean, median or 5th percentile has decreased this year from the last year.
- Rightfully or not, the paired t-test, the Wilcoxon rank sum test may be used.
Issues of these tests

- Either paired $t$-test, Wilcoxon rank sum test, or two-sample $t$, aims for changes in the median or percentiles.
- The data are clustered: observations on SSUs from the same PSU are not independent.
- Ignoring cluster structure tends to under-estimate the uncertainty, leads to inflated type I errors of standard tests:
  - wrongfully set-off the alarm more often than anticipated.
Solution for independent clustered samples

- Chen et al. (2018) considered the situation where the data are clustered, but the samples from different years are independent.
- The goal is the same: develop valid lumber-quality monitoring procedures.
- They proposed test statistics based on EL under the DRM.
- The technical issue is to approximate the sample distribution of the proposed test statistics when the null hypothesis is true.
- Their main contribution: turn the test problem into an equivalent confidence interval problem.
- Use cluster-based resampling scheme to construct confidence intervals.
Rotating sampling plans are often used in applications, and it is the one adopted in forestry applications in Canada.

In the beginning, a number of PSUs (mills) are randomly selected from the population.

Suppose the number of PSUs is 30 in the first year. The rotating sampling plan may replace 6 PSUs in the original sample with 6 new PSUs randomly selected from the rest of the population.

Repeat the above scheme for each of the next 5 years. In the end, all PSUs in the original sample will be rotated out.
Longitudinal and cross-sectional random effects

- The measurements on the SSUs from the same PSUs are still dependent.
- Under the rotating sampling plan, year one and year two share 24 SSUs. In forestry terms, 24 mills contribute samples in both years.
- The measurements from different years but the same mill, are intuitively dependent, likely positively correlated in some way.
- Hence, the resampling approach of Chen et al. (2018) is not suitable for rotating sampling plan.
  - Remark: resampling technique is used to approximate sampling distributions of test statistics.
Permutation test

- In abstract, suppose the data set can be presented as \{z_1, z_2, \ldots, z_n\}, and a statistic is denoted as

\[ T(z_1, z_2, \ldots, z_n). \]

- Let \( \pi(\cdot) \) be a permutation operation:

\[ \{\pi(1), \pi(2), \ldots, \pi(n)\} = \{1, 2, \ldots, n\} \]

- In many situations, the sample distributions of

\[ T(z_1, z_2, \ldots, z_n) \text{ and } T(z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(n)}) \]

are the same, for a class of permutations.

- For instance, when \( T(z_1, z_2, \ldots, z_n) = \bar{z}_n \), the sample distributions of

\[ T(z_1, z_2, \ldots, z_n) \text{ and } T(z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(n)}) \]

are the same, for any permutation.
Two-sample problem

- Suppose, \( \{z_1, z_2, \ldots, z_{m+n}\} \) is a pooled sample in two-sample problem (both iid):

  \[
  \{z_1, z_2, \ldots, z_{m+n}\} = \{x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n\}.
  \]

- Let \( T \) be the commonly used \( t \)-statistic:

  \[
  T = \frac{\bar{y}_n - \bar{x}_m}{\sqrt{n^{-1}s^2_y + m^{-1}s^2_x}}.
  \]

  - If two populations have the same distribution, then \( T(z_1, z_2, \ldots, z_{m+n}) \) and \( T(z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(m+n)}) \) have the same distribution for any permutation \( \pi(\cdot) \).
  - If two populations have different distributions, then they have different distributions in general.
Two-sample problem

- Let two population distribution be called $F(x)$ and $G(y)$.
- Suppose $G(y) = F(x - \delta)$ for some large $\delta$ value. If so, the observed value of this $t$ statistic has a high probability being very large.
- In comparison, after a permutation $\pi(\cdot)$,

$$T(z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(m+n)})$$

mostly likely becomes smaller.
- Hence, when most $T(z_{\pi(1)}, \ldots, z_{\pi(m+n)})$ values are smaller than $T(z_1, z_2, \ldots, z_{m+n})$, we suspect that $G(y) = F(x - \delta)$ for some meaningfully size $\delta$.
- The permutation test simply calculate enough $T(z_{\pi(1)}, \ldots, z_{\pi(m+n)})$ values and find the percentage of times when it is larger than $T(z_1, z_2, \ldots, z_{m+n})$.
- When this percentage is very low, say below 5%, reject the null hypothesis.
Back to rotating sampling design

