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-N"urnberg
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
extttAchim.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, 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 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 the 'classical' test procedures.

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

$Date: 2005/12/19 15:34:04 $ $Revision: 1.26 $

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 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 (Fisher, 1935). The pros and cons of both approaches in different fields of application have been widely discussed (e.g. by Ludbrook and Dudley, 1998; Berger, 2000; Shuster, 2005). 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 (Edgington, 1987; Good, 2000; Pesarin, 2001; Ernst, 2004). 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 Strasser and Weber, 1999; Janssen and Pauls, 2003) 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 Strasser and Weber (1999). 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 compute or approximate the conditional null distribution. The classical procedures, such as a permutation t test, are part of this framework and, even more interestingly, new test procedures can be embedded into the same theory whose main ideas are sketched in Section 2.

Currently, the statistician's toolbox consists of rather inflexible 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 (Hothorn et al., 2005)in the R system for statistical computing (R Development Core Team, 2005).

2 A Conceptual Lego System

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

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

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 $h(\mathbf{Y}_i) = h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))$ must depend 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 in Section 3.

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 $\mu \in \mathbb{R}^{pq \times 1}$ 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 Strasser and Weber (1999):

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

where $\otimes$ denote 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 an observed linear statistic $\mathbf{t} \in \mathbb{R}^{pq \times 1}$ into the real line can be of arbitrary form. An obvious choice is the maximum of the absolute values of the standardized linear statistic

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

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 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 trackable for small and moderate sample sizes. Conditional Monte-Carlo procedures ('resampling') 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 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.

Less well known is the fact that 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 ($pq = 1$). The evaluation of multivariate normal distributions is possible by means of quasi-randomized Monte-Carlo procedures (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.

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 interesting from our point of view, how new conditional test procedures can be constructed conceptually *and* practically. Therefore, each inference problem goes along with R code necessary to perform the appropriate conditional test using the *coin* functionality which enables the data analyst to benefit from this simple methodology in every day's data analysis. All analyses are reproducible from the *coin* package vignette available from `http://CRAN.R-project.org`.

Genetic Components of Alcoholism. Various studies have linked alcohol dependence phenotypes to chromosome 4. One candidate gene is *NACP* (non-amyloid component of plaques), coding for alpha synuclein. Bönsch et al. (2005) found longer alleles of *NACP*-REP1 in alcohol-dependent patients compared with healthy controls and report that the allele lengths show some association with levels of expressed alpha synuclein mRNA (see Figure 1).

[Figure 1 about here.]

Our first attempt to test for different levels of gene expression in the three groups is the classical Kruskal-Wallis test. Here, the transformation g is a dummy coding of the allele length ($g(\mathbf{X}_i) = (0, 1, 0)^\top$ for intermediate length, for example) and the value of the influence function $h(\mathbf{Y}_i)$ is the rank of $\mathbf{Y}_i$ among the ranks of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Thus, the observed linear statistic $\mathbf{t}$ is the vector of rank sums in each of the three groups and the test statistic is a quadratic form $(\mathbf{t} - \mu)\Sigma^+(\mathbf{t} - \mu)^\top$ utilzing the conditional expectation μ and covariance matrix Σ.

In order to compute the linear statistic we need to define an influence function performing a ranking of the expression levels. Under the null hypothesis, the c_{quad}-type Kruskal-Wallis test statistic tends to a χ^2 distribution with two degrees of freedom (the rank of the conditional covariance matrix Σ) from which a p-value can be computed. In R, the function `independence_test` takes a formula describing the inference problem, i.e., the independence of expression levels (`elevel`) and allele lengths (`length`), the influence function is specified via the `ytrafo` argument and we ask for a c_{quad}-type test statistic (`teststat`) as follows:

```
R> independence_test(elevel ~ length, data = alpha,
+     ytrafo = function(data) trafo(data, numeric_trafo = rank),
+     teststat = "quadtype")

        Asymptotic General Independence Test

data:  elevel by groups short, intermediate, long
T = 8.8302, df = 2, p-value = 0.01209
```

