A Lego-System 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]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 and Computer Science
Eindhoven University of Technology
HG 9.25, P.O. Box 513
5600 MB Eindhoven, The Netherlands
markvdw@win.tue.nl

Abstract

Conditioning on the observed data is an important and flexible design principle for statistical test procedures. Although generally applicable, most text book and software implementations are limited to the treatment of special cases. 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 a software implementation that enables the data analyst to compute on the theoretical concepts as closely as possible. We re-analyze data where non-standard inference procedures are required utilizing the *coin* add-on package in the R system for statistical computing and show what one can gain from going beyond pre-packaged test procedures.

KEY WORDS: Permutation tests; Multiple testing; Independence; Software.

Version:

$Id: LegoCondInf.Rnw,v 1.15 2005/11/29 10:09:48 hothorn Exp $

1 Introduction

The distribution of a test statistic under the circumstances of a certain 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 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 in the 1930s (Fisher, 1935).

The pros and cons of both approaches have been discussed in extensio (e.g. by Ludbrook and Dudley, 1998; Berger, 2000; Shuster, 2005) and we refrain from stepping into this mostly philosophical debate, noting that most conditional test procedures are asymptotically equivalent to their unconditional counterparts anyway.

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 (Edgington, 1987; Good, 2000; Pesarin, 2001; Ernst, 2004). Such a classification into inference problems (categorical data analysis, multivariate analysis, K-sample location problems, correlation etc.) each being associated with a 'natural' form of the test statistic obstructs our view on the common foundations of all permutation tests. Theoretical advances in the last decade (Strasser and Weber, 1999; Janssen and Pauls, 2003) helped us to understand the strong connections between the 'classical' permutation tests and 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 Strasser and Weber (1999). This unifying theory is based on a flexible form of multivariate linear statistics for the general independence problem those conditional expectation and covariance is trackable. 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.

It is one mission, if not *the* mission, of statistical computing to transform new theoretical developments into flexible software tools for the data analyst. Currently, the statisticians toolbox consists of rather inflexible spanners, such as `wilcox.test` for the Wilcoxon-Mann-Whitney test or `mantelhaen.test` for the Cochran-Mantel-Haenszel chi-squared test in S languages. The implementation of permutation tests with user interfaces designed to deal with special cases (see the Tables in Oster, 2002, 2003, for an overview on procedures implemented in StatXact, LogXact, Strata, SAS and Testimate) leads to the classical 'cook book' statistics. Such cook books, and thus software implementations, typically teach recipes and hide the concepts which are necessary to go beyond the implemented

procedures when the data analyst is faced with non-standard inference problems or wants to perform a test not supported by the preferred software package.

With this work, we add an adjustable spanner to the statisticians toolbox which helps to address both the common as well as new or unusual inference problems with the appropriate conditional test procedures. The *coin* add-on package to the R system for statistical computing (R Development Core Team, 2005) essentially is a software instance of the Strasser-Weber framework for the generalized independence problem which allows for computations directly on the theory those main concepts are sketched in Section 2. In the main part of this paper we show how one can build permutation tests 'on the fly' by plugging together Lego bricks for the multivariate linear statistic, the test statistic and the conditional null distribution.

2 Conditional Inference

To fix ideas we assume that we are provided with 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. Strasser and Weber (1999) suggest to derive scalar test statistics for testing H_0 from multivariate linear statistics of the form

$$\mathbf{T} = \text{vec}\left(\sum_{i=1}^{n} g(\mathbf{X}_i) h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))^\top\right) \in \mathbb{R}^{pq}. \tag{1}$$

Here, $g : \mathcal{X} \to \mathbb{R}^p$ is a transformation of the $\mathbf{X}$ measurements and the *influence function* $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ depends on the responses $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ in a permutation symmetric way. We will give specific examples how to choose g and h for specific inference problems later on.

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 $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$.

