`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

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

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

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

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 **?**:

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

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.

```
> library(coin)

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

> 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))))
> 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

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

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

> expectation(it)

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

> 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
```

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

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

> statistic(it)

[1] 22.85242

> pvalue(it)

[1] 4.334659e-05
```

Independence in Contingency Tables

```
> data(jobsatisfaction)
> jobsatisfaction

, , Gender = Female

             Job.Satisfaction
Income        Very Dissatisfied A Little Dissatisfied Moderately Satisfied
  <5000                       1                     3                   11
  5000-15000                  2                     3                   17
  15000-25000                 0                     1                    8
  >25000                      0                     2                    4
             Job.Satisfaction
Income        Very Satisfied
  <5000                    2
  5000-15000               3
  15000-25000              5
  >25000                   2

, , Gender = Male

             Job.Satisfaction
Income        Very Dissatisfied A Little Dissatisfied Moderately Satisfied
  <5000                       1                     1                    2
  5000-15000                  0                     3                    5
  15000-25000                 0                     0                    7
  >25000                      0                     1                    9
             Job.Satisfaction
Income        Very Satisfied
  <5000                    1
```

```
  5000-15000                1
  15000-25000               3
  >25000                    6
```

```
> it <- cmh_test(jobsatisfaction)
> it

        Asymptotical Generalised Cochran-Mantel-Haenszel Test

data:  Job.Satisfaction by groups <5000, 5000-15000, 15000-25000, >25000 stratified by Gende
T = 10.2001, df = 9, p-value = 0.3345

> statistic(it, "standardized")

            Very Dissatisfied A Little Dissatisfied Moderately Satisfied
<5000               1.3112789            0.69201053           -0.2478705
5000-15000          0.6481783            0.83462550            0.5175755
15000-25000        -1.0958361           -1.50130926            0.2361231
>25000             -1.0377629           -0.08983052           -0.5946119
            Very Satisfied
<5000           -0.9293458
5000-15000      -1.6257547
15000-25000      1.4614123
>25000           1.2031648
```

Ordered Alternatives

```
> lbl_test(jobsatisfaction)

        Asymptotical Linear-by-Linear Association Test

data:  Job.Satisfaction (ordered) by groups <5000 < 5000-15000 < 15000-25000 < >25000 strati
T = 6.6235, df = 1, p-value = 0.01006

> lbl_test(jobsatisfaction, scores = list(Job.Satisfaction = c(1,
+     3, 4, 5), Income = c(3, 10, 20, 35)))

        Asymptotical Linear-by-Linear Association Test

data:  Job.Satisfaction (ordered) by groups <5000 < 5000-15000 < 15000-25000 < >25000 strati
T = 6.1563, df = 1, p-value = 0.01309
```

Multivariate Problems