The results are equivalent to the results reported by `kruskal.test`, the 'classical' interface to the Kruskal-Wallis test in R

```
R> kruskal.test(elevel ~ length, data = alpha)

        Kruskal-Wallis rank sum test

data:  elevel by length
Kruskal-Wallis chi-squared = 8.8302, df = 2, p-value =
0.01209
```

However, going beyond the functionality implemented in `kruskal.test` would require extensive programming but is easily possible with the *coin* functionality being available. For example, ignoring the ordinal structure of the allele length is only suboptimal, especially when we have an ordered alternative in mind. Ordinal variables can be incorporated into the general framework via linear-by-linear association tests (Agresti, 2002). When $\mathbf{X}$ is measured at K levels associated with a score vector $\gamma \in \mathbb{R}^{K \times 1}$, the linear statistic reads

$$\mathbf{T}_\gamma = \text{vec}\left(\sum_{i=1}^{n} \gamma^\top g(\mathbf{X}_i) h(\mathbf{Y}_i)^\top\right).$$

Here, the mid-points of the intervals used to categorize the allele lengths are a possible choice for the score vector γ and the linear-by-linear association test can be performed by attaching the scores to the variable `length`:

```
R> independence_test(elevel ~ length, data = alpha,
+     ytrafo = function(data) trafo(data, numeric_trafo = rank),
+     scores = list(length = c(2, 7, 11)))

        Asymptotic General Independence Test

data:  elevel by
         groups short < intermediate < long
T = 2.9263, p-value = 0.003430
```

The smaller p-value corresponds well with Figure 1, i.e., the impression that the expression levels increase with increasing allele lengths.

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 Cochran-Mantel-Haenszel test:

```
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., $h(\mathbf{Y}_i) = (1, 0, 0)^\top$ and $g(\mathbf{X}_i) = (1, 0, 0, 0)^\top$ for a non-smoking Alzheimer patient. Consequently, the linear multivariate statistic $\mathbf{T}$ based on g and h is the 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.

When we investigate the association between smoking and Alzheimer's disease separately for women and men it turns out that the deviation from independence is due to men only

```
R> pvalue(independence_test(as.table(alzheimer[, , "Male"]),
+     teststat = "quadtype"))

[1] 3.169418e-06

R> pvalue(independence_test(as.table(alzheimer[, , "Female"]),
+     teststat = "quadtype"))

[1] 0.09060652
```

and thus we focus on the male patients in the following.

The form of the deviation from independence is of special interest. However, a c_{quad}-type test statistic is not particular useful for this purpose because the contributions of all cells in the contingency table are collapsed in this quadratic form. Instead, we define the test statistic as the maximum of the standardized contingency table via

```
R> males <- as.table(alzheimer[, , "Male"])
R> it_alzmax <- independence_test(males, teststat = "maxtype")
R> it_alzmax

        Asymptotic General Independence Test

data:  disease by groups None, <10, 10-20, >20
T = 4.9504, p-value = 1.148e-05
```

The standardized contingency table sheds some light on the deviations from independence

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

      Alzheimer's Other dementias Other diagnoses
None    2.5900465       -2.340275      -0.1522407
<10     2.9713093       -2.056864      -0.8446233
10-20  -0.7765307       -1.237441       2.1146396
>20    -3.6678046        4.950373      -1.5303056
```

and leads to the impression that patients suffering from Alzheimer's disease smoked less cigarettes than expected under independence and, to an even 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 95% quantile of the permutation null distribution which is available from

```
R> qperm(it_alzmax, 0.95)

[1] 2.812946
```

Alternatively and more conveniently, we can to switch to the p-value scale:

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

      Alzheimer's Other dementias Other diagnoses
None  0.092422483    1.708823e-01       0.9999984
<10   0.031642807    3.075011e-01       0.9719619
10-20 0.981646338    8.417658e-01       0.2748693
>20   0.002842801    7.808092e-06       0.6621122
```

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 and a small number of heavy smokers suffering from Alzheimer's disease.

The levels of smoking arise from a underlying discrete variable and we should make use of this information by applying a linear-by-linear association test. A natural choice of the scores are the mid-points of the internals used to discretize the number of cigarettes per day and we can set up a linear-by-linear association test with $c_{\max}$-type test statistic by attaching scores to variable `smoking`:

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

[1] 7.446659e-05
99 percent confidence interval:
 6.425674e-05 8.467643e-05
```

