WIRTSCHAFTS UNIVERSITÄT WIEN

A Lego System for Conditional Inference

Torsten Hothorn, Kurt Hornik, Mark van de Wiel, Achim Zeileis

`http://statmath.wu-wien.ac.at/~zeileis/`

Overview

- Conditional inference
    - A conceptual Lego system
    - Independence problem
    - A unified framework
    - History
    - From conceptual to computational Lego bricks
- Playing Lego
    - Independent *k* samples: Genetic components of alcoholism
    - Contingency tables: Smoking and Alzheimer's disease
    - Multivariate response: Photococarcinogenicity experiments
    - Independent 2 samples: Contaminated fish consumption
    - Maximally selected statistics: Tree pipit abundance
    - Generalized maximally selected statistics: High- and low-risk groups of rectal cancer patients
- Concluding remarks

A conceptual Lego system

Hothorn, Hornik, van de Wiel, and Zeileis (2006) discuss a unified approach to conditional inference in the independence problem based on the theory of Strasser and Weber (1999). This theory unifies a wide collection of classical and modern non-parametric test procedures.

The theory utilizes various components that can be put together like Lego bricks for a specific problem:

- influence function for response,
- transformation of explanatory variable,
- aggregation to test statistic,
- type of null distribution (exact, asymptotic, approximate).

Hothorn *et al.* (2006) provide an implementation in the R package **coin** enabling the construction of known and new test procedures “on the fly”.

Independence problem

Null hypothesis: Independence of two variables Y and X (both possibly multivariate).

$$H_0 : D(Y|X) = D(Y).$$

Two models are typically distinguished:

- Population model: X codes well-defined populations from which random samples can be drawn.
- Randomization model: X is the randomization result (e.g., treatment arm in a clinical trial).

A class of linear statistics

For Y and X from populations $\mathcal{Y}$ and $\mathcal{X}$ a linear statistic for assessing departures from H_0 can be defined:

$$T = \operatorname{vec}\left(\sum_{i=1}^{n} w_i g(X_i) h(Y_i)^\top\right) \in \mathbb{R}^{pq}$$

with

- weights $w_i \in \mathbb{R}$,
- transformation $g : \mathcal{X} \to \mathbb{R}^p$,
- influence function $h : \mathcal{Y} \to \mathbb{R}^q$.

Problem: The distribution of T depends on the joint distribution of Y and X and is thus typically unknown in practice (unless further assumptions are imposed).

Conditional null distribution

Solution: Use conditional distribution of T given the observed data.

Under H_0, all permutations S of Y yield the conditional distribution of T. It has mean $\mu \in \mathbb{R}^{pq}$:

$$\begin{aligned} \mu = \mathbb{E}(T|S) &= \operatorname{vec}\left(\left(\sum_{i=1}^{n} w_i g(X_i)\right)\mathbb{E}(h|S)^\top\right), \\ \mathbb{E}(h|S) &= w_+^{-1}\sum_i w_i h(Y_i), \end{aligned}$$

where $w_+ = \sum_{i=1}^n w_i$.

This can be easily computed for a given problem.

Conditional null distribution

Similarly, the conditional covariance matrix $\Sigma \in \mathbb{R}^{pq \times pq}$ under H_0 is:

$$\begin{aligned}
\Sigma = \mathbb{V}(T|S) &= \frac{w_+}{w_+ - 1} \mathbb{V}(h|S) \otimes \left(\sum_i w_i g(X_i) \otimes w_i g(X_i)^\top \right) - \\
&\quad \frac{1}{w_+ - 1} \mathbb{V}(h|S) \otimes \left(\sum_i w_i g(X_i) \right) \otimes \left(\sum_i w_i g(X_i) \right)^\top, \\
\mathbb{V}(h|S) &= w_+^{-1} \sum_i w_i \left(h(Y_i) - \mathbb{E}(h|S) \right) \left(h(Y_i) - \mathbb{E}(h|S) \right)^\top,
\end{aligned}$$

where $\otimes$ denotes the Kronecker product.