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 Strasser and Weber (1999):

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

$$- \quad \frac{1}{n-1}\mathbb{V}(h|S) \otimes \left(\sum_i g(\mathbf{X}_i)\right) \otimes \left(\sum_i g(\mathbf{X}_i)\right)^\top$$

where $\otimes$ denote the Kronecker product. The conditional expectation of the influence function is

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

with corresponding $q \times q$ covariance matrix $\mathbb{V}(h|S)$ given by

$$n^{-1} \sum_i \left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) - \mathbb{E}(h|S)\right) \left(h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) - \mathbb{E}(h|S)\right)^\top .$$

The key step for the construction of test statistics from the multivariate linear statistic $\mathbf{T}$ is its standardization utilizing the the conditional expectation μ and covariance matrix Σ. 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

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

utilizing the conditional expectation μ and covariance matrix Σ. A prominent alternative are quadratic forms $c_{\mathrm{quad}}(\mathbf{t}, \mu, \Sigma) = (\mathbf{t} - \mu)\Sigma^+(\mathbf{t} - \mu)^\top$ 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 less than 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 trackable for small to moderate sample sizes.

The conditional distribution can be approximated by the limit distribution under all circumstances. Strasser and Weber (1999) 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 $n \to \infty$. 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 itself being approximated by means of quasi-randomized Monte-Carlo procedures in the multivariate setting (Genz, 1992). For quadratic forms c_{quad} which follow a χ^2 distribution with degrees of freedom given by the rank of Σ (e.g. Theorem 6.20, Rasch, 1995), exact probabilities can be computed efficiently.

Conditional Monte-Carlo procedures can also be used to approximate the exact distribution up to any desired accuracy by evaluating the test statistic for a random sample from the set all permutations S. It is important to note that the presence of a grouping of the observations into blocks, only permutations within blocks are eligible and that the conditional expectation and covariance matrix need to be computed separately for each block.

3 Playing Lego with *coin*

The *coin* package implements software infrastructure for the main components of the theoretical framework sketched above, namely the linear statistic $\mathbf{T}$ (1) with user-defined transformations g and influence functions h, functions for the computation of the conditional expectation μ and covariance matrix Σ as in (2) and utilities for the computation of the conditional distribution of $c_{\max}$ or c_{quad} test statistics. Thus, the flexibility of the theoretical framework is translated and preserved in a software instance which enables the data analyst to benefit from this conceptually simple methodology in every day's data analysis.

In the following, we will adress some inference problems which require functionality not available in standard software packages. The data are included in the *coin* package and our analyses can be reproduced from the package vignette.

Smoking and Alzheimer's Disease. Salib and Hillier (1997) report results of a case-control study on Alzheimer's disease and smoking behavior of 198 patients suffering from Alzheimer's disease and 164 controls. The data shown in Table 1 have been re-constructed from Table 4 in Salib and Hillier (1997). The authors conclude that 'cigarette smoking is less frequent in men with Alzheimer's disease'.

[Table 1 about here.]

Ignoring the ordinal structure of the smoking behavior, the null hypothesis of independence between smoking and disease status treating gender as a block factor with a c_{quad}-type test statistic, i.e. the conditional version of the Cochran-Mantel-Haenszel test

```
R> data("alzheimer", package = "coin")
R> it_alz <- independence_test(alzheimer, teststat = "quadtype")
R> it_alz

        Asymptotic General Independence Test

data:  disease by
         groups None, <10, 10-20, >20
         stratified by gender
T = 23.3163, df = 6, p-value = 0.0006972
```

suggests that there is a clear deviation from independence. By default, the influence function h and the transformation g are dummy codings of the disease status $\mathbf{Y}$ and the smoking behavior $\mathbf{X}$, i.e. $h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) = (1, 0, 0)$ and $g(\mathbf{X}_i) = (1, 0, 0, 0)$ for a non-smoking Alzheimer patient. Consequently, the linear multivariate statistic $\mathbf{T}$ based on g and h is the (vectorized) contingency table of both variables

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

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

with conditional expectation `expectation(it_alz)` and conditional covariance `covariance(it_alz)` which are available for standardizing the contingency table $\mathbf{T}$. The conditional distribution is approximated by its limiting χ^2 distribution by default. This leads to exactly the same p-value as the unconditional test

```
R> pvalue(it_alz)

[1] 0.0006971815

R> mantelhaen.test(alzheimer)$p.value

[1] 0.0006971815
```