The single-step adjusted p-values

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

  Alzheimer's Other dementias Other diagnoses
 0.0001379306    7.877645e-05       0.9334802
```

support the conclusion that smoking is associated with both other dementia and 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 2 depicts the group effects for all three response variables.

[Figure 2 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> 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 = 9.456e-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 $h(\mathbf{Y}_1) = (-1.08, -0.56, 5)^\top$ and $g(\mathbf{X}_1) = (1, 0, 0)^\top$. 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 covariance structure Σ 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 $\mathbf{T}$

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

  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           0.13581           0.18968 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           0.13202           0.18542 0.9999
B           0.00000           0.00008 1.0000
C           0.00000           0.00000 0.9989
```

The more powerful step-down adjusted p-values (Algorithm 2.8 in Westfall and Young, 1993) are

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

  Surv(time, event) Surv(dmin, tumor)  ntumor
A           0.08236           0.09850 0.95520
B           0.00000           0.00004 0.88758
C           0.00000           0.00000 0.91998
```

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 three applications, standard transformations for g and h such as dummy codings, ranks and logrank scores have been applied. In the third application, we will show how one can utilize the Lego system 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 3). 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 test) and may be viewed as a multivariate form of the Wilcoxon-Mann-Whitney test.

[Figure 3 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> 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 4. It turns out that using the normal approximation would be sufficient for all practical purposes in this application.

[Figure 4 about here.]

4 Discussion

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 some examples). In practical set ups, an implementation of this two-step procedure requires a certain amount of programming and computing time. Sometimes, permutation tests are even regarded as being 'computationally impractical' for larger sample sizes (Balkin and Mallows, 2001).

