coin: A computational framework for conditional inference

Torsten Hothorn[1], Kurt Hornik[2], Mark 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]Institut 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 and Computer Science
Eindhoven University of Technology
HG 9.25, P.O. Box 513
5600 MB Eindhoven, The Netherlands
markvdw@win.tue.nl

1 Introduction

2 Permutation Tests

$(\mathbf{Y}_i, \mathbf{X}_i, w_i, b_i), i = 1, \dots, n)$

$$\mathbf{T} = \mathrm{vec}\left(\sum_{i=1}^{n} w_i g(\mathbf{X}_i) h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^p q \qquad (1)$$

The conditional expectation $\mu \in \mathbb{R}^{pq}$ 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 **?**:

$$
\begin{aligned}
\mu &= \mathbb{E}(\mathbf{T}|S) = \operatorname{vec}\left(\left(\sum_{i=1}^{n} w_i g(\mathbf{X}_i)\right) \mathbb{E}(h|S)^\top\right), \\
\Sigma &= \mathbb{V}(\mathbf{T}|S) \\
&= \frac{\mathbf{w}.}{\mathbf{w}. - 1} \mathbb{V}(h|S) \otimes \left(\sum_i w_i g(\mathbf{X}_i) \otimes w_i g(\mathbf{X}_i)^\top\right) \\
&- \frac{1}{\mathbf{w}. - 1} \mathbb{V}(h|S) \otimes \left(\sum_i w_i g(\mathbf{X}_i)\right) \otimes \left(\sum_i w_i g(\mathbf{X}_i)\right)^\top
\end{aligned} \tag{2}
$$

where $\mathbf{w}. = \sum_{i=1}^{n} w_i$ denotes the sum of the case weights, and $\otimes$ is the Kronecker product. The conditional expectation of the influence function is

$$
\mathbb{E}(h|S) = \mathbf{w}.^{-1} \sum_i w_i h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) \in \mathbb{R}^q
$$

with corresponding $q \times q$ covariance matrix

$$
\begin{aligned}
\mathbb{V}(h|S) = \mathbf{w}.^{-1} \sum_i w_i \left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) - \mathbb{E}(h|)\right) \\
\left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \dots, \mathbf{Y}_n)) - \mathbb{E}(h|S)\right)^\top.
\end{aligned}
$$

Having the conditional expectation and covariance at hand we are able to standardize a linear statistic $\mathbf{T} \in \mathbb{R}^{pq}$ of the form (1). Univariate test statistics c mapping an observed linear statistic $\mathbf{t} \in \mathbb{R}^{pq}$ into the real line can be of arbitrary form. An obvious choice is the maximum of the absolute values of the standardized linear statistic

$$
c_{\max}(\mathbf{t}, \mu, \Sigma) = \max \left| \frac{\mathbf{t} - \mu}{\operatorname{diag}(\Sigma)^{1/2}} \right|
$$

utilizing the conditional expectation μ and covariance matrix Σ. The application of a quadratic form $c_{\text{quad}}(\mathbf{t}, \mu, \Sigma) = (\mathbf{t} - \mu)\Sigma^+(\mathbf{t} - \mu)^\top$ is one alternative, although computationally more expensive because the Moore-Penrose inverse Σ^+ of Σ is involved.

The conditional distribution and thus the P-value of the statistics $c(\mathbf{t}, \mu, \Sigma)$ can be computed in several different ways. For some special forms of the linear statistic, the exact distribution of the test statistic is trackable. Conditional Monte-Carlo procedures can be used to approximate the exact distribution. **?** proved (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 $n, s \to \infty$. Thus, the asymptotic conditional distribution of test statistics of the form $c_{\max}$ is normal and can be computed directly in the univariate case ($pq = 1$) or approximated by means of quasi-randomized Monte-Carlo procedures in the multivariate setting (**?**). For quadratic forms c_{quad} which follow a χ^2 distribution with degrees of freedom given by the rank of Σ (Theorem 6.20, **?**), exact probabilities can be computed efficiently.

3 Examples

Independent K-Sample Problems $\mathbf{Y}$ is univariate numeric (or censored) and $\mathbf{X}$ a factor at K levels. g is the dummy matrix and h by be arbitrary.

```
R> library(coin)

Loading required package: survival
Loading required package: splines
Loading required package: mvtnorm

R> YOY <- data.frame(length = c(46, 28, 46, 37, 32, 41, 42, 45,
+     38, 44, 42, 60, 32, 42, 45, 58, 27, 51, 42, 52, 38, 33, 26,
+     25, 28, 28, 26, 27, 27, 27, 31, 30, 27, 29, 30, 25, 25, 24,
+     27, 30), site = factor(c(rep("I", 10), rep("II", 10), rep("III",
+     10), rep("IV", 10))))
R> kruskal_test(length ~ site, data = YOY)

        Asymptotical Kruskal-Wallis Test

data:  length by groups I, II, III, IV
T = 22.8524, df = 3, p-value = 4.335e-05

R> it <- independence_test(length ~ site, data = YOY, ytrafo = function(data) trafo(data,
+     numeric_trafo = rank), teststat = "quadtype")
R> statistic(it, "linear")

    [,1]
I    278
II   307
III  119
IV   116

R> expectation(it)

    [,1]
I    205
II   205
III  205
IV   205

R> covariance(it)

          [,1]      [,2]      [,3]      [,4]
[1,] 1019.0385 -339.6795 -339.6795 -339.6795
[2,] -339.6795 1019.0385 -339.6795 -339.6795
[3,] -339.6795 -339.6795 1019.0385 -339.6795
[4,] -339.6795 -339.6795 -339.6795 1019.0385
```

```
R> statistic(it, "standardized")

          [,1]
I     2.286797
II    3.195250
III -2.694035
IV  -2.788013

R> statistic(it)

[1] 22.85242

R> pvalue(it)

[1] 4.334659e-05
```

Independence in Contingency Tables

Ordered Alternatives

Multivariate Problems