The form of the deviation from independence is of special interest, however, a chi-squared statistic is not particular useful for this purpose. Instead, we define the test statistic as the maximum of the standardized contingency table via

```
R> it_alzmax <- independence_test(alzheimer, teststat = "maxtype")
R> it_alzmax

        Asymptotic General Independence Test

data:  disease by
         groups None, <10, 10-20, >20
         stratified by gender
T = 3.5106, p-value = 0.005076
```

which leads to virtually the same p-value. The standardized contingency table sheds some light on the deviations from independence

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

      Alzheimer's Other dementias Other diagnoses
None     1.840717      -2.1858441       0.2464727
<10      1.911821      -0.2588064      -1.7091807
10-20   -2.128934      -0.5147297       2.6917719
>20     -1.248560       3.5106083      -2.2187050
```

and leads to the impression that patients suffering from Alzheimer's disease smoked less cigarettes than expected under independence and, to a much larger degree, patients with other dementias smoked much more than expected. However, interpreting the standardized contingency table either requires knowledge about the distribution of the standardized statistics, i.e. via an approximation of the 97.5% quantile of the conditional null distribution (two-sided test) which is available from

```
R> qperm(it_alzmax, 0.975)

[1] 3.039035
```

Alternatively and more conveniently, we can to switch to the p-value scale. Here, we choose step-down adjusted resampling-based p-values (Westfall and Young, 1993). First, we approximate the conditional distribution by 50.000 Monte-Carlo replications

```
R> it_alzMC <- independence_test(alzheimer, distribution = approximate(B = 50000))
```

with global p-value

```
R> pvalue(it_alzMC)

[1] 0.0057
99 percent confidence interval:
 0.004869862 0.006625664
```

(the normal approximation is rather accurate) and apply Algorithm 2.8 in Westfall and Young (1993) to obtain p-values adjusted for multiple comparisons

```
R> pvalue(it_alzMC, method = "step-down")

      Alzheimer's Other dementias Other diagnoses
None      0.31918         0.18538         0.84776
<10       0.30764         0.96212         0.34756
10-20     0.20134         0.95458         0.07074
>20       0.59570         0.00570         0.19754
```

The above results support the conclusion that the rejection of the null hypothesis of independence is due to a large number of heavy smokers with other dementias but seems rather unrelated to Alzheimer's disease itself.

Of course, ignoring the ordinal structure of one of the variables is only sub-optimal. Ordinal variables can be incorporated into the general framework via linear-by-linear association tests (Agresti, 2002). When $\mathbf{Y}$ is measured at J levels and $\mathbf{X}$ at K levels, $\mathbf{Y}$ and $\mathbf{X}$ are associated with score vectors $\xi \in \mathbb{R}^J$ and $\gamma \in \mathbb{R}^K$, respectively. The linear statistic is now a linear combination of the linear statistic $\mathbf{T}$ of the form

$$\mathbf{MT} = \text{vec}\left(\sum_{i=1}^{n} \gamma^\top g(\mathbf{X}_i) \left(\xi^\top h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)\right)^\top\right) \in \mathbb{R} \text{ with } \mathbf{M} = \xi \otimes \gamma.$$

For smoking, a natural choice of the scores are the midpoints of the internals used to discretize the number of cigarettes per day and we can setup a linear-by-linear association test with $c_{\max}$ type test statistic via

```
R> it_alzL <- independence_test(alzheimer, scores = list(smoking = c(0,
+     5, 15, 25)))
R> pvalue(it_alzL)
```

```
[1] 0.005686773
99 percent confidence interval:
 0.005374770 0.005998777
```

and the single-step adjusted p-values

```
R> pvalue(it_alzL, method = "single-step")
```

```
 Alzheimer's Other dementias Other diagnoses
  0.05112845     0.005616577       0.7946195
```

for the standardized linear statistic

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