Aggregation to test statistic

To aggregate an observed linear statistic T to a scalar test statistic, the following strategies seem natural:

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

where Σ^{+} is the Moore-Penrose inverse of Σ.

Asessing the test statistic

Various approaches can be used to assess the significance of *c*.

- **Exact:** Direct computation of *c* for all permutations *S* is typically burdensome but special algorithms are available for certain problems (e.g., shift algorithm for 2-sample problems).
- **Approximate:** Compute *c* for a sufficiently large number of permutations from *S*, drawn using Monte Carlo methods.
- **Asymptotic:** Compute the conditional asymptotic distribution of *c* based on the asymptotic conditional distribution of *T*.
  $T \sim \mathcal{N}(\mu, \Sigma)$ for $n \to \infty$.

History

The ideas underlying this unified theory are not new. In fact, permutation methods have been discussed in the literature since the 1930s.

Example: For a 2-sample problem, $g(X)$ is typically chosen as the indicator function for the two samples. If $h(Y) = Y$, this yields a t statistic (using the 1-sample standard deviation).

- The exact unconditional distribution under the assumption of normality was famously derived by Gosset in 1908.
- In the 1930s, Fisher suggested to use the exact conditional distribution instead.
- Already in 1937 Pitman and Welch published results about the asymptotic properties of the conditional approach in Biometrika.

History

Problem: Hard to compute and thus not used for a long time.

Idea in mid-1900s: Use $h(Y) = \text{rank}(Y)$, then the exact conditional distribution (for data without ties) can be computed by recursion formulas.

Justification: Ranks introduce robustness (for certain types of departures from normality) in the procedures.

Since late 1900s: Increased interest again in conditional inference methods. Permutations become feasible much more generally by using new algorithms and increased computing power of modern PCs.

Problem: Many implementations of permutation tests are focused on specific test problems.

From conceptual to computational Lego bricks

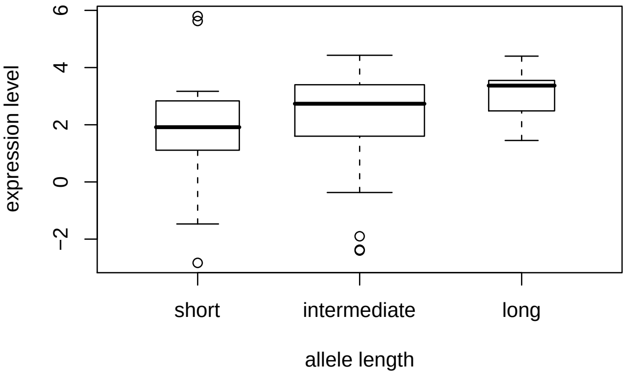
In R, package **coin** provides an implementation that reflects the flexibility of the conceptual tools. The workhorse function is

```
independence_test(
 formula        y ~ x | block
 ytrafo         influence function for Y
 xtrafo         transformation of X
 teststat       "max" or "quad"
 distribution   exact(), approximate() or asymptotic()
)
```

This can be employed for computing well-known and new test procedures *without* explicitly implementing the specific null distribution.

Genetic components of alcoholism

Bönsch *et al.* (2005) study the association of allele length and expression levels of alpha synuclein mRNA, a gene linked to alcoholism. Allele length was discretized: short (0–4, $n = 24$), intermediate (5–9, $n = 58$), long (10–12, $n = 15$).



Genetic components of alcoholism

Use Kruskal-Wallis test for assessing the association:

```
R> library("coin")
R> independence_test(elevel ~ alength, data = alpha,
+    ytrafo = rank, teststat = "quad")

Asymptotic General Independence Test

data:  elevel by
 alength (short, intermediate, long)
chi-squared = 8.83, df = 2, p-value = 0.01209
```

`xtrafo` is chosen as the indicator function for the categorical variable `alength` by default.

