Stats paper proposal
Tim B

In this document I chart out several ideas for a next statistics paper. The following keywords are relevant: randomization tests, conditional randomization/permutation tests, e-values, multiple hypothesis testing, sample splitting, robustness, and online learning. I propose two main directions and a few tertiary directions. I also suggest a CRISPR genomics application.

Direction 1: A more general framework for randomization tests

Motivation and background

Randomization tests are tests in which a 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 genetics to assess the association of SNPs (Johnson et al. 2010) and to test the differential expression of genes (Maleki et al. 2020) and in neuroscience to improve the robustness of GLMs in the analysis of brain imaging data (Winkler et al. 2014). 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, ML-based holdout randomization test (Tansey et al. 2021) – can result in non-reproducible p-values.

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

As a brief review, let T^* be the test statistic computed on the raw data, and let $T_1, T_2 \ldots, T_B$ be the test statistics recomputed on the permuted (or resampled, etc.) data. We assume for simplicity that there are no ties among the T_is, although this assumption has minimal impact on the theory. Randomization tests satisfy the following key invariance property: the vector $(T^*, T_1, T_2, \ldots, T_B)$ is *exchangeable* under the null hypothesis. This means that the cumulative distribution function of $(T^*, T_1, T_2, \ldots, T_B)$ is invariant to permutations. Exchangeability implies that the empirical p-value

$$p_B = \frac{\sum_{i=1}^{B} \mathbb{I}\left(T^* \leq T_i\right)}{B}$$

is valid, i.e., p stochastically dominates the uniform distribution. (One often adds a 1 to both the numerator and denominator to ensure finite-sample correctness, but for large B, the two expressions coincide almost exactly.) The standard practice is to reject the null hypothesis at level $\alpha \in (0,1)$ if $p < \alpha$, adjusting for multiple testing (typically via BH) to control FDR in high multiplicity settings.

Linear e-values

We propose a more general framework for randomization testing based on e-values. For $i \in \{1, \ldots, B\}$, let $T_{(1)}, \ldots, T_{(B)}$ be the order statistics of $T_1, \ldots, T_B$. Define $I_i = \mathbb{I}(T^* \leq T_{(i)})$. For given constants $a_0, a_1, \ldots, a_B \in \mathbb{R}$, define the test statistic e by

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

where e is a nonnegative random variable such that $\mathbb{E}[e] = 1$ under the null hypothesis. We call e a "linear e-value" so as to distinguish it from other, more general e-values. We state two key propositions that enable us to construct linear e-values with ease. The first proposition derives the expectation of I_i under the null hypothesis by applying a well-known fact about the ranks of exchangeable variables (recorded in Kuchibhotla 2020).

Proposition 1 *For $i \in \{1, \ldots, B\}$, $\mathbb{E}[I_i] = i/(B+1)$.*

Proof: The vector $(T^*, T_1, \ldots, T_n)$ is exchangeable. Therefore, by Corollary 1 of (Kuchibhotla 2020),

$$\mathbb{P}\left[\text{rank}(T^*) \leq i\right] = \frac{i}{1+B},$$

where

$$\text{rank}(T^*) = |\{j \in \{1, \ldots, B\} : T_j \leq T^*\}| + 1$$

is the rank of T^*. But $\text{rank}(T^*) = i$ if and only if $T^* \leq T_{(i)}$. Therefore, $\mathbb{P}(T^* \leq T_{(i)}) = i/(B+1)$, implying the conclusion. □

The next proposition derives a simple expression for the rth power of the sum of the I_is (proof deferred to appendix).

Proposition 2 *For $r, B \in \mathbb{N}$,*

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

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

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

In other words, 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. This result, though simple, is to the best of our knowledge new.

Examples of linear e-values

Using the above propositions, we construct several example linear e-values.

Example 1: Warmup. Setting $a_i = (B+1)/i$ and $c = 0$ in definition (1), we obtain

$$e = (B+1) \sum_{i=1}^{B} \frac{I_i}{i}.$$

The variable e clearly is nonnegative. Moreover,

$$\mathbb{E}(e) = (B+1) \sum_{i=1}^{B} \frac{i}{i(B+1)} = 1.$$

Therefore, e is a linear e-value.

