A Lego System for Conditional Inference

Torsten Hothorn[1], Kurt Hornik[2],

Mark A. van de Wiel[3] and Achim Zeileis[2]

[1] Institut für Medizininformatik, Biometrie und Epidemiologie

Friedrich-Alexander-Universität Erlangen-Nürnberg

Waldstraße 6, D-91054 Erlangen, Germany

Torsten.Hothorn@R-project.org

[2] Department für Statistik und Mathematik, Wirtschaftsuniversität Wien

Augasse 2-6, A-1090 Wien, Austria

Kurt.Hornik@R-project.org

Achim.Zeileis@R-project.org

[3] Department of Mathematics, Vrije Universiteit

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

mark.vdwiel@vumc.nl

Abstract

Conditioning on the observed data is an important and flexible design principle for statistical test procedures. Although generally applicable, permutation tests currently in use are limited to the treatment of special cases, such as contingency tables or K-sample problems. A new theoretical framework for permutation tests opens up the way to a unified and

generalized view. We argue that the transfer of such a theory to practical data analysis has important implications in many applications and requires tools that enable the data analyst to compute on the theoretical concepts as closely as possible. We re-analyze four data sets by adapting the general conceptual framework to these challenging inference problems and utilizing the **coin** add-on package in the R system for statistical computing to show what one can gain from going beyond the 'classical' test procedures.

KEY WORDS: Permutation tests; Independence; Asymptotic distribution; Software.

1 INTRODUCTION

The distribution of a test statistic under the circumstances of a null hypothesis clearly depends on the unknown distribution of the data and thus is unknown as well. Two concepts are commonly applied to dispose of this dependency. Unconditional tests impose assumptions on the distribution of the data such that the null distribution of a test statistic can be derived analytically. In contrast, conditional tests replace the unknown null distribution by the conditional null distribution, i.e., the distribution of the test statistic given the observed data. The latter approach is known as *permutation testing* and was developed by R. A. Fisher more than 70 years ago (**?**). The pros and cons of both approaches in different fields of application have been widely discussed (e.g. by **???**). Here, we focus on the practical aspects of permutation testing rather than dealing with its methodological foundations.

For the construction of permutation tests it is common exercise to 'recycle' test statistics well known from the unconditional world, such as linear rank

statistics, ANOVA F statistics or χ^2 statistics for contingency tables, and to replace the unconditional null distribution with the conditional distribution of the test statistic under the null hypothesis (**????**). Because the choice of the test statistic is the only 'degree of freedom' for the data analyst, the classical view on permutation tests requires a 'cook book' classification of inference problems (categorical data analysis, multivariate analysis, K-sample location problems, correlation, etc.), each being associated with a 'natural' form of the test statistic.

The theoretical advances of the last decade (notably **???**) give us a much better understanding of the strong connections between the 'classical' permutation tests defined for different inference problems. As we will argue in this paper, the new theoretical tools open up the way to a simple construction principle for test procedures in new and challenging inference problems. Especially attractive for this purpose is the theoretical framework for permutation tests developed by **?**. This unifying theory is based on a flexible form of multivariate linear statistics for the general independence problem.

This framework provides us with a conceptual Lego system for the construction of permutation tests consisting of Lego bricks for linear statistics suitable for different inference problems (contingency tables, multivariate problems, etc.), different forms of test statistics (such as quadratic forms for global tests or test statistics suitable for multiple comparison procedures), and several ways to derive the conditional null distribution (by means of exact computations or approximations). The classical procedures, such as a permutation t test, are part of this framework and, even more interesting, new test procedures can be embedded into the same theory whose main ideas are sketched in Section **??**.

Currently, the statistician's toolbox consists of rather specialized spanners, such as the Wilcoxon-Mann-Whitney test for comparing two distributions or the Cochran-Mantel-Haenszel χ^2 test for independence in contingency tables. With this work, we add an adjustable spanner to the statistician's toolbox which helps

to address both the common as well as new or unusual inference problems with the appropriate conditional test procedures. In the main part of this paper we show how one can construct and implement permutation tests 'on the fly' by plugging together Lego bricks for the multivariate linear statistic, the test statistic and the conditional null distribution, both conceptually and practically by means of the **coin** add-on package (**?**) in the R system for statistical computing (**?**).

2 A CONCEPTUAL LEGO SYSTEM

To fix notations, we assume that we are provided with independent and identically distributed observations $(\mathbf{Y}_i, \mathbf{X}_i)$ for $i = 1, \ldots, n$. The variables $\mathbf{Y}$ and $\mathbf{X}$ from sample spaces $\mathcal{Y}$ and $\mathcal{X}$ may be measured at arbitrary scales and may be multivariate as well. We are interested in testing the null hypothesis of independence of $\mathbf{Y}$ and $\mathbf{X}$

$$H_0 : D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y})$$

against arbitrary alternatives. **?** suggest to derive *scalar* test statistics for testing H_0 from *multivariate* linear statistics of the form

$$\mathbf{T} = \text{vec}\left(\sum_{i=1}^{n} g(\mathbf{X}_i) h(\mathbf{Y}_i)^\top\right) \in \mathbb{R}^{pq \times 1}.$$

Here, $g : \mathcal{X} \to \mathbb{R}^{p \times 1}$ is a transformation of the $\mathbf{X}$ measurements and $h : \mathcal{Y} \to \mathbb{R}^{q \times 1}$ is called *influence function*. The function $h(\mathbf{Y}_i) = h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))$ may depend on the full vector of responses $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, however only in a permutation symmetric way, i.e., the value of the function must not depend on the order in which $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ appear. We will give several examples how to choose g and h for specific inference problems in Section **??**.