```
 Alzheimer's Other dementias Other diagnoses
   -2.334202        3.087989      -0.6459272
```

support the conclusion that smoking is associated with other dementia and, therefore, smoking is less frequent in patients suffering from Alzheimer's disease.

Photococarcinogenicity Experiments. The effect on tumor frequency and latency in photococarcinogenicity experiments, where carcinogenic doses of ultraviolet radiation (UVR) are administered, are measured by means of (at least) three response variables: the survival time, the time to first tumor and the total number of tumors of animals in different treatment groups. The main interest is testing the global null of no treatment effect with respect to survival time, time to first tumor or number of tumors (Molefe et al., 2005, analyze the detection time of tumors in addition, this data is not given here). In case the global null hypothesis can be rejected, the deviations from the partial hypotheses are of special interest.

Molefe et al. (2005) report data of an experiment where 108 animals were exposed to different levels of UVR exposure (group A: topical vehicle and 600 Robertson–Berger units of UVR, group B: no topical vehicle and 600 Robertson–Berger units of UVR and group C: no topical vehicle and 1200 Robertson–Berger units of UVR). The data are taken from Tables 1 to 3 in Molefe et al. (2005), where a parametric test procedure is proposed. Figure 1 depicts the group effects for all three response variables.

[Figure 1 about here.]

First, we construct a global test for the null hypothesis of independence of treatment and *all* three response variables. A $c_{\max}$-type test based on the standardized multivariate linear statistic and an approximation of the conditional distribution utilizing the asymptotic distribution simply reads

```
R> data("photocar", package = "coin")
R> it_ph <- independence_test(Surv(time, event) + Surv(dmin,
+     tumor) + ntumor ~ group, data = photocar)
R> it_ph
```

```
        Asymptotic General Independence Test

data:  Surv(time, event), Surv(dmin, tumor), ntumor by groups A, B, C
T = 7.0777, p-value = 6.963e-12
```

Here, the influence function h consists of the logrank scores of the survival time and time to first tumor as well as the number of tumors, i.e. for the first animal in the first group $h(\mathbf{Y}_1, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)) = (-1.08, -0.56, 5)$ and $g(\mathbf{X}_1) = (1, 0, 0)$. The multivariate statistic is the sum of each of the three elements of the influence function h in each of the groups, i.e.

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

  Surv(time, event) Surv(dmin, tumor) ntumor
A         -8.894531         -9.525269    276
B        -18.154654        -17.951560    274
C         27.049185         27.476828    264
```

It is important to note that this global test utilizes the complete correlation structure

```
R> cov2cor(covariance(it_ph))

                    A:Surv(time, event) B:Surv(time, event)
A:Surv(time, event)                1.00               -0.50
B:Surv(time, event)               -0.50                1.00
C:Surv(time, event)               -0.50               -0.50
A:Surv(dmin, tumor)                0.66               -0.33
B:Surv(dmin, tumor)               -0.33                0.66
C:Surv(dmin, tumor)               -0.33               -0.33
A:ntumor                          -0.05                0.03
B:ntumor                           0.03               -0.05
C:ntumor                           0.03                0.03
                    C:Surv(time, event) A:Surv(dmin, tumor)
A:Surv(time, event)               -0.50                0.66
B:Surv(time, event)               -0.50               -0.33
C:Surv(time, event)                1.00               -0.33
A:Surv(dmin, tumor)               -0.33                1.00
B:Surv(dmin, tumor)               -0.33               -0.50
C:Surv(dmin, tumor)                0.66               -0.50
A:ntumor                           0.03                0.25
B:ntumor                           0.03               -0.13
C:ntumor                          -0.05               -0.13
                    B:Surv(dmin, tumor) C:Surv(dmin, tumor)
A:Surv(time, event)               -0.33               -0.33
B:Surv(time, event)                0.66               -0.33
C:Surv(time, event)               -0.33                0.66
```

```
A:Surv(dmin, tumor)               -0.50              -0.50
B:Surv(dmin, tumor)                1.00              -0.50
C:Surv(dmin, tumor)               -0.50               1.00
A:ntumor                          -0.13              -0.13
B:ntumor                           0.25              -0.13
C:ntumor                          -0.13               0.25
                    A:ntumor B:ntumor C:ntumor
A:Surv(time, event)    -0.05     0.03     0.03
B:Surv(time, event)     0.03    -0.05     0.03
C:Surv(time, event)     0.03     0.03    -0.05
A:Surv(dmin, tumor)     0.25    -0.13    -0.13
B:Surv(dmin, tumor)    -0.13     0.25    -0.13
C:Surv(dmin, tumor)    -0.13    -0.13     0.25
A:ntumor                1.00    -0.50    -0.50
B:ntumor               -0.50     1.00    -0.50
C:ntumor               -0.50    -0.50     1.00
```