Example 2: p^{shift}-values. We recover standard p-values (up to a translation) as a special case of the proposed framework. Set $c = 3/2$ and $a_i = -1/B$ in definition (1). Then we obtain

$$e = (3/2) - \frac{1}{B} \sum_{i=1}^{B} I_i$$

as our test statistic. This statistic clearly is nonnegative. (In fact, it is greater than or equal to 1/2.) Moreover, we can calculate its expectation as follows:

$$\mathbb{E}[e] = \frac{3}{2} - \frac{1}{B} \sum_{i=1}^{B} \frac{i}{B+1} = \frac{3}{2} - \frac{1}{B(B+1)} \sum_{i=1}^{B} i = \frac{3}{2} - \frac{1}{B(B+1)} \left(\frac{B^2+B}{2} \right) = \frac{3}{2} - \frac{1}{2} = 1.$$

Therefore, e is a valid linear e-value. We recognize $p_B = \frac{1}{B}\sum_{i=1}^{B} I_i$ as the standard, right-tailed randomization test p-value, which follows an approximate uniform distribution under the null hypothesis. Hence, $e = (3/2) - p_B$ follows an approximate $U(1/2, 3/2)$ distribution. We can leverage this additional distributional information on e to improve power in hypothesis testing (as we will see later).

This example motivates our definition of a "p^{shift}-value." A p^{shift}-value is a linear e-value that is a p-value under a translation and/or scaling. Formally, let e be a linear e-value. We say that e is a p^{shift}-value if there exist scalars $k_0, k_1 \in \mathbb{R}$ such that

$$\mathbb{P}(k_0 + k_1 e \leq \alpha) \leq \alpha.$$

p^{shift}-values have the same statistical properties as p-values; therefore, p^{shift}-values can be used to control the type-I error rate or FWER.

Powerful linear e-values via analytic p-to-e calibrators.

We outline a general and flexible strategy for producing powerful linear e-values. Vovk and Wang introduced the notion of a *p-to-e calibrator*, which is a function that transforms a p-value into an e-value (Vovk and Wang 2021). A p-to-e calibrator f must satisfy the following properties: (i) f is defined on $[0, 1]$; (ii) f is nonnegative; (iii) f is decreasing; and (iv) f integrates to (at most) 1. Vovk and Wang furthermore defined the notion of an "admissible" p-to-e calibrator, which roughly speaking is a calibrator that is "good" (or at least "not bad"), i.e. it yields powerful e-values. An admissible calibrator f satisfies the following additional properties: (i) f is left-continuous; (ii) $f(0) = \infty$; and (iii) f integrates to *exactly* 1. To illustrate the above definitions, let f be an admissible p-to-e calibrator, and let $P \sim U(0, 1)$ be a uniformly distributed random variable with cumulative density function $G(p) = p$. We have by LOTUS that $\mathbb{E}[f(P)] = \int_0^1 f(p)dG(p) = \int_0^1 f = 1$.

Vovk, Wang, and Ramdas introduced several examples of p-to-e calibrators, which we report here. First, for given $\kappa \in (0, 1)$, define the function $f_\kappa : [0, 1] \to \mathbb{R}^{\geq 0}$ by $f_\kappa(p) = \kappa p^{\kappa-1}$. It is easy to check that f_κ satisfies the properties listed above and is therefore an admissible p-to-e calibrator. Next, define $g : [0, 1] \to \mathbb{R}^{\geq 0}$ by

$$g(p) = \frac{1 - p + p \ln(p)}{p \log^2(p)}.$$

The function g also is an admissible p-to-e calibrator, although this is a bit harder to check.

We propose a general procedure for constructing linear e-values using p-to-e calibrators. Let $f : [0, 1] \to \mathbb{R}$ be an analytic p-to-e calibrator. Let $f^{(j)}$ denote the jth derivative of f. Furthermore, let $x_0 \in (0, 1)$, and let $r \in \mathbb{N}$. The rth degree Taylor expansion of f at x_0 is

$$f(x) = \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j + O(x^{r+1}).$$

For example, if $f_\kappa(x) = \kappa x^{\kappa-1}$ (as above), then

$$f_\kappa^{(j)}(x) = \prod_{l=0}^{j} (\kappa - l) x^{\kappa-j-1}.$$

Next, let $p_B = (1/B) \sum_{i=1}^{B} I_i$ be the right-tailed randomization test p-value. Evaluating f at p_B yields

$$f(p_B) = \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!}(p_B - x_0)^j + O(p_B^{r+1}).$$

