Rank-transformed subsampling: inference for exchangeable $p$-values

Rajen D. Shah (Statistical Laboratory, University of Cambridge)

Warwick Statistics CRiSM Seminar
26 April 2023
Joint work with:

Richard Guo
Randomised tests are useful
Drawbacks of randomised tests
Rank-transformed subsampling
Applications
Randomised tests
Suppose we are interested in testing a null hypothesis $H_0$ given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$. 
Suppose we are interested in testing a null hypothesis $H_0$ given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

**Steps:**

1. **‘Hunt’**: Use Part A to determine which test to use to target the alternative the data appear to have come from.

2. **Test**: Apply the test to Part B.
Sample splitting (Moran, 1973; Cox, 1975)

Suppose we are interested in testing a null hypothesis $H_0$ given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^c$.

A  ‘Hunt’: Use Part A to determine which test to use to target the alternative the data appear to have come from.

B  Test: Apply the test to Part B.

In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the ‘hunting’ in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.
Suppose we are interested in testing a null hypothesis $H_0$ given iid data.

Different tests may be particularly powerful against different alternatives $P \in H_0^C$.

1. **‘Hunt’**: Use Part A to determine which test to use to target the alternative the data appear to have come from.

2. **Test**: Apply the test to Part B.

In the second step, we may treat the chosen test as fixed in advance and there is no need to account for the ‘hunting’ in step 1.

Hunt step can be as elaborate as needed in order to find an appropriate test.

Strategy particularly useful when $H_0 = \cap_{\delta \in \mathcal{D}} H_0(\delta)$, so $H_1 = \cup_{\delta \in \mathcal{D}} H_0^C(\delta)$. 
Testing for clustering structure

Clustering algorithms cannot be used directly, as they may return clusters when none are truly present. We can formalise our null hypothesis as testing for unimodality.

Various notions exist in multiple dimensions including linear unimodality:

$X \in \mathbb{R}^p$ is unimodal if $a^T X$ is unimodal $\forall a \neq 0$.

That is, $H_0: \bigcap \{ a \neq 0 : a^T X \text{ is unimodal} \}$, $H_1: \exists a \neq 0$ such that $a^T X$ is not unimodal.
Testing for clustering structure

Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.
Clustering algorithms cannot be used directly, as they may return clusters when none are truly present.

We can formalise our null hypothesis as testing for **unimodality**.

Various notions exist in multiple dimensions including *linear unimodality*:

\[ X \in \mathbb{R}^p \text{ is unimodal if } a^T X \text{ is unimodal } \forall a \neq 0. \]

That is,

\[ H_0 : \bigcap_{a \neq 0} \{ a^T X \text{ is unimodal} \}, \]

\[ H_1 : \exists a \neq 0 \text{ such that } a^T X \text{ is not unimodal}. \]
Testing for clustering structure
Testing for clustering structure
Testing for clustering structure

- Rank-transformed subsampling

26 April 2023
Testing for clustering structure

- Rank-transformed subsampling

26 April 2023 11 / 32
Consider testing the conditional independence $H_0 : X \independent Y \mid Z$ given iid copies $(X_i, Y_i, Z_i)_{i=1}^n$.

Suppose instead we had samples from a distribution with density

$$p(x, y, z) \frac{q(z)}{p(z \mid x)} = p(y \mid x, z)p(z \mid x)p(x) \frac{q(z)}{p(z \mid x)}$$

$$= p(y \mid x, z)q(z)p(x)$$

under $H_0 = p(y \mid z)q(z)p_x(x)$.

Thus the reweighted marginal distribution of $(X, Y)$ would be $p(y)p(x)$, so $X \independent Y$.

In the reweighted distribution, the null of conditional independence becomes the simpler null of independence.
Consider testing the conditional independence $H_0 : X \perp \perp Y | Z$ given iid copies $(X_i, Y_i, Z_i)_{i=1}^n$.

Suppose instead we had samples from a distribution with density

$$p(x, y, z) \frac{q(z)}{p(z|x)} = p(y|x, z) p(z|x) p(x) \frac{q(z)}{p(z|x)}$$

$$= p(y|x, z) q(z) p(x)$$

under $H_0 = p(y|z) q(z) p(x)$.