when p-values are computed via quasi-randomized Monte-Carlo procedures in the multivariate setting (Genz, 1992). Alternatively, a test statistic based on the quadratic form c_{quad} directly incorporates the covariance matrix and leads to a very similar p-value.

The deviations from the partial null hypotheses, i.e. independence of each single response and treatment groups, can be inspected by the standardized linear statistic $\mathbf{T}$

```
  Surv(time, event) Surv(dmin, tumor)     ntumor
A         -2.327338         -2.178704  0.2642120
B         -4.750336         -4.106039  0.1509783
C          7.077674          6.284743 -0.4151904
```

or by means of adjusted p-values

```
R> pvalue(it_ph, method = "single-step")
```

```
  Surv(time, event) Surv(dmin, tumor)  ntumor
A           0.13585           0.18977 0.99989
B           0.00002           0.00034 1.00000
C           0.00000           0.00000 0.99859
```

Of course, the goodness of the asymptotic procedure can be checked against the Monte-Carlo approximation which is computed by

```
R> it <- independence_test(Surv(time, event) + Surv(dmin,
+     tumor) + ntumor ~ group, data = photocar, distribution = approximate(50000))
R> pvalue(it, method = "single-step")
```

```
  Surv(time, event) Surv(dmin, tumor)  ntumor
A           0.13304            0.1861 0.99994
B           0.00000            0.0003 1.00000
C           0.00000            0.0000 0.99878
```

The more powerful step-down adjusted p-values are

```
R> pvalue(it, method = "step-down")
```

```
  Surv(time, event) Surv(dmin, tumor)  ntumor
A           0.08322           0.09918 0.95434
B           0.00000           0.00018 0.88940
C           0.00000           0.00000 0.91862
```

Clearly, the rejection of the global null hypothesis is due to the group differences in both survival time and time to first tumor whereas no treatment effect on the total number of tumors can be observed.

Contaminated Fish Consumption. In the former two applications, standard transformations for g and h such as dummy codings and logrank scores have been applied. In the third application, we will show how one can utilize the *coin* functionality to implement a newly invented test procedure.

Rosenbaum (1994) proposed to compare groups by means of a *coherence criterion* and studied a dataset of subjects who ate contaminated fish for more than three years in the 'exposed' group and a control group. Three response variables are available: the mercury level of the blood, the percentage of cells with structural abnormalities and the proportion of cells with asymmetrical or incomplete-symmetrical chromosome aberrations (see Figure 2). The observations are partially ordered: an observation is said to be smaller than another when all three variables are smaller. The rank score for observation i is the number of observations that are larger (following the above criterion) than observation i minus the number of observations that are smaller. The distribution of the rank scores in both groups is to be compared and the corresponding test is called 'POSET-test' (partially ordered sets).

[Figure 2 about here.]

The coherence criterion can be formulated in a simple function

```
R> coherence <- function(data) {
+     x <- t(as.matrix(data))
+     matrix(apply(x, 2, function(y) sum(colSums(x <
+         y) == nrow(x)) - sum(colSums(x > y) == nrow(x))),
+         ncol = 1)
+ }
```

which is now defined as influence function h via the `ytrafo` argument

```
R> data("mercuryfish", package = "coin")
R> poset <- independence_test(mercury + abnormal + ccells ~
+     group, data = mercuryfish, ytrafo = coherence,
+     distribution = exact())
```

Once the transformations g (a zero-one coding of the exposed and control group) and h (the coherence criterion) are defined, we enjoy the whole functionality of the framework, including an exact two-sided p-value

```
R> pvalue(poset)

[1] 4.486087e-06
```

and density (`dperm`), distribution (`pperm`) and quantile functions (`qperm`) of the conditional distribution. When only a small number of observations is available, it might be interesting to compare the exact conditional distribution and its approximation via the limiting distribution. For the `mercuryfish` data, the relevant parts of both distribution functions are shown in Figure 3. It turns out the the normal approximation would be sufficient for all practical purposes in this application.