For $j \in \{0, \ldots, r\}$, set $c_j = f^{(j)}(x_0)/(j!)$. Applying Proposition 2, can write the above as a linear combination of the I_is:

$$f(p_B) = \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!} x_0^j + \sum_{i=1}^{B} \left[\sum_{j=1}^{r} \sum_{k=1}^{j} \frac{f^{(j)}(x_0) x_0^{j-k} (-1)^{j-k} M(B,k,i)}{k!(j-k)!} \right] I_i + O(p_B^{r+1})$$
$$:= \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!} x_0^j + \sum_{i=1}^{B} W(f, x_0, r, B, i) I_i + O(p_B^{r+1}).$$

Omitting the remainder term, we take

$$e = \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!} x_0^j + \sum_{i=1}^{B} W(f, x_0, r, B, i) I_i \tag{2}$$

as our linear e-value. In words, the test statistic e is a linear e-value that approximates (to arbitrary precision) the p-to-e calibrator f evaluated at the empirical p-value p_B. The expectation of e is approximately 1 under the null hypothesis, assuming that the degree r of the Taylor polynomial is large and the p-value p_B is approximately uniformly distributed. However, we can normalize e to ensure that it has unit expectation. The mean μ of e is

$$\mu = \sum_{j=0}^{r} \frac{f^{(j)}(x_0)}{j!} x_0^j + \frac{1}{B+1} \sum_{i=1}^{B} W(f, x_0, r, B, i) i.$$

Therefore, the normalized linear e-value e_{norm} defined by $e_{\text{norm}} = e/\mu$ has expectation 1.

We note that we can calculate e in software in a highly efficient way. To calculate e, we must calculate $W(f, x_0, r, B, i)$ for all $i \in \{1, \ldots, B\}$; recall that f is the p-to-e calibrator, x_0 is the point about which we perform the Taylor expansion, B is the number of permuted (or resampled, etc.) test statistics, r is the degree of the Taylor polynomial, and i is the index of the resampled test statistic. For a given choice of (f, x_0, r, B), we can compute the coefficients $\{W(f, x_0, r, B, i)\}_{i=1}^{B}$ once and save them; then, we can load $\{W(f, x_0, r, B, i)\}_{i=1}^{B}$ and evaluate (2) each time we need to compute an e-value. Additionally, we can save the coefficients generated by certain, popular choices for (f, x_0, r, B, i) (e.g., $f(x) = (1/2)x^{-1/2}$, $x_0 = 1/2$, $r = 20$, and $B \in \{500, 1000, 5000\}$) in the software package itself. Overall, calculating (2) should be about the same speed as calculating a standard p-value.

Advantages to the proposed framework and connections to the standard approach

We highlight three advantages of the proposed framework.

1. **Robustness.** To achieve FDR control in the standard randomization test framework, we must ensure that the p-values are uniformly distributed (or more precisely, that the p-values stochastically dominate the uniform distribution) under the null hypothesis. This task can be quite challenging in practice: QQ-plots often reveal that p-values in the tail of the distribution are inflated. In settings in which we test many (e.g., tens of thousands or hundreds of thousands of) hypotheses, this p-value inflation can lead to a severe loss of FDR control.

The proposed framework provides a simple procedure to protect against this common source of type-I error inflation. Consider a linear e value

$$e = c + \sum_{i=1}^{B} a_i I_i.$$

Let $K(t) = (1/\sqrt{2\pi}) \exp(-(1/2)t^2)$ be the Gaussian kernel. Moreover, let $h > 0$ be a tunable kernel bandwidth. For $i \in \{1, \ldots, B\}$, define the weight $w_i > 0$ by

$$w_i = K(|i - B/2|/h).$$

In other words, the weights are defined by a Gaussian kernel over the ranks of the T_is. Observe that (i) w_i is a fixed (i.e., non-random) number, and (ii) as i tends toward 1 or B (i.e., away from the midpoint of $B/2$), w_i decreases. Next, define the weighted test statistic e_{weight} by

$$e_{\text{weight}} = c + \sum_{i=1}^{B} w_i a_i I_i.$$

The mean μ of e_{weight} is $\mu = c + \sum_{i=1}^{B} (w_i)(a_i)(i)/(B+1)$. Finally, define the normalized test statistic e_{norm} by

$$e_{\text{norm}} = e_{\text{weight}}/\mu.$$

e_{norm} is a valid linear e-value. Moreover, e_{norm} downweights the importance of resampled statistics in the tail of the sampling distribution, thereby protecting against miscalibration due to misspecified tails. To the best of our knowledge, there is no clear analog of this procedure for p-values.