- Let $x_{k,i}$ be the observed values on the $i$th cluster in year $k$.
- Under rotating sampling plan, we may image the observations are as follows:
  \[
  \begin{align*}
  x_{1,1} \ldots x_{1,6} & \\
  x_{1,7} \ldots x_{1,12} & \quad x_{2,7} \ldots x_{2,12} \\
  x_{1,13} \ldots x_{1,18} & \quad x_{2,13} \ldots x_{2,18} \quad x_{3,13} \ldots x_{3,18} \\
  x_{1,19} \ldots x_{1,24} & \quad x_{2,19} \ldots x_{2,24} \quad x_{3,19} \ldots x_{3,24} \\
  x_{1,25} \ldots x_{1,30} & \quad x_{2,25} \ldots x_{2,30} \quad x_{3,25} \ldots x_{3,30} \\
  x_{2,31} \ldots x_{2,36} & \quad x_{3,31} \ldots x_{3,36} \\
  \ldots & \quad x_{3,37} \ldots x_{3,42} 
  \end{align*}
  \]
Invariance

- Suppose the population distributions are the same for years 1 and 2.
- Then the following data have the “same distribution” as the data in the last slide:
  \[ x_{1,1}, \ldots, x_{1,6} \]
  \[ x_{2,7}, \ldots, x_{2,12} \]
  \[ x_{1,13}, \ldots, x_{1,18} \]
  \[ x_{1,19}, \ldots, x_{1,24} \]
  \[ x_{1,25}, \ldots, x_{1,30} \]
  \[ x_{2,31}, \ldots, x_{2,36} \]
  \[ x_{3,37}, \ldots, x_{3,42} \]
  \[ x_{3,1}, \ldots, x_{1,12} \]
  \[ x_{2,13}, \ldots, x_{2,18} \]
  \[ x_{2,19}, \ldots, x_{2,24} \]
  \[ x_{2,25}, \ldots, x_{2,30} \]
  \[ x_{3,31}, \ldots, x_{3,36} \]
  \[ x_{3,37}, \ldots, x_{3,42} \]
  \[ x_{3,13}, \ldots, x_{3,18} \]
  \[ x_{3,19}, \ldots, x_{3,24} \]
  \[ x_{3,25}, \ldots, x_{3,30} \]
  \[ x_{3,31}, \ldots, x_{3,36} \]
  \[ x_{3,37}, \ldots, x_{3,42} \]
Invariance

- Suppose the population distributions are the same for years 1 and 2.
- Then the following data have the “same distribution” as the data in the last slide:
  \[
  x_{1,1} \cdots x_{1,6} \quad x_{1,7} \cdots x_{1,12} \\
  x_{2,7} \cdots x_{2,12} \\
  x_{1,13} \cdots x_{1,18} \quad x_{2,13} \cdots x_{2,18} \quad x_{3,13} \cdots x_{3,18} \\
  x_{1,19} \cdots x_{1,24} \quad x_{2,19} \cdots x_{2,24} \quad x_{3,19} \cdots x_{3,24} \\
  x_{1,25} \cdots x_{1,30} \quad x_{2,25} \cdots x_{2,30} \quad x_{3,25} \cdots x_{3,30} \\
  x_{2,31} \cdots x_{2,36} \quad x_{3,31} \cdots x_{3,36} \quad x_{3,37} \cdots x_{3,42} \\
  \ldots
  \]
  - We permuted positions of \( x_{1,7} \cdots x_{1,12} \) and \( x_{2,7} \cdots x_{2,12} \).
Invariance

- The invariance property presented in the last slide is valid more broadly:
  - we may exchange only $x_{1,7}$ and $x_{2,7}$, or only $x_{1,9}$ and $x_{2,9}$, or both of them;
  - we may exchange $x_{1,7}$ and $x_{2,7}$ and $x_{1,24}$ and $x_{2,24}$.
  - we may exchange any number of pairs $x_{1,j}$ and $x_{2,j}$.