[Figure 3 about here.]

4 Conclusion

Conditioning on the observed data is a simple, yet powerful, design principle for statistical tests. Conceptually, one only needs to choose an appropriate test statistic and evaluate it for all admissible permutations of the data (Ernst, 2004, gives an example with Hotelling's T^2). In practical setups, an implementation of this procedure requires a certain amount of programming and computing time. Often, permutation tests are regarded as being 'computationally impractical' for larger sample sizes (Balkin and Mallows, 2001). Therefore, popular software packages offer implementations for the most prominent conditional tests, such as the permutation t test, where fast algorithms for the computation of conditional p-values are available and the limiting distribution is known.

The permutation test framework by Strasser and Weber (1999) makes at least two important contributions: analytic formulae for the conditional expectation and covariance and the limiting normal distribution of a class of multivariate linear statistics. Thus, test statistics can be defined for appropriately standardized linear statistics and a fast approximation of the conditional distribution is available, especially for large sample sizes.

The *coin* package is an attempt to translate the theoretical concepts of Strasser and Weber (1999) into software as closely as possible preserving the simplicity and flexibility of conditional inference. Basically, the package implements *one* function for computing the linear statistic $\mathbf{T}$, *one* function for the conditional expectation μ and covariance Σ and plug-ins for several test statistics c. Moreover, normal, χ^2 or Monte-Carlo approximations of the conditional distribution only need to be implemented *once*.

But who stands to benefit from such a software infrastructure? We argue that better data analysis is possible in cases when the appropriate conditional test is not available from standard software packages. Statisticians can modify

existing test procedures or even try new ideas by computing directly on the theory. A high-level Lego-system is attractive for software developers, because only the transformation g and influence function h need to be newly implemented, but the burden of implementing a Monte-Carlo procedure, or even thinking about asymptotics, is waived. Since the *coin* package consists of only a few core functions that need to be tested, the setup of quality assurance tools is rather simple in this case (Bergmann et al., 2000, the need for such tests is obvious,[). Many text books (e.g. Hollander and Wolfe, 1999) or software manuals (first of all the excellent StatXact handbook by Mehta and Patel, 2003) include examples and results of the associated test procedures which have been reproduced with *coin*.

Since the *coin* package is part of the Comprehensive R Archive Network (CRAN, http://CRAN.R-project.org) we have been able to help several people asking 'Is the xyz-test available in R' on the `r-help` email list with the answer 'No, but its only those two lines of R code in *coin*'. With the *coin* functionality being available we are no longer limited to already implemented test procedures nor are forced to self-implementation. Instead, the appropriate conditional test procedure for the problem at hand is only a matter of choosing appropriate transformation and influence functions.

References

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken, New Jersey: John Wiley & Sons, 2nd ed.

Balkin, S. D. and Mallows, C. L. (2001), "An Adjusted, Asymmetric Two-Sample t Test," *The American Statistician*, 55, 203–206.

Berger, V. W. (2000), "Pros and cons of permutation tests in clinical trials," *Statistics in Medicine*, 19, 1319–1328.

Bergmann, R., Ludbrook, J., and Spooren, W. P. (2000), "Different Outcomes of the Wilcoxon-Mann-Whitney Test From Different Statistics Packages," *The American Statistician*, 54, 72–77.

Edgington, E. S. (1987), *Randomization Tests*, New York, USA: Marcel Dekker.

Ernst, M. D. (2004), "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19, 676–685.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh, UK: Oliver and Boyd.

Genz, A. (1992), "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, 1, 141–149.