2. **Multiple hypothesis testing under arbitrary dependence.** Wang and Ramdas recently designed an analog of the BH procedure for e-values (Wang and Ramdas 2020). The method, called "e-BH," takes as input a set of e-values and outputs a discovery set with guaranteed FDR control. Remarkably, the procedure controls FDR for any dependence structure between the e-values. The user optionally can supply information about the marginal or joint distribution of the e-values (e.g., that the PRDS property obtains) to boost power. Simulation studies and theoretical analyses indicate that e-BH is highly-powered. Linear e-values as proposed can be passed directly into the e-BH procedure.

3. **De-randomizing sample splitting methods.** Certain randomization tests, such as the holdout randomization test (HRT) of Tansey et al., require sample splitting. Sample splitting poses a challenge: if one splits the sample only once, the resulting p-value is non-reproducible. On the other hand, if one splits the sample multiple times, thereby producing multiple p-values, one cannot easily combine these p-values into a single p-value. e-values help to resolve this difficulty (Vovk 2020). Let $e_1, \ldots, e_m$ be the e-values produced by m sample splits. Then $\bar{e} = (1/m) \sum_{i=1}^{m} e_i$ is itself an e-value. As m tends to infinity, $\bar{e}$ converges (under some conditions) to a constant, de-randomizing the sample splitting procedure.

Importantly, the above benefits can be mixed-and-matched at our discretion. Figure 1 summarizes the relationships between the test statistics considered in this proposal.

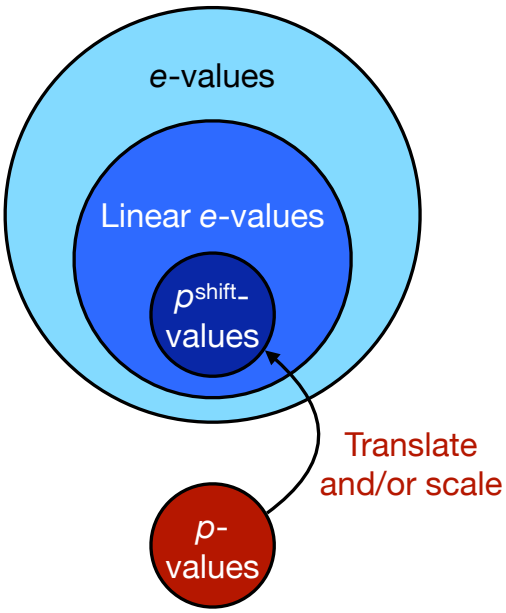

Figure 1: Connection between the various test statistics considered in this proposal: p^{shift}-values $\subset$ linear e-values $\subset$ e-values. Meanwhile, p^{shift}-values are scaled and/or translated p-values with unit expectation.

Direction 2: A class of fast and powerful test statistics for randomization tests

Coming soon...

Tertiary directions

We note several possible other directions that could be of interest.

1. **An analog of Selective SeqStep for e-values.** Selective SeqStep is a multiple testing correction procedure for controlling FDR proposed by Barber and Candes (Barber and Candés 2015). Li and Candes recommend using Selective SeqStep instead of BH to correct p-values outputted by the CRT, as Selective SeqStep better handles discrete p-values (Li and Candès 2021). Wang and Ramdas developed an analog of BH for e-values; might it be possible to develop an analog of Selective SeqStep for e-values? Such an extension would enable users of the proposed framework to choose between different multiple testing correction procedures, which would be helpful.

2. **Testing and training on the same data.** Our goal in applying the CRT is to test for conditional independence of X and Y given Z. Under what conditions can we train the regression function $\hat{E}(X|Z)$ and test the conditional independence hypothesis using the same data? Intuitively, the variance of the regression function $\hat{E}(X|Z)$ must be small. Can we derive a practical (e.g., bootstrap-based) methodology to help users determine whether

or not to sample split?

3. **Double robustness and double regression.** Under the right choice of the test statistic, the CRT is doubly-robust in an asymptotic (i.e., large n, fixed p) regime. Can we derive more precise, non-asymptotic or high-dimensional statements about double robustness a la double machine learning (DML; Chernozhukov et al. 2018) or the generalized covariance measure (GCM; Shah and Peters 2020)? More generally, CRT, GCM, and DML at their core are methods for causal inference based on flexible double regression. What are the connections and relative strengths and weaknesses of these approaches?