Genetic components of alcoholism

Convenience interface:

```
R> kt <- kruskal_test(elevel ~ alength, data = alpha)
R> kt

Asymptotic Kruskal-Wallis Test

data:  elevel by
 alength (short, intermediate, long)
chi-squared = 8.83, df = 2, p-value = 0.01209
```

The underlying conceptual components can be easily recovered:

```
R> statistic(kt)

[1] 8.83

R> pvalue(kt)

[1] 0.01209
```

Genetic components of alcoholism

```
R> statistic(kt, type = "linear")

short          900.5
intermediate 2878.5
long           974.0

R> expectation(kt)

       short intermediate         long
        1176         2842          735

R> covariance(kt)

              short intermediate  long
short         14305       -11366 -2939
intermediate -11366        18469 -7104
long          -2939        -7104 10043
```

Genetic components of alcoholism

Question: The Kruskal-Wallis test has long been available in R (in `kruskal.test()`), so what is the advantage of using **coin**?

Answer: Going beyond the classical functionality is easy in **coin** (and would otherwise require extensive programming), e.g.:

- Use original observations instead of ranks.
- Use the resampling distribution instead of the asymptotic distribution.
- Exploit the ordered nature of the allele length using numeric scores (interval midpoints), similar to linear-by-linear association tests.

Genetic components of alcoholism

Use original observations instead of ranks:

```
R> independence_test(elevel ~ alength, data = alpha,
+     teststat = "quad")

Asymptotic General Independence Test

data:  elevel by
 alength (short, intermediate, long)
chi-squared = 5.056, df = 2, p-value = 0.07981
```

The default `ytrafo` is to use the identity for numeric variables like `elevel`.

Genetic components of alcoholism

Use the resampling distribution:

```
R> set.seed(123)
R> pvalue(independence_test(elevel ~ alength,
+     data = alpha, teststat = "quad",
+     distribution = approximate(B = 19999)))

[1] 0.07835
99 percent confidence interval:
 0.07354 0.08337
```

Genetic components of alcoholism

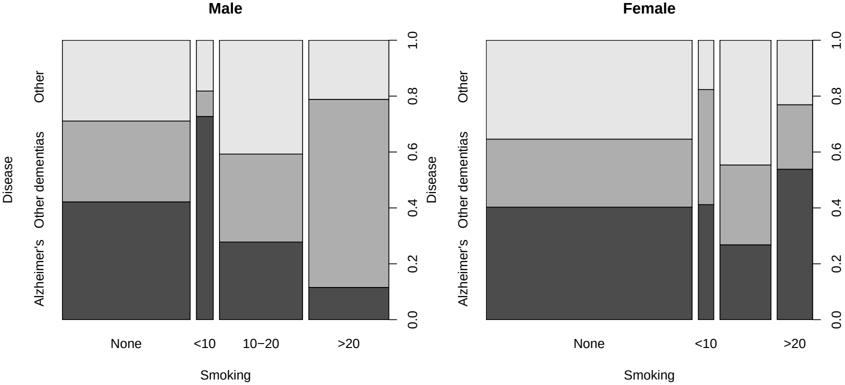
Use numeric scores for ordered alternative:

```
R> mpoints <- function(x) c(2, 7, 11)[unlist(x)]
R> independence_test(elevel ~ alength, data = alpha,
+     teststat = "quad", xtrafo = mpoints,
+     distribution = approximate(B = 19999))
```

```
Approximative General Independence Test

data:  elevel by
 alength (short, intermediate, long)
chi-squared = 4.626, p-value = 0.02915
```

Smoking and Alzheimer’s disease

Salib and Hillier (1997) report results of a case-control study on Alzheimer’s disease and smoking behaviour of 198 patients and 164 controls.



Smoking and Alzheimer's disease

Use the Chochran-Mantel-Haenszel test for assessing the independence between smoking behaviour and disease status, treating gender as a block factor.

```
R> cmh <- independence_test(disease ~ smoking | gender,
+     data = alzheimer, teststat = "quad")
R> cmh
```

```
Asymptotic General Independence Test

data:  disease by
 smoking (None, <10, 10-20, >20)
 stratified by gender
chi-squared = 23.32, df = 6, p-value = 0.0006972
```

The default `xtrafo` and `ytrafo` are indicator functions for both categorical variables `disease` and `smoking`.

