Statistics paper proposal

Tim B

In this document I chart out several ideas for a next statistics (as opposed to genomics) paper. The following keywords are applicable: randomization tests, conditional randomization/permutation tests, e-values, multiple hypothesis testing, sample splitting, kernels, online model training. I propose two main directions and a few tertiary directions. I also suggest a genomics application.

Direction 1: A more general framework for randomization tests

Motivation

Randomization tests are statistical tests in which a test statistic is recomputed over permuted, resampled, rotated, or otherwise transformed versions of the data to produce an empirical null distribution against which a statistic computed on the raw data is compared. Randomization tests are ubiquitous throughout all of statistics and science. For example, randomization tests routinely are used in neuroscience to perform nonparametric inference on GLMs (Winkler et al. 2014) and in genetics to test the association of SNPs (Johnson et al. 2010) and for other purposes. See Dobriban 2021 for a recent review and unifying theoretical analysis of randomization tests. Despite their widespread popularity, randomization tests pose several practical challenges, especially in high-multiplicity settings: miscalibrated p-values can cause type-I error inflation; combining p-values across many hypothesis tests can result in excessively conservative or liberal discovery sets, especially when tests are dependent; and sample splitting – a procedure required by certain randomization tests, such as the recently-proposed holdout randomization test (Tansey et al. 2021) – can result in non-reproducible p-values.

Building closely on the work of several authors (e.g., Wang and Ramdas 2020; Vovk 2020; Vovk and Wang 2021) we propose a simple new framework for randomization tests that helps to resolve these challenges. The framework leverages e-values, test statistics that by definition have unit expectation under the null. We recover the standard, p-value based approach to randomization tests a special case of our framework. Although our theory applies broadly, we focus mostly on (marginal) permutation tests, conditional randomization tests, and conditional permutation tests as illustrative examples.

Let T^* be the test statistic computed on the raw data, and let $T_1, T_2 \ldots, T_B$ be the *ordered* test statistics recomputed on the permuted (or resampled, etc.) data. (We assume for the time being that there are no ties.) Define $I_i = \mathbb{I}(T^* \leq T_i)$. For given constants $a_0, a_1, \ldots, a_B \in \mathbb{R}$, define e by

$$e = a_0 + \sum_{i=1}^{B} a_i I_i,$$

where e is subject to the constraint that $\mathbb{E}[e] = 1$ under the null hypothesis. In other words, e is an e-value that is a linear combination of the I_is. We call e a "randomization test e-value (RT e-value)" so as to distinguish it from other, more general e-values.

Constructing RT e-values

We consider methods for constructing RT e-values, leveraging two key properties of the I_is. First, because T^* is i.i.d. with the T_is, we have that $\mathbb{P}(T^* \leq T_i) = i/B$. Therefore, $\mathbb{E}[I_i] = i/B$. We state the second key property as a proposition.

Proposition 1 *For $r, B \in \mathbb{N}$, we have that*

$$\left(\sum_{i=1}^{B} I_i\right)^r = \sum_{i=1}^{B} \left[(B-i+1)^r - (B-i)^r\right] I_i.$$

Equivalently, the rth power of the empirical right-sided p-value $p_B := \frac{1}{B}\sum_{i=1}^{B} I_i$ is

$$p_B^r = \sum_{i=1}^{B} \left[(1 - i/B + 1/B)^r - (1 - i/B)^r\right] I_i := \sum_{i=1}^{B} M(B, r, i) I_i.$$

Finally, for given $x_0 \in \mathbb{R}$ and coefficients $c_0, c_1, \ldots, c_r \in \mathbb{R}$, the rth degree polynomial $\sum_{j=0}^{r} c_j (p_B - x_0)^j$ is given by

$$\sum_{j=0}^{r} c_j (p_B - x_0)^j = \left(a_0 + \sum_{j=1}^{r} a_j x_0^j\right) + \sum_{i=1}^{B} \left[\sum_{j=1}^{r}\sum_{k=1}^{j} a_j x_0^{j-k} \binom{j}{k} M(B, k, i)\right] I_i.$$

This proposition states that the rth power of the sum of the I_is is a simple linear combination of the I_is, not a messy multinomial expression as one might initially expect. Equivalently, the rth power of the empirical right-sided p-value (i.e., $p_B = (1/B)\sum_{i=1}^{B} I_i$) is a simple linear combination of the I_is. Finally, the rth degree polynomial of p_B centered at x_0 with coefficients $c_0, \ldots, c_r$ is a straightforward extension of the above. The proof of Proposition 1 is a combinatorial argument (see appendix). This result is very simple, but we have not seen it elsewhere.

We provide two quick examples of specific RT e-values. Then, we describe a general and flexible procedure for producing powerful RT e-values.

Example 1: A quick algebraic example.