4. **Theory for p-value vs. e-value calibration**. It is known that exchangeability of the resampled test statistics implies implies validity of the resulting p-value. Can we answer the following questions? (i) Is exchangeability *necessary* for the p-value to be valid? (ii) What are the conditions under which an e-value is valid? In particular, does e-value validity require exchangeability? Intuitively, we can construct the e-value in such a way that only some (rather than all) quantiles of the distribution are used. (iii) Is it possible for us to obtain a bound on the type-I error of an e-value a la Berrett et al. (2020) or Kim et al. (2021)? This statement might be in terms of expectation rather than probability.

5. **The local holdout permutation test**. A simple, new conditional independence test that results from combining the holdout randomization test (Tansey et al. 2021) with the local permutation test (Kim et al. 2021) is what I call the "local holdout permutation test." The idea is as follows: (1) split the data in half, (2) train a classifier on the first half, (3) run the local permutation test on the second half, using empirical risk as the test statistic.

6. **Extensions to other rank-based tests**. Can we extend the ideas here to other rank-based tests? For example, the Mann-Whitney U test, whose test statistics basically is a sum of ranks?

References

Barber, Rina Foygel and Emmanuel J. Candés (2015). "Controlling the false discovery rate via knockoffs". In: *Annals of Statistics* 43.5, pp. 2055–2085 (cit. on p. 7).

Berrett, Thomas B., Yi Wang, Rina Foygel Barber, and Richard J. Samworth (2020). "The conditional permutation test for independence while controlling for confounders". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.1, pp. 175–197 (cit. on p. 8).

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters". In: *Econometrics Journal* 21.1, pp. C1–C68 (cit. on p. 8).

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).

Kim, Ilmun, Matey Neykkov, Sivaraman Balakrishnan, and Larry Wasserman (2021). "Local permutation tests for conditional independence". In: (cit. on p. 8).

Kuchibhotla, Arun Kumar (2020). "Exchangeability, Conformal Prediction, and Rank Tests". In: pp. 1–36 (cit. on p. 2).

Li, Shuangning and Emmanuel J. Candès (2021). "Deploying the Conditional Randomization Test in High Multiplicity Problems". In: pp. 1–43 (cit. on p. 7).

Maleki, Farhad, Katie Ovens, Daniel J. Hogan, and Anthony J. Kusalik (2020). "Gene Set Analysis: Challenges, Opportunities, and Future Research". In: *Frontiers in Genetics* 11.June, pp. 1–16 (cit. on p. 1).

Shah, Rajen D. and Jonas Peters (2020). "The hardness of conditional independence testing and the generalised covariance measure". In: *Annals of Statistics* 48.3, pp. 1514–1538 (cit. on p. 8).

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 pp. 1, 8).

Vovk, Vladimir (2020). "A note on data splitting with e-values". In: pp. 1–11 (cit. on pp. 1, 6).

Vovk, Vladimir and Ruodu Wang (2021). "E-values: Calibration, combination and applications". In: *Annals of Statistics* 49.3, pp. 1736–1754 (cit. on pp. 1, 4).

Wang, Ruodu and Aaditya Ramdas (2020). "False discovery rate control with e-values". In: pp. 1–32 (cit. on pp. 1, 6).

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

Proposition 2 proof

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

$$\prod_{i=1}^{B} I_i^{k_i} = I_{\tau(k_1,\ldots,k_B)}. \tag{3}$$

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

$$\prod_{i=1}^{B} I_i^{k_i} = \prod_{i=1}^{N} I_{\sigma(i)}^{k_{\sigma(i)}}, \tag{4}$$

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 4 therefore reduces to

$$\prod_{i=1}^{N} I_{\sigma(i)}^{k_{\sigma(i)}} = I_{\sigma(1)}. \tag{5}$$

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 (4) and (5) yields the conclusion (3).