Smoking and Alzheimer’s disease

The linear statistic is simply the underlying contingency table:

```
R> statistic(cmh, type = "linear")

      Alzheimer's Other dementias Other
None          126              79   104
<10            15               8     5
10-20          30              33    47
>20            27              44    20
```

If performed separately for both genders, it turns out that there is some association for the male but not for the female patients.

Smoking and Alzheimer’s disease

Hence, we use a maximum-type test for the male patients only to gain insights into the pattern of association.

```
R> alzmax <- independence_test(disease ~ smoking,
+     data = alzheimer,
+     subset = alzheimer$gender == "Male",
+     teststat = "max")
R> alzmax

Asymptotic General Independence Test

data:  disease by smoking (None, <10, 10-20, >20)
maxT = 4.95, p-value = 1.030e-05
```

Smoking and Alzheimer's disease

The table of standardized statistics is

```
R> statistic(alzmax, type = "standardized")
```

```
      Alzheimer's Other dementias   Other
None       2.5900          -2.340 -0.1522
<10        2.9713          -2.057 -0.8446
10-20     -0.7765          -1.237  2.1146
>20       -3.6678           4.950 -1.5303
```

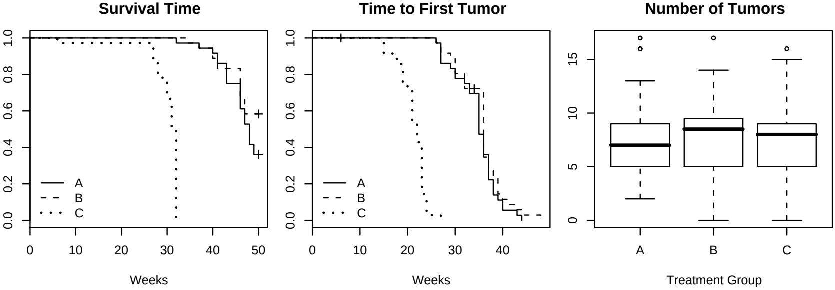
with critical value

```
R> qperm(alzmax, 0.95)
```

```
[1] 2.815
```

Photococarcinogenicity experiments

Molefe *et al.* (2005) study the effect of phototoxic doses of ultraviolet radiation on tumor frequency and latency. At least three responses are of interest: survival time, time to first tumor, and number of tumors. Three different doses are applied: A (600 RBu, with topical vehicle, $n = 36$), B (600 RBu, without topical vehicle, $n = 36$), C (1200 RBu, without topical vehicle, $n = 36$).



Photococarcinogenicity experiments

Global test of all three endpoints using maximum statistic:

```
R> phc <- independence_test(
+    Surv(time, event) + Surv(dmin, tumor) + ntumor ~ group,
+    data = photocar, teststat = "max")
R> phc

Asymptotic General Independence Test

data:  Surv(time, event), Surv(dmin, tumor), ntumor
 by group (A, B, C)
maxT = 7.078, p-value = 6.55e-12
```

Photococarcinogenicity experiments

Again, the source of deviation can be identified by comparing the individual standardized statistics with their 95% critical value:

```
R> statistic(phc, type = "standardized")
```

```
  Surv(time, event) Surv(dmin, tumor)  ntumor
A            -2.327            -2.179  0.2642
B            -4.750            -4.106  0.1510
C             7.078             6.285 -0.4152
```

```
R> qperm(phc, 0.95)
```

```
[1] 2.714
```