Example 2: Recovering p-values as a special case.

Example 3: Powerful RT e-values via p-to-e calibrators.

Direction 2: A class of fast and powerful test statistics for the conditional randomization (permutation) test

References

Dobriban, Edgar (2021). "Consistency of invariance-based randomization tests". In: pp. 1–35 (cit. on p. 1).

Johnson, Randall C. et al. (2010). "Accounting for multiple comparisons in a genome-wide association study (GWAS)". In: *BMC Genomics* 11.1, pp. 2–7 (cit. on p. 1).

Tansey, Wesley, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei (2021). "The Holdout Randomization Test for Feature Selection in Black Box Models Wesley". In: *Journal of Computational and Graphical Statistics*, pp. 1–12 (cit. on p. 1).
Vovk, Vladimir (2020). "A note on data splitting with e-values". In: pp. 1–11 (cit. on p. 1).
Vovk, Vladimir and Ruodu Wang (2021). "E-values: Calibration, combination and applications". In: *Annals of Statistics* 49.3, pp. 1736–1754 (cit. on p. 1).
Wang, Ruodu and Aaditya Ramdas (2020). "False discovery rate control with e-values". In: pp. 1–32 (cit. on p. 1).
Winkler, Anderson M., Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols (2014). "Permutation inference for the general linear model". In: *NeuroImage* 92, pp. 381–397 (cit. on p. 1).

Appendices

We provide a proof of Proposition 1. Define the set $C(r, B)$ by

$$C(r, B) := \left\{ (k_1, \ldots, k_B) \in \{0, \ldots, r\}^B : \sum_{i=1}^{B} k_i = r \right\},$$

i.e., $C(r, B)$ is the set of length-B tuples of integers from 0 to r such that the elements of the tuple sum to r. Next, let the function $\tau : C(r, B) \to \{1, \ldots, B\}$ be defined by

$$\tau(k_1, \ldots, k_B) = \min \{ i \in \{1, \ldots, B\} : k_i \geq 1 \},$$

i.e., $\tau(k_1, \ldots, k_B)$ is the position of the minimal nonzero element of a $(k_1, \ldots, k_B)$. Finally, let τ^{-1} be the pre-image of of τ. It is easy to see that, for $i \in \{0, \ldots, B\}$,

$$\tau^{-1}(i) = \{(0, \ldots, 0, k_i, k_{i+1}, \ldots, k_B) \in C(r, B) : k_i \geq 1\}.$$

In other words, $\tau^{-1}(i)$ is the set of tuples whose first nonzero entry is the ith entry.

Before proceeding, we establish an important property of the I_is. If at least one of the k_is is nonzero, we have that

$$\prod_{i=1}^{B} I_i^{k_i} = I_{\tau(k_1, \ldots, k_B)}. \tag{1}$$

This equality holds for the following reason. Assume without loss of generality that there are $N \in \{1, \ldots, B\}$ nonzero k_is. Let $\sigma : \{1, \ldots, N\} \to \{1, \ldots, B\}$ give the position of the ith nonzero k_i (so that $\sigma(1)$ is the position of the first nonzero k_i, $\sigma(2)$ is the position of the second, etc.). We can write

$$\prod_{i=1}^{B} I_i^{k_i} = \prod_{i=1}^{N} I_{\sigma(i)}^{k_{\sigma(i)}}, \tag{2}$$

i.e., we can remove all I_is that are raised to the power of zero. Because the I_is are Bernoulli random variables, we have that $I_{\sigma(i)}^{k_{\sigma(i)}} = I_{\sigma(i)}$. Next, recall that $I_\sigma(i) = \mathbb{I}\left(T^* \leq T_{\sigma(i)}\right)$, where

$T_{\sigma(1)} < T_{\sigma(2)} < \cdots < T_{\sigma(n)}$. If $T^* \leq T_{\sigma(i)}$, then by transitivity, $T^* \leq T_{\sigma(2)} < \cdots < T_{\sigma(n)}$, implying $I_{\sigma(i)} = 1$ for all $i \in \{1, \ldots, N\}$. Therefore, $I_{\sigma(1)} = 1 = I_{\sigma(1)} \ldots I_{\sigma(n)}$. On the other hand, if $T^* > T_{\sigma(i)}$ then $T_{\sigma(1)} = 0$, implying $I_{\sigma(1)} = 0 = I_{\sigma(1)} \ldots I_{\sigma(n)}$. Combining these cases, we conclude that $I_{\sigma(1)} = I_{\sigma(1)} \ldots I_{\sigma(B)}$. Equation 2 therefore reduces to

$$\prod_{i=1}^{N} I_{\sigma(i)}^{k_{\sigma(i)}} = I_{\sigma(1)}. \tag{3}$$

Finally, because $\sigma(1)$ is the position of the first nonzero k_i, we have that $\sigma(1) = \tau(k_1, \ldots, k_B)$. Combining this fact with (2) and (3) yields the conclusion (1).