The distribution of $\mathbf{T}$ depends on the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$, which is unknown under almost all practical circumstances. At least under the null

hypothesis one can dispose of this dependency by fixing $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and conditioning on all possible permutations S of the responses $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Tests that have been constructed by means of this conditioning principle are called *permutation tests*.

The conditional expectation $\mu \in \mathbb{R}^{pq \times 1}$ and covariance $\Sigma \in \mathbb{R}^{pq \times pq}$ of $\mathbf{T}$ under H_0 given all permutations $\sigma \in S$ of the responses are derived by **?**:

$$\begin{aligned}
\mu = \mathbb{E}(\mathbf{T}|S) &= \operatorname{vec}\left(\left(\sum_{i=1}^{n} g(\mathbf{X}_i)\right)\mathbb{E}(h|S)^\top\right) \\
\Sigma = \mathbb{V}(\mathbf{T}|S) &= \frac{n}{n-1}\mathbb{V}(h|S) \otimes \left(\sum_i g(\mathbf{X}_i) \otimes g(\mathbf{X}_i)^\top\right) \\
&- \frac{1}{n-1}\mathbb{V}(h|S) \otimes \left(\sum_i g(\mathbf{X}_i)\right) \otimes \left(\sum_i g(\mathbf{X}_i)\right)^\top
\end{aligned}$$

where $\otimes$ denotes the Kronecker product, and the conditional expectation of the influence function is $\mathbb{E}(h|S) = n^{-1}\sum_i h(\mathbf{Y}_i)$ with corresponding $q \times q$ covariance matrix

$$\mathbb{V}(h|S) = n^{-1}\sum_i \left(h(\mathbf{Y}_i) - \mathbb{E}(h|S)\right)\left(h(\mathbf{Y}_i) - \mathbb{E}(h|S)\right)^\top.$$

The key step for the construction of test statistics based on the multivariate linear statistic $\mathbf{T}$ is its standardization utilizing the conditional expectation μ and covariance matrix Σ. Univariate test statistics c mapping a linear statistic $\mathbf{T} \in \mathbb{R}^{pq \times 1}$ into the real line can be of arbitrary form. Obvious choices are the maximum of the absolute values of the standardized linear statistic or a quadratic form:

$$\begin{aligned}
c_{\max}(\mathbf{T}, \mu, \Sigma) &= \max\left|\frac{\mathbf{T} - \mu}{\operatorname{diag}(\Sigma)^{1/2}}\right|, \\
c_{\text{quad}}(\mathbf{T}, \mu, \Sigma) &= (\mathbf{T} - \mu)\Sigma^+(\mathbf{T} - \mu)^\top,
\end{aligned}$$

involving the Moore-Penrose inverse Σ^+ of Σ.

The conditional distribution $\mathbb{P}(c(\mathbf{T}, \mu, \Sigma) \leq z|S)$ is the number of permutations $\sigma \in S$ of the data with corresponding test statistic not exceeding z divided

by the total number of permutations in S. For some special forms of the multivariate linear statistic the exact distribution of some test statistics is tractable for small and moderate sample sizes. In principle, resampling procedures can always be used to approximate the exact distribution up to any desired accuracy by evaluating the test statistic for a random sample from the set of all permutations S. It is important to note that in the presence of a grouping of the observations into independent blocks, only permutations within blocks are eligible and that the conditional expectation and covariance matrix need to be computed separately for each block.

Less well known is the fact that a normal approximation of the conditional distribution can be computed for arbitrary choices of g and h. **?** showed in their Theorem 2.3 that the conditional distribution of linear statistics $\mathbf{T}$ with conditional expectation μ and covariance Σ tends to a multivariate normal distribution with parameters μ and Σ as $n \to \infty$. Thus, the asymptotic conditional distribution of test statistics of the form $c_{\max}$ is normal and can be computed directly in the univariate case ($pq = 1$) and by numerical algorithms in the multivariate case (**?**). For quadratic forms c_{quad} which follow a χ^2 distribution with degrees of freedom given by the rank of Σ (see **?**, Chapter 29), exact probabilities can be computed efficiently.

3 PLAYING LEGO

The Lego system sketched in the previous section consists of Lego bricks for the multivariate linear statistic $\mathbf{T}$, namely the transformation g and influence function h, multiple forms of the test statistic c and several choices of approximations of the null distribution. In this section, we will show how classical procedures, starting with the conditional Kruskal-Wallis test and the Cochran-Mantel-Haenszel test, can be embedded into this general theory and, much more