Thus the reweighted marginal distribution of $(X, Y)$ would be $p(y)p(x)$, so $X \perp \perp Y$.

In the reweighted distribution, the null of conditional independence becomes the simpler null of independence.

If we know $p(z|x)$ we can always obtain a sample from the reweighted distribution through e.g. rejection sampling.
Testing generalised conditional independencies (Robins, 1999)

\[ p(y|x, z)p(z|x)p(x) \]

\[ p(y|x, z)q(z)p(x) \]
Testing generalised conditional independencies (Robins, 1999)

\[ p(y|x, z)p(z|x)p(x) \]

\[ p(y|x, z)q(z)p(x) \]

Reweighting by \( \frac{q(a_2)}{p(a_2|a_1,l)} \)

\( A_1, A_2 \): 1st and 2nd treatments. \( L, Y \): 1st and 2nd outcomes, confounded by e.g. health status \( U \).
Drawbacks of randomised tests
Replicability and Power

**Replicability:** Conclusions may depend delicately on the random seed used.

**Power loss:** Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.

Test statistics $T^{(1)}, \ldots, T^{(L)}$ are exchangeable.
Replicability and Power

**Replicability:** Conclusions may depend delicately on the random seed used.

**Power loss:** Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.

![Diagram]

Test statistics $T^{(1)}, \ldots, T^{(L)}$ are exchangeable.

Suppose $T^{(1)} \sim N(\mu, 1)$ and we want to test

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0.$$  

Consider aggregating them by, e.g.,

$$S = \left( T^{(1)} + \cdots + T^{(L)} \right) / L.$$
Replicability: Conclusions may depend delicately on the random seed used.

Power loss: Tests may not be making full use of the data.

Consider repeatedly applying the same randomised procedure to the same data.

\[ T^{(1)}, \ldots, T^{(L)} \]

Test statistics \( T^{(1)}, \ldots, T^{(L)} \) are exchangeable.

Suppose \( T^{(1)} \sim \mathcal{N}(\mu, 1) \) and we want to test

\[ H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0. \]

Consider aggregating them by, e.g.,

\[ S = \left( T^{(1)} + \cdots + T^{(L)} \right) / L. \]

Let us further model \( (T^{(1)}, \ldots, T^{(L)}) \) as jointly normal with correlation \( \rho > 0 \).
Toy example: replicability

Single-split test  
Reject $H_0$ when $T^{(1)}$ is large

Aggregated test  
Reject $H_0$ when $S = \frac{T^{(1)} + \cdots + T^{(L)}}{L}$ is large.

"Not replicated" = one acceptance and one rejection in two runs.
Toy example: Power

**Single-split test**  
Reject $H_0$ when $T^{(1)}$ is large

**Aggregated test**  
Reject $H_0$ when $S = \left( T^{(1)} + \cdots + T^{(L)} \right) / L$ is large.

\[
\begin{array}{c}
\text{method} \\
\text{red: aggregated (L=200)} \\
\text{orange: single}
\end{array}
\]
Complex dependence in practice

In the toy example, the aggregated test was based on

\[ S \sim \mathcal{N}(\mu, 1/L + \rho(L - 1)/L). \]
Complex dependence in practice

In the toy example, the aggregated test was based on

\[ S \sim \mathcal{N}(\mu, 1/L + \rho(L - 1)/L). \]

In reality, however, the dependence among \( T^{(1)}, \ldots, T^{(L)} \) can be complex and there is no good description or approximation (beyond symmetry).

![Distribution of S in a real hunt-and-test example case from Kim & Ramdas (2020)](image)

Toy example: Power

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-split test</td>
<td>Reject $H_0$ when $T^{(1)}$ is large</td>
</tr>
<tr>
<td>Aggregated test</td>
<td>Reject $H_0$ when $S = \frac{T^{(1)} + \cdots + T^{(L)}}{L}$ is large.</td>
</tr>
</tbody>
</table>

The conservative rule rejects when $S > 2z_\alpha$. 