Good, P. I. (2000), *Permutation Tests: A Practical Guide to Resampling Methods for Testing*, New York, USA: Springer-Verlag.

Hollander, M. and Wolfe, D. A. (1999), *Nonparametric statistical inference*, New York: John Wiley & Sons, 2nd ed.

Janssen, A. and Pauls, T. (2003), "How Do Bootstrap and Permutation Tests Work?" *The Annals of Statistics*, 31, 768–806.

Ludbrook, J. and Dudley, H. (1998), "Why Permutation Tests are Superior to t and F Tests in Biomedical Research," *The American Statistician*, 52, 127–132.

Mehta, C. R. and Patel, N. R. (2003), *StatXact-6: Statistical Software for Exact Nonparametric Inference*, Cytel Software Cooperation, Cambridge, USA.

Molefe, D. F., Chen, J. J., Howard, P. C., Miller, B. J., Sambuco, C. P., Forbes, P. D., and Kodell, R. L. (2005), "Tests for effects on tumor frequency and latency in multiple dosing photococarcinogenicity experiments," *Journal of Statistical Planning and Inference*, 129, 39–58.

Oster, R. A. (2002), "An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods," *The American Statistician*, 56, 235–246.

— (2003), "An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods–Part II," *The American Statistician*, 57, 201–213.

Pesarin, F. (2001), *Multivariate Permutation Tests: With Applications to Biostatistics*, Chichester, UK: John Wiley & Sons.

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.

Rasch, D. (1995), *Mathematische Statistik*, Heidelberg, Leipzig: Johann Ambrosius Barth Verlag.

Rosenbaum, P. R. (1994), "Coherence in Observational Studies," *Biometrics*, 50, 368–374.

Salib, E. and Hillier, V. (1997), "A case-control study of smoking and Alzheimer's disease," *International Journal of Geriatric Psychiatry*, 12, 295–300.

Shuster, J. J. (2005), "Diagnostics for Assumptions in Moderate to Large Simple Clinical Trials: Do They Really Help?" *Statistics in Medicine*, 24, 2431–2438.

Strasser, H. and Weber, C. (1999), "On the asymptotic theory of permutation statistics," *Mathematical Methods of Statistics*, 8, 220–250.

Westfall, P. H. and Young, S. S. (1993), *Resampling-based Multiple Testing*, New York: John Wiley & Sons.

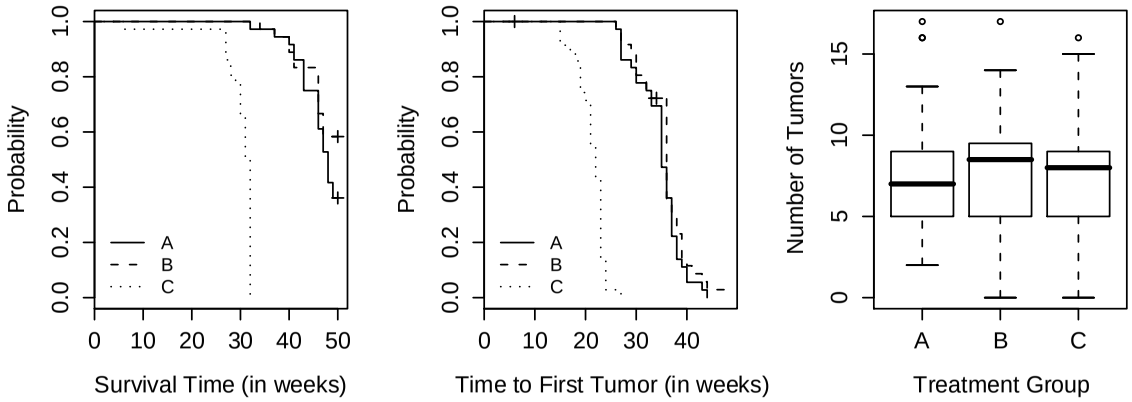

Figure 1: `photocar` data: Kaplan-Meier estimates of time to death and time to first tumor as well as boxplots of the total number of tumors in three treatment groups.