- In summary, permuting any number of pairs of clusters of the same PSUs but these two years will not alter the joint distribution
  - under the null hypothesis that population distributions are the same for years 1 and 2.
Permutation test

- Years 1 and 2 share 28 SSUs under the rotating sampling plan in our forestry example.
- There are one million+ ($2^{28}$) of different permutations that do not change the null joint distribution.
- Let $T$ be a statistic based on this data set, then its null sample distribution remains the same after such permutations.
- In other words, the observed $T$ is an ordinary value in a million+ possibilities under the null hypothesis.
- If we arbitrarily identify 5% of these values to form a rejection region, the type I error is 5%.
Good permutation test

• The distribution of $T$ is different with different permutations in general under the alternative hypothesis.
• Suppose a $T$ is such that its sample distribution is “stochastically increased” when the alternative hypothesis is true.
• We may identify the percentage of the permuted $T$ values larger than the original $T$ value.
• A $T$ based permutation test rejects the null hypothesis when this percentage is low.
• With a good choice of $T$, the null will be rejected with higher probability when the alternative is true.
How do we choose $T$?

- Recall that we wish to detect changes in the population distribution in the direction that mean, median or percentile is lower in year 2 compared to year 1.
- The two-sample t-statistic (not t-test) and the Wilcoxon rank sum statistic (not w-test) both have the required properties in general.
- Let us use simulation to demonstrate their performance either utilizing or not utilizing the permutation scheme.
- See simulation results in the next slide, though I will skip detailed simulation settings here.
## Rejection rates when data are clustered normal

| $(\mu_0, \mu_1)$ | $(\sigma_1, \sigma_2, \sigma_3) = (1, 1, 2)$ | $(\sigma_1, \sigma_2, \sigma_3) = (1, 2, 3)$ |  |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | t.test | w.test | t.test | w.test | Non-Perm | Permutation | t.test | w.test | t.test | w.test | Non-Perm | Permutation |
| $(8.0, 8.4)$    | 0.3     | 0.6     | 0.1     | 0.1     | 4.4     | 4.8         | 1.7     | 1.6     | 4.4     | 4.8         | 1.7     | 1.6         |
| $(8.0, 8.0)$    | 9.0     | 9.6     | 5.1     | 4.9     | 14.2    | 13.8        | 6.2     | 6.4     | 14.2    | 13.8        | 6.2     | 6.4         |
| $(8.0, 7.6)$    | 45.8    | 44.3    | 28.8    | 28.9    | 31.6    | 30.9        | 16.5    | 16.3    | 31.6    | 30.9        | 16.5    | 16.3        |
| $(8.0, 7.2)$    | 85.2    | 84.2    | 73.4    | 72.7    | 61.4    | 60.3        | 37.5    | 37.6    | 61.4    | 60.3        | 37.5    | 37.6        |
|                 | $K + 1 = 5$, $r = 5$, $n = 36$ | $K + 1 = 5$, $r = 5$, $n = 36$ |  |
| $(8.0, 8.0)$    | 18.6    | 17.4    | 5.1     | 5.3     | 21.8    | 20.9        | 4.4     | 4.0     | 21.8    | 20.9        | 4.4     | 4.0         |
| $(8.0, 7.6)$    | 64.3    | 62.4    | 39.8    | 39.4    | 46.6    | 45.7        | 18.8    | 18.8    | 46.6    | 45.7        | 18.8    | 18.8        |
| $(8.0, 7.2)$    | 95.8    | 95.6    | 84.6    | 82.7    | 75.2    | 74.7        | 45.0    | 44.9    | 75.2    | 74.7        | 45.0    | 44.9        |
|                 | $K + 1 = 5$, $r = 10$, $n = 36$ | $K + 1 = 5$, $r = 5$, $n = 36$ |  |
| $(8.0, 8.0)$    | 9.9     | 8.8     | 4.5     | 5.0     | 13.3    | 13.5        | 5.3     | 5.3     | 13.3    | 13.5        | 5.3     | 5.3         |
| $(8.0, 7.6)$    | 56.3    | 55.2    | 39.9    | 38.5    | 38.2    | 36.6        | 20.2    | 19.7    | 38.2    | 36.6        | 20.2    | 19.7        |
| $(8.0, 7.2)$    | 94.1    | 93.7    | 87.4    | 86.0    | 69.7    | 68.4        | 48.2    | 46.7    | 69.7    | 68.4        | 48.2    | 46.7        |
|                 | $K + 1 = 5$, $r = 5$, $n = 48$ | $K + 1 = 5$, $r = 5$, $n = 48$ |  |
Let $\hat{\xi}_{1,p}$ and $\hat{\xi}_{2,p}$ be the estimated $p$th quantiles for years 1 and 2.