![Graph showing power methods for different ρ values]
Toy example: Power

Single-split test  Reject $H_0$ when $T^{(1)}$ is large
Aggregated test  Reject $H_0$ when $S = (T^{(1)} + \cdots + T^{(L)}) / L$ is large.

The conservative rule rejects when $S > 2z_\alpha$.

Other conservative approaches similarly lose power. (Vovk & Wang, 2020; Vovk et al., 2021; DiCiccio et al., 2020; Meinshausen et al., 2009;...
Rank-transformed subsampling
Have exchangeable test statistics $T^{(1)}, \ldots, T^{(L)}$.

**A1** Under $P \in H_0$, $T_n$ is asymptotically $U(0, 1)$. (Also works for $T_n \xrightarrow{d} \mathcal{N}(0, 1)$).
Have exchangeable test statistics $T^{(1)}, \ldots, T^{(L)}$.

**A1** Under $P \in H_0$, $T_n$ is asymptotically $U(0,1)$. (Also works for $T_n \xrightarrow{d} \mathcal{N}(0,1)$).

Choose a deterministic **aggregation function** $S : \mathbb{R}^L \rightarrow \mathbb{R}$ to give $S_n := S(T^{(1)}, \ldots, T^{(L)})$.

**A2** Under $P \in H_0$, $S_n$ converges to (unknown) distribution $G_P$ with bounded density.

We wish to construct a test / form a $p$-value based on $S_n$. 
Subsampling \((\text{e.g. Politis et al. (1999)})\)

We use **subsampling** to estimate the asymptotic distribution \(G_P\).

Choose \(b = 1, \ldots, B\) subsamples of size \(m := \lfloor n / \log n \rfloor\).

\[
\hat{G}_n := \text{ECDF}\{\hat{S}_1, \ldots, \hat{S}_B\}.
\]

By standard consistency of subsampling \((\text{e.g. Politis et al. (1999)})\) for \(P \in H_0\),

\[
\|\hat{G}_n - G_P\|_\infty \xrightarrow{p} 0.
\]
We use **subsampling** to estimate the asymptotic distribution $G_P$.

Choose $b = 1, \ldots, B$ subsamples of size $m := \lfloor n / \log n \rfloor$.

\[
\hat{G}_n := \text{ECDF} \{\hat{S}_1, \ldots, \hat{S}_B\}.
\]

By standard consistency of subsampling (e.g. Politis et al. (1999)) for $P \in H_0$, \[\|\hat{G}_n - G_P\|_\infty \xrightarrow{P} 0.\]

But $\hat{G}_n$ will continue to approximate $G_P$ under local alternatives.
We have not yet used that we *know* the asymptotic null distribution of $T_n^{(1)}$ (A1).
We have not yet used that we *know* the asymptotic null distribution of $T^{(1)}_n$ (A1).

Replace each $T^{(l)}_{m,b}$ by its normalised rank within the matrix:

$$\tilde{T}^{(l)}_{m,b} := \frac{\#\{T^{(l')}_{m,b'} \leq T^{(l)}_{m,b}\}}{BL} = \hat{F}_n \left(T^{(l)}_{m,b}\right),$$

where $\hat{F}_n$ is the ECDF of $\{T^{(l)}_{m,b}\}$.

We use $\tilde{G}_n := \text{empirical measure of } \{\tilde{S}_b\}$ as our reference for testing.
Rank transform: illustration

$H_0$

$H_1$

before rank transform   after rank transform
Theorem (Size control)

Under (A1) and (A2), for \( P \in H_0, \mathbb{P}_P \left\{ S_n < \tilde{G}_n^{-1}(\alpha) \right\} \to \alpha. \)
Theory

Theorem (Size control)

Under (A1) and (A2), for $P \in H_0$, $\mathbb{P}_P \left\{ S_n < \tilde{G}_n^{-1}(\alpha) \right\} \to \alpha$.