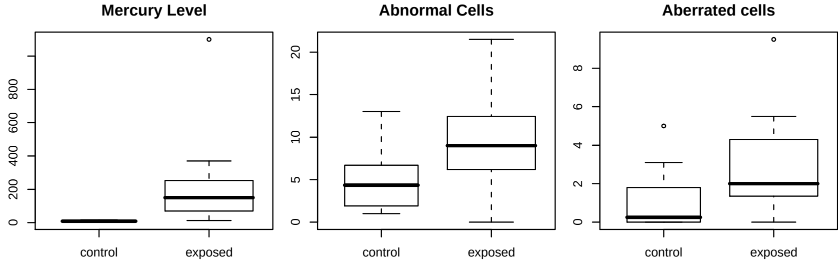
Photococarcinogenicity experiments

Equivalently, we can switch to the *p*-value scale for each statistic:

```
R> phc_pval <- pvalue(phc, method = "single-step")
R> round(phc_pval, digits = 3)

  Surv(time, event) Surv(dmin, tumor) ntumor
A             0.136             0.189  1.000
B             0.000             0.000  1.000
C             0.000             0.000  0.999
```

Contaminated fish consumption

Rosenbaum (1994) studies subjects who ate contaminated fish for more than three years in the exposed group ($n = 23$) and a control group ($n = 16$). Three responses are available: mercury level of the blood, percentage of abnormal cells, percentage of cells with chromosome aberrations.



Contaminated fish consumption

Rosenbaum (1994) proposed to compare the groups using a *coherence criterion*: An observation is said to be smaller than another when all variables are smaller. The rank score is the number of observations smaller minus the number larger.

The resulting univariate score induces a partial ordering, hence the resulting test is called POSET (partially ordered sets) test.

In this situation—univariate response (after transformation) in two samples—the exact conditional distribution of the test statistic can be efficiently obtained using the Streitberg-Röhmel shift algorithm.

Contaminated fish consumption

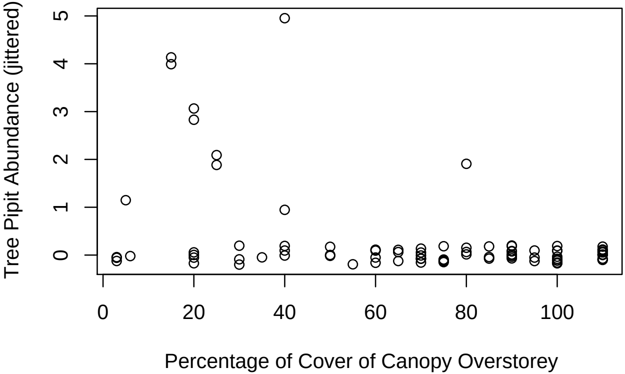
```
R> coherence <- function(data) {
+     x <- t(as.matrix(data))
+     f <- function(y)
+       sum(colSums(x < y) == nrow(x)) -
+       sum(colSums(x > y) == nrow(x))
+     apply(x, 2, f)
+  }
```

```
R> independence_test(mercury + abnormal + ccells ~ group,
+     data = mercuryfish, ytrafo = coherence,
+     distribution = exact())
```

```
Exact General Independence Test

data:  mercury, abnormal, ccells by group (control, exposed)
Z = -4.258, p-value = 4.486e-06
alternative hypothesis: two.sided
```

Tree pipit abundance

Müller and Hothorn (2004) study various habitat factors influencing the abundance of tree pipits in oak forests. The cover of canopy overstorey is of particular interest.



Tree pipit abundance

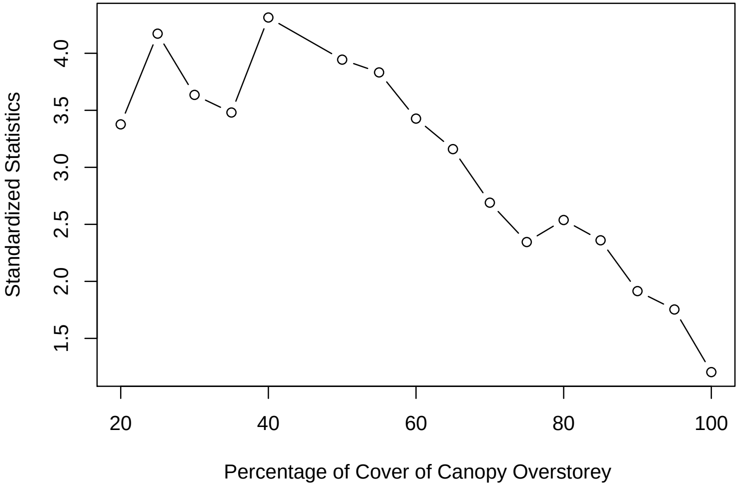
This suggests that there is a step-shaped relationship between the mean number of tree pipits and the cover of canopy overstorey (rather than a linear association), i.e., a cutpoint.