Define a test statistic $T = \hat{\xi}_{1,p} - \hat{\xi}_{2,p}$.

So long as the quantile estimators have good properties, this $T$ is a good test statistic.

We may estimate quantiles based on plain empirical distributions (EM) or based on DRM through empirical likelihood (EL), or likelihood ratio (ELR).

Each of these three statistics leads to a permutation test.

See simulation results based on simulated data.
Rejection rate of tests for equal percentiles with clustered normal data

<table>
<thead>
<tr>
<th></th>
<th>((\sigma_1, \sigma_2, \sigma_3) = (1, 1, 2))</th>
<th></th>
<th>((\sigma_1, \sigma_2, \sigma_3) = (1, 2, 3))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(5^{th}) percentile</td>
<td>(50^{th}) percentile</td>
<td>(5^{th}) percentile</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>EL</td>
<td>ELR</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>5.4</td>
<td>4.5</td>
<td>4.7</td>
</tr>
<tr>
<td>3</td>
<td>17.7</td>
<td>19.2</td>
<td>19.1</td>
</tr>
<tr>
<td>4</td>
<td>41.8</td>
<td>49.5</td>
<td>48.4</td>
</tr>
</tbody>
</table>

\(K + 1 = 5, \ r = 10, \ n = 36\)

<table>
<thead>
<tr>
<th></th>
<th>(5^{th}) percentile</th>
<th>(50^{th}) percentile</th>
<th>(5^{th}) percentile</th>
<th>(50^{th}) percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>EL</td>
<td>ELR</td>
<td>EM</td>
</tr>
<tr>
<td>2</td>
<td>5.7</td>
<td>5.6</td>
<td>5.9</td>
<td>5.2</td>
</tr>
<tr>
<td>3</td>
<td>13.2</td>
<td>13.9</td>
<td>13.8</td>
<td>18.8</td>
</tr>
<tr>
<td>4</td>
<td>52.1</td>
<td>59.7</td>
<td>58.2</td>
<td>79.0</td>
</tr>
</tbody>
</table>

\(K + 1 = 5, \ r = 5, \ n = 48\)

<table>
<thead>
<tr>
<th></th>
<th>(5^{th}) percentile</th>
<th>(50^{th}) percentile</th>
<th>(5^{th}) percentile</th>
<th>(50^{th}) percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>EL</td>
<td>ELR</td>
<td>EM</td>
</tr>
<tr>
<td>2</td>
<td>5.0</td>
<td>4.2</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>3</td>
<td>21.4</td>
<td>24.9</td>
<td>24.6</td>
<td>34.6</td>
</tr>
<tr>
<td>4</td>
<td>49.3</td>
<td>60.9</td>
<td>59.4</td>
<td>79.7</td>
</tr>
</tbody>
</table>
Summary on normal clustered data

- Without permutation, both $t$ and $w$ tests have inflated type I errors.
- Their permutation tests have well controlled type I errors.
- To test for change in quantiles, we should use EM, EL or similar test statistics.
- They all have well controlled type I errors and satisfactory power properties.
- The notion of DRM is not discussed in this presentation. I would like to say that it helps.
Other simulation results

- We also generated clustered data based on rotating sampling plan from Gamma distribution, or from real data via unequal probability sampling, the conclusions are the same:
  - Without permutation, both $t$ and $w$ tests have inflated type I errors.
  - Their permutation tests have well controlled type I errors and good power properties.
Discussions

- The permutation principle works as long as we can identify test statistics which are invariant under the null hypothesis on a rich group of permutations.
- Accidentally missing out a PSU, imperfect rotating scheme, unequal cluster sizes can all in principle be accommodated in our proposed method.
- We have R-code to conduct each of these 5 tests.
Thank you and questions are welcome.