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

$$\begin{aligned}\left(\sum_{i=1}^{B} I_i\right)^r &= \sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} \prod_{i=1}^{B} I_i^{k_i} \\ &= \sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} I_{\tau(k_1,\ldots,k_B)}, \end{aligned} \tag{6}$$

where the second equality follows from (3). The final term in (6) 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

$$\begin{aligned}\sum_{(k_1,\ldots,k_B):k_1+\cdots+k_B=r} \binom{r}{k_1, k_2, \ldots, k_B} I_{\tau(k_1,\ldots,k_B)} &= \sum_{i=1}^{B} \sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1, k_2, \ldots, k_B} I_i \\ &= \sum_{i=1}^{B} I_i \sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1, k_2, \ldots, k_B}. \end{aligned} \tag{7}$$

We evaluate the inner sum (7), which is the coefficient corresponding to I_i in the linear combination. We have that

$$\sum_{(k_1,\ldots,k_B)\in\tau^{-1}(i)} \binom{r}{k_1,k_2,\ldots,k_B} = \sum_{(0,\ldots,0,k_i,\ldots,k_B):k_i\geq 1,k_i+\cdots+k_B=r} \binom{r}{k_1,k_2,\ldots,k_B}$$

$$= \sum_{(k_1,\ldots,k_{B-i+1}):k_1\geq 1,k_1+\cdots+k_{B-i+1}=r} \binom{r}{k_1,k_2,\ldots,k_{B-i+1}} = \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \binom{r}{j,l_1,\ldots,l_{B-i}}$$

$$= \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r!}{j!l_1!\ldots l_{B-i}!} = \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r(r-1)\ldots(r-j+1)(r-j)!}{j!l_1!\ldots l_{B-i}!}$$

$$= \sum_{j=1}^{r} \sum_{l_1+\cdots+l_{B-i}=r-j} \frac{r!}{(r-j)!j!}\binom{r-j}{l_1,\ldots,l_{B-i}} = \sum_{j=1}^{r} \frac{r!}{(r-j)!j!} \sum_{l_1,\ldots,l_{B-i}} \binom{r-j}{l_1,\ldots,l_{B-i}}$$

$$= \sum_{j=1}^{r} \binom{r}{j}(B-i)^{r-j} = \sum_{j=1}^{r} \binom{r}{j}(B-i)^{r-j}1^j = \sum_{j=0}^{r} \binom{r}{j}(B-i)^{r-j}1^j - 1(B-i)^r$$

$$= (B-i+1)^r - (B-i)^r. \quad (8)$$

Combining (6), (7), and (8), 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

$$(p_B - x_0)^j = \sum_{k=0}^{j} \binom{j}{k} p_B^k x_0^{j-k}(-1)^{j-k} = x_0^j(-1)^j + \sum_{k=1}^{j} \binom{j}{k} p_B^k x_0^{j-k}(-1)^{j-k}$$

$$= x_0^j(-1)^j + \sum_{k=1}^{j} \binom{j}{k} \left[\sum_{i=1}^{B} M(B,k,i) I_i\right] x_0^{j-k}(-1)^{j-k}$$

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

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

$$\sum_{j=0}^{r} c_j\,(p_B - x_0)^j = c_0 + \sum_{j=1}^{r} c_j (p_B - x_0)^j$$

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

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

completing the proof.

A Beginning theory on calibration

Let $W_1, \ldots, W_n$ be exchangeable. For $t \in \mathbb{R}$, we have that

$$\mathbb{P}(\text{rank}(W_n) \leq t) = \frac{\lfloor t \rfloor}{n}.$$

Define $P_n := \text{rank}(W_n)/n$. Then

$$\mathbb{P}(P_n \leq \alpha) = \mathbb{P}(\text{rank}(W_n)/n \leq \alpha) = \mathbb{P}(\text{rank}(W_n) \leq n\alpha) = \frac{\lfloor n\alpha \rfloor}{n} \leq \frac{n\alpha}{n} = \alpha.$$

Therefore, P_n is a valid p-value. Moreover, P_n converges in distribution to $U[0,1]$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}(P_n \leq \alpha) = \alpha.$$

By way of proof, let $\epsilon > 0$ be given. Let $N = 1/\epsilon$, and let $n > N$ for $n \in \mathbb{N}$. Then

$$\left| \alpha - \frac{\lfloor n\alpha \rfloor}{n} \right| = \left| \frac{n\alpha}{n} - \frac{\lfloor n\alpha \rfloor}{n} \right| \leq 1/n \text{ (because } x - \lfloor x \rfloor \leq 1 \text{)} < \epsilon.$$

Conversely, let $(W_1, \ldots, W_n)$ be a sequence of random variables. Define $P_n := \text{rank}(W_n)/n$, as before. Assume that

$$\lim_{n\to\infty} \mathbb{P}\left(P_n \leq \alpha\right) = \lim_{n\to\infty} \mathbb{P}(\text{rank}(W_n)/n \leq \alpha) = \alpha.$$

We want to show that $(W_1, \ldots, W_n)$ are exchangeable.