If the cutpoint c was known, its significance could be assessed in the conditional inference framework by using the indicator function $g_c(X) = I(X \leq c)$.

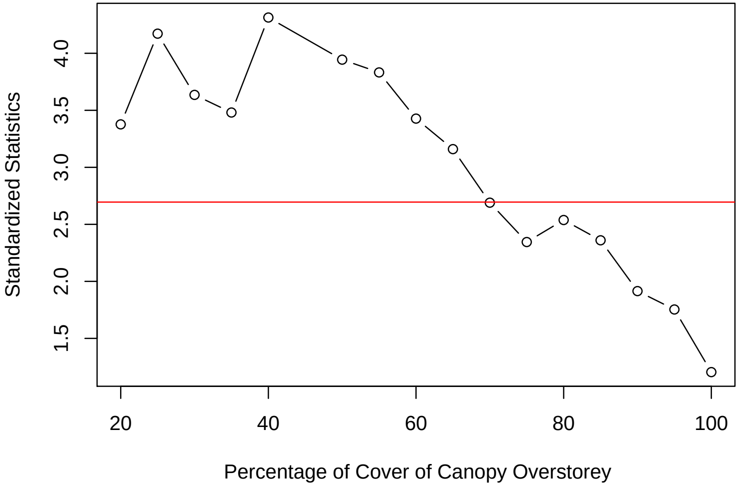
A straightforward idea to assess all conceivable cutpoints $c_1, \dots, c_\ell$ is to use *maximally selected statistics*, i.e., compute all 2-sample test statistics and reject if the maximum is too large.

This is again a special case of the conditional inference framework when using a maximum statistic and the multivariate transformation $g(X) = (g_{c_1}(X), \dots, g_{c_\ell}(X))^\top$.

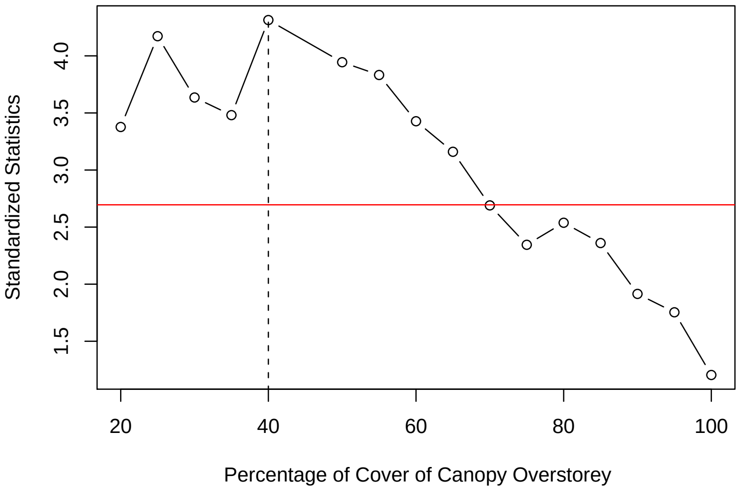
Tree pipit abundance



Tree pipit abundance



Tree pipit abundance



Tree pipit abundance

Thus, maximally selected statistics can be used to assess *if* and *where* a cutpoint exists.

```
R> tp <- maxstat_test(counts ~ coverstorey,
+     data = treepipit)
R> tp

Asymptotic Maxstat Test

data:  counts by coverstorey
maxT = 4.314, p-value = 0.0001545
sample estimates:
$cutpoint
[1] 40
```

High- and low-risk groups of rectal cancer patients

Sauer *et al.* (2004) study the association of survival times of $n = 349$ rectal cancer patients and their TNM classification (ordinal assessments of tumors, lymph nodes, metastases).