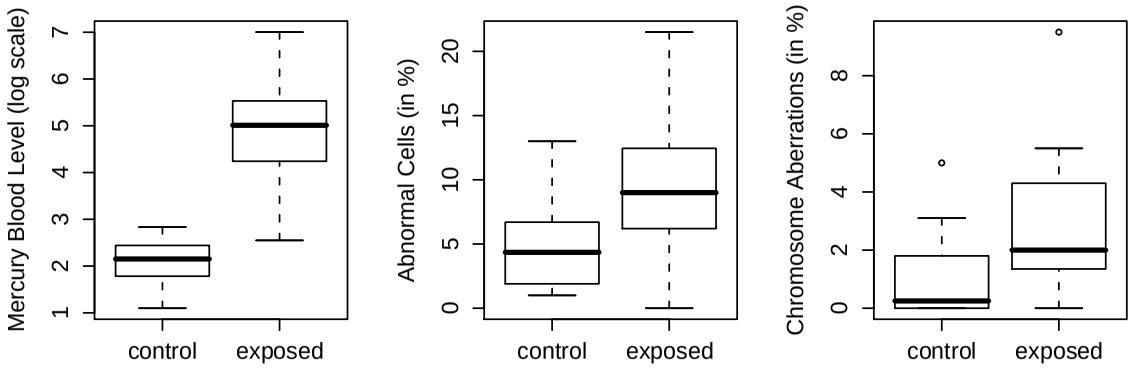

Figure 2: `mercuryfish` data: Distribution of all three response variables in the exposed group and control group.

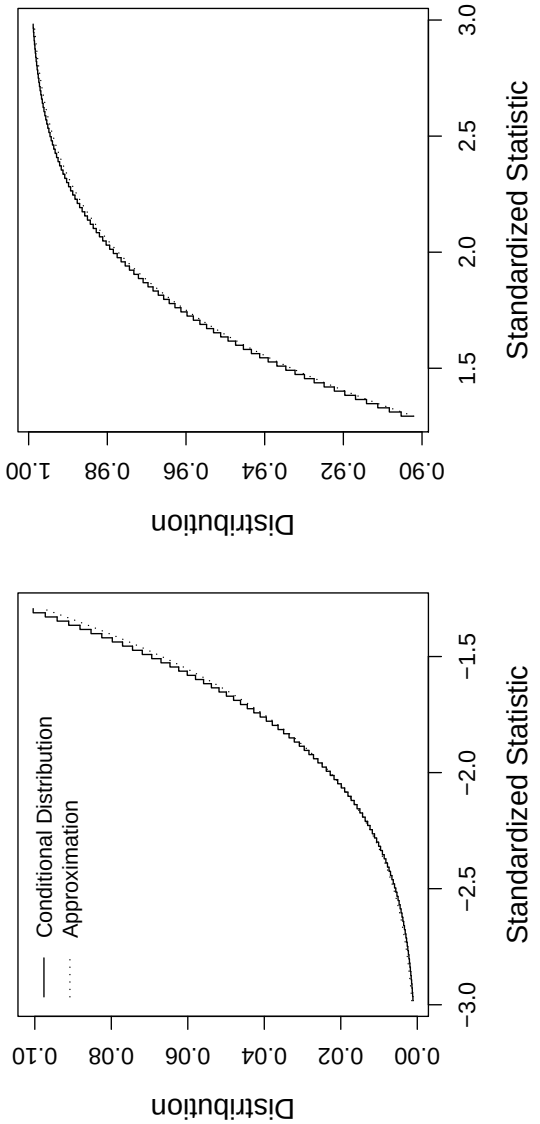

Figure 3: `mercuryfish` data: Conditional distribution and asymptotic normal approximation for the POSET test.

Table 1: `alzheimer` data: Smoking and Alzheimer's disease.

	No. of cigarettes daily			
	None	<10	10–20	>20
Female				
Alzheimer's	91	7	15	21
Other dementias	55	7	16	9
Other diagnoses	80	3	25	9
Male				
Alzheimer's	35	8	15	6
Other dementias	24	1	17	35
Other diagnoses	24	2	22	11