Having established this lemma, we can evaluate the rth power of the sum of the I_is. The multinomial theorem states that

$$\begin{aligned}\left(\sum_{i=1}^{B} I_i\right)^r &= \sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} \prod_{i=1}^{B} I_i^{k_i} \\ &= \sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} I_{\tau(k_1,\ldots,k_B)}, \end{aligned} \tag{4}$$

where the second equality follows from (1). The final term in (4) is simply a linear combination of the I_is. We therefore can factor out the terms in the sum corresponding to I_i for each i, yielding

$$\begin{aligned}\sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} I_{\tau(k_1,\ldots,k_B)} &= \sum_{i=1}^{B} \sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1, k_2, \ldots, k_B} I_i \\ &= \sum_{i=1}^{B} I_i \sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1, k_2, \ldots, k_B}. \end{aligned} \tag{5}$$

We evaluate the inner sum (5), which is the coefficient corresponding to I_i in the linear combination. We have that

$$\begin{aligned}&\sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1, k_2, \ldots, k_B} = \sum_{(0,\ldots,0,k_i,\ldots,k_B):k_i\geq 1, k_i+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} \\ &= \sum_{(k_1,\ldots,k_{B-i+1}):k_1\geq 1, k_1+\cdots+k_{B-i+1}=r} \binom{r}{k_1, k_2, \ldots, k_{B-i+1}} = \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \binom{r}{j, l_1, \ldots, l_{B-i}} \\ &= \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r!}{j! l_1! \ldots l_{B-i}!} = \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r(r-1)\ldots(r-j+1)(r-j)!}{j! l_1! \ldots l_{B-i}!} \\ &= \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r!}{(r-j)! j!} \binom{r-j}{l_1, \ldots, l_{B-i}} = \sum_{j=1}^{r} \frac{r!}{(r-j)! j!} \sum_{l_1,\ldots,l_{B-i}} \binom{r-j}{l_1, \ldots, l_{B-i}} \\ &= \sum_{j=1}^{r} \binom{r}{j} (B-i)^{r-j} = \sum_{j=1}^{r} \binom{r}{j} (B-i)^{r-j} 1^j = \sum_{j=0}^{r} \binom{r}{j} (B-i)^{r-j} 1^j - 1(B-i)^r \\ &= (B-i+1)^r - (B-i)^r. \end{aligned} \tag{6}$$

Combining (4), (5), and (6), we conclude that

$$\left(\sum_{i=1}^{B} I_i\right)^r = \sum_{i=1}^{B} \left[(B-i+1)^r - (B-i)^r\right] I_i.$$

Next, setting $p_B = \frac{1}{B}\sum_{i=1}^{B} I_i$, we obtain

$$p_B^r = \left(\frac{1}{B}\sum_{i=1}^{B} I_i\right)^r = \sum_{i=1}^{B} \left[(1-i/B+1/B)^r - (1-i/B)^r\right] I_i = \sum_{i=1}^{B} M(B,r,i) I_i.$$

We next consider the rth degree polynomial of p_B^r. First, for $j \in \mathbb{N}$, and $x_0 \in \mathbb{R}$, we have that

$$\begin{aligned}(p_B - x_0)^r &= \sum_{k=0}^{j} \binom{j}{k} p_B^k x_0^{j-k} = x_0^j + \sum_{k=1}^{j} \binom{j}{k} p_B^k x_0^{j-k} \\ &= x_0^j + \sum_{k=1}^{j} \binom{j}{k} \left[\sum_{i=1}^{B} M(B,k,i) I_i\right] x_0^{j-k} = x_0^j + \sum_{i=1}^{B} \left[\sum_{k=1}^{j} x_0^{j-k} \binom{j}{k} M(B,k,i)\right] I_i.\end{aligned}$$

Finally, let $c_0, c_1, \ldots, c_r \in \mathbb{R}$ be polynomial coefficients. We have that

$$\begin{aligned}\sum_{j=0}^{r} c_j (p_B - x_0)^j = c_0 + \sum_{j=1}^{r} c_j (p_B - x_0)^j &= c_0 + \sum_{l=1}^{r} c_j \left[x_0^j + \sum_{i=1}^{B} \left[\sum_{k=1}^{j} x_0^{j-k} \binom{j}{k} M(B,k,i)\right] I_i\right] \\ &= \left(c_0 + \sum_{j=1}^{r} c_j x_0^j\right) + \sum_{i=1}^{B} \left[\sum_{j=1}^{r}\sum_{k=1}^{j} c_j x_0^{j-k} \binom{j}{k} M(B,k,i)\right] I_i,\end{aligned}$$

completing the proof.