Current practice in TNM classification is to distinguish stage I vs. II cancer by the T category, II vs. III by N (N $\leq$ N0), III vs. IV by M.

Instead of using these fixed interactions, consider all ordered interactions in a generalized maximally selected statistic. Only T and N can be used because all patients belong to M category M0.

- influence function *h*: logrank scores for censored response,
- transformation *g*: all binary partitions in the two ordered covariates (T and N category) that are ordered in T given N and vice versa.

High- and low-risk groups of rectal cancer patients

```
R> independence_test(Surv(time, event) ~ tn,
+     data = preOP, xtrafo = ordered_splits,
+     distribution = approximate(B = 9999))
```

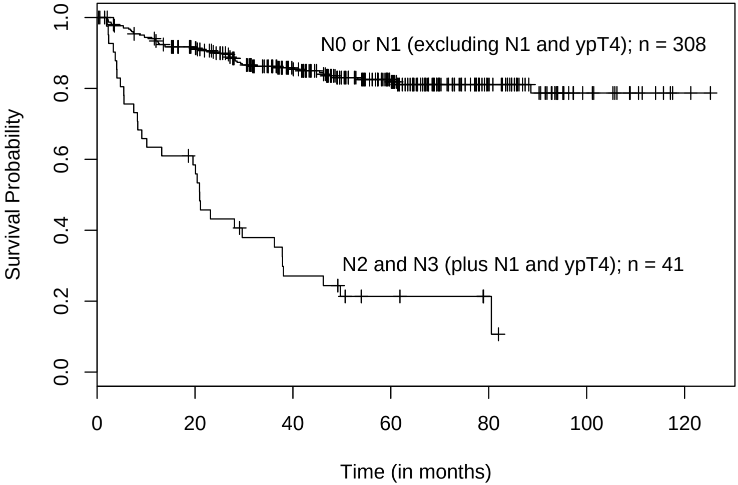
```
Approximative General Independence Test

data:  Surv(time, event) by tn
maxT = 8.69, p-value < 2.2e-16
```

The binary partition leading to the maximal standardized statistic is essentially a cutpoint in the N category

- low risk: N0 or N1 (excluding N1 and ypT4),
- high risk: N2 and N3 (plus N1 and ypT4).

However, just one patient is in group “N1 and ypT4”.

High- and low-risk groups of rectal cancer patients



Software

Package **coin** provides `independence_test()` as the workhorse function, based on C routines for computing the linear statistic, its expectation and covariance. Only a single implementation of the Monte Carlo and asymptotic null distribution is used.

Convenience interfaces facilitate application of classical tests (previously available in R) in a flexible conditional-inference framework.

Most analyses discussed above can be reproduced via

```
R> vignette("LegoCondInf", package = "coin")
```

The package is available from the Comprehensive R Archive Network at

```
http://CRAN.R-project.org/package=coin
```

Special cases

The following classical tests are special cases of the framework implemented in **coin**:

2- und *k*-sample permutation test, Wilcoxon-Mann-Whitney rank sum test, van Elteren test, van der Waerden test, Median test, Kruskal-Wallis test, Ansari-Bradley test, Fligner-Killeen test, Pearson's χ^2 test, generalized Cochran-Mantel-Haenszel test, linear-by-linear association test, logrank test, maximally selected statistics, Spearman test, Friedman test, Wilcoxon signed rank test, Page test, McNemar test, Cochran's *Q*, Quade test, Anderson test, Wilcoxon-Nemenyi-McDonald-Thompson test, Nemenyi-Damico-Wolfe-Dunn test, Rosenbaum's POSET test, ...

References

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250.
Preprint available at `http://epub.wu-wien.ac.at/`

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**, 257–263.
Preprint available at `http://epub.wu-wien.ac.at/`

Hothorn T, Zeileis A (2008). "Generalized Maximally Selected Statistics." *Biometrics*, forthcoming.
Preprint available at `http://epub.wu-wien.ac.at/`