Theorem (Local alternatives)

Suppose additionally that for every sequence $P_n \in \mathcal{P}$ (where $\mathcal{P}$ is the set of possible distributions) such that $P_n \overset{d}{\to} P \in H_0$, it holds that

$$\left( F_{n,P_n}(T^{(1)}_n), \ldots, F_{n,P_n}(T^{(L)}_n) \right) \overset{d}{\to} (C^{(1)}, \ldots, C^{(L)}) \quad (3.1)$$

for some $(C^{(1)}, \ldots, C^{(L)})$ whose distribution does not depend on the sequence $P_n$. Then, for every sequence $P_n \in \mathcal{P}$ such that $P_n \overset{d}{\to} P$, we have that

$$\tilde{G}_n^{-1}(1 - \alpha) \overset{p}{\to} G_{P}^{-1}(1 - \alpha),$$

i.e. the critical value of our test converges to that of the ‘oracle’ test.
Applications
Testing unimodality

Hunt: 2-means clustering, Test: $T_n = \text{asymptotic dip test } p\text{-value}$. $L = 50$ splits.

Setting: Mixture of two $d$-dimensional (unit ball, multivariate $t$) distributions separated $\tau$ away.

Aggregation: Consider $S = \text{avg}$, $S = \text{min}$.

SigClust is a competing method based on multivariate normal mixture.

Adaptive algorithm version available that adapts to the aggregation function with better performance.
Gene expression of cancer subtypes

Three types of renal cell carcinoma: clear cell (ccRCC), papillary (PRCC) and chromophobe (ChRCC).

ICGC/TCGA Pan-Cancer dataset: Expression levels of 1,000 genes. $L = 6000$ splits.
Generalised conditional independence

Reweighting by \( \frac{q(a_2)}{p(a_2|a_1, l)} \) to give distribution \( Q \)

\( A_1, A_2 \): 1st and 2nd treatments. \( L, Y \): 1st and 2nd outcomes, confounded by e.g. health status \( U \).

\( H_0 \): \( A_1 \) has no direct effect on \( Y \).
Generalised conditional independence

Reweighting by \( \frac{q(a_2)}{p(a_2 | a_1, l)} \) to give distribution \( Q \)

\( A_1, A_2: \) 1st and 2nd treatments. \( L, Y: \) 1st and 2nd outcomes, confounded by e.g. health status \( U. \)

\( H_0: \) \( A_1 \) has no direct effect on \( Y. \)

We can construct tests based on \( \text{Cov}_Q(A_1, Y) = 0. \)

**IPW test** (Robins, 1999) proposes

\[
Z_i := \frac{Y_i(A_{1,i} - \mathbb{E}A_{1,i})}{p(A_{2,i} | A_{1,i}, L_i)}, \quad T := \sum_i Z_i / \sqrt{\sum_i Z_i^2} \xrightarrow{d} \mathcal{N}(0, 1).
\]
Simple setting

**Setting:** \( A_1 \sim \text{Ber}(1/2), \ A_2 \sim \text{Ber}(\expit(2A_1 - L + 2)) \) and

\[
U \sim \mathcal{N}_4(0, \Sigma_{ij} = 2^{-|i-j|}), \quad L = A_0 + \beta_{U,L}^T U + \varepsilon_L, \quad Y = \tau A_1 - A_2 + \beta_{U,Y}^T U + \varepsilon_Y.
\]

Also consider alternative to rejection sampling: ‘distinct replacement sampling’ (DRPL) (Thams et al., 2021).

\[ S = \text{avg}(L=20) \]

\( S \) is the average of permutation \( p \)-values computed on \( L = 20 \) accepted/resampled data.
More complex setting

A more difficult setting where IPW is inapplicable.

$T_n$: permutation $p$-value with HSIC (Gretton et al., 2012) as the statistic on the accepted/resampled data.
Randomised tests can be useful for a variety of applications:

- Testing for clustering structure, testing for the presence of signal in high-dimensional data, goodness-of-fit testing, (nonparametric) variable significance testing,…
- Testing under distributional shifts
- Testing or confidence interval construction based on double / debiased machine learning

Replicability and power issues may hamper their adoption in practice.

- Conservative aggregation rules improve replicability, but power can degrade significantly.

While a naive subsampling also suffers from power loss, rank-transformed subsampling uses knowledge of the null distribution to avoid these issues.

Thank you for listening.