The permutation test framework by Strasser and Weber (1999) helps us to take a fresh look at conditional inference procedures and 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.

It is one mission, if not *the* mission, of statistical computing to transform new theoretical developments into flexible software tools for the data analyst. The *coin* package is an attempt to translate the theoretical concepts of Strasser and Weber (1999) into software tools as closely as possible preserving the simplicity and flexibility of the theory. With this package, the rather inflexible spanners currently in use, such as `wilcox.test` for the Wilcoxon-Mann-Whitney test or `mantelhaen.test` for the Cochran-Mantel-Haenszel χ^2 test in S languages and

`NPAR1WAY` for linear rank statistics in SAS (see the Tables in Oster, 2002, 2003, for an overview on procedures implemented in StatXact, LogXact, Stata, SAS and Testimate), are extended by `independence_test`, a much more flexible and adjustable spanner.

But who stands to benefit from such a software infrastructure? We argue that an improved 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 both researchers and 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.

With a unifying conceptual framework in mind and a software implementation, such as *coin*, at hand, we are no longer limited to already published and implemented permutation test procedures and are free to define our own transformations and influence functions, can choose several forms of suitable test statistics and utilize several methods for the computation or approximation of the conditional distribution of the test statistic of interest. Thus, the construction of an appropriate permutation test, for both classical new inference problems, is only a matter of putting together adequate Lego bricks.

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.

Bönsch, D., Lederer, T., Reulbach, U., Hothorn, T., Kornhuber, J., and Bleich, S. (2005), "Joint Analysis of the NACP-REP1 Marker Within the Alpha Synuclein Gene Concludes Association with Alcohol Dependence," *Human Molecular Genetics*, 14, 967–971.

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.

Hothorn, T., Hornik, K., van de Wiel, M., and Zeileis, A. (2005), *coin: Conditional Inference Procedures in a Permutation Test Framework*, R package version 0.4-1, `http://CRAN.R-project.org`.

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.

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.

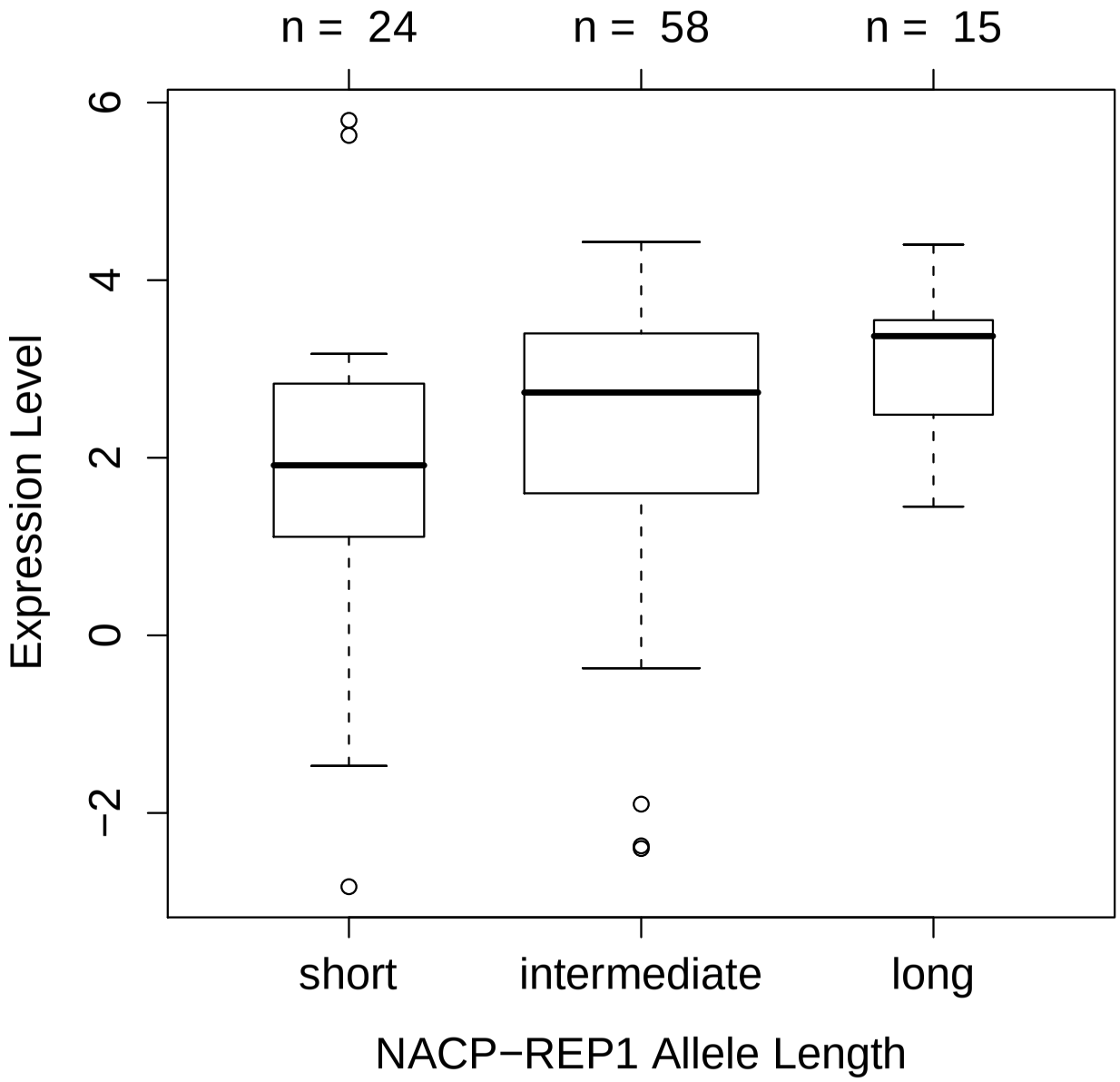

Figure 1: `alpha` data: Distribution of levels of expressed alpha synuclein mRNA in three groups defined by the *NACP*-REP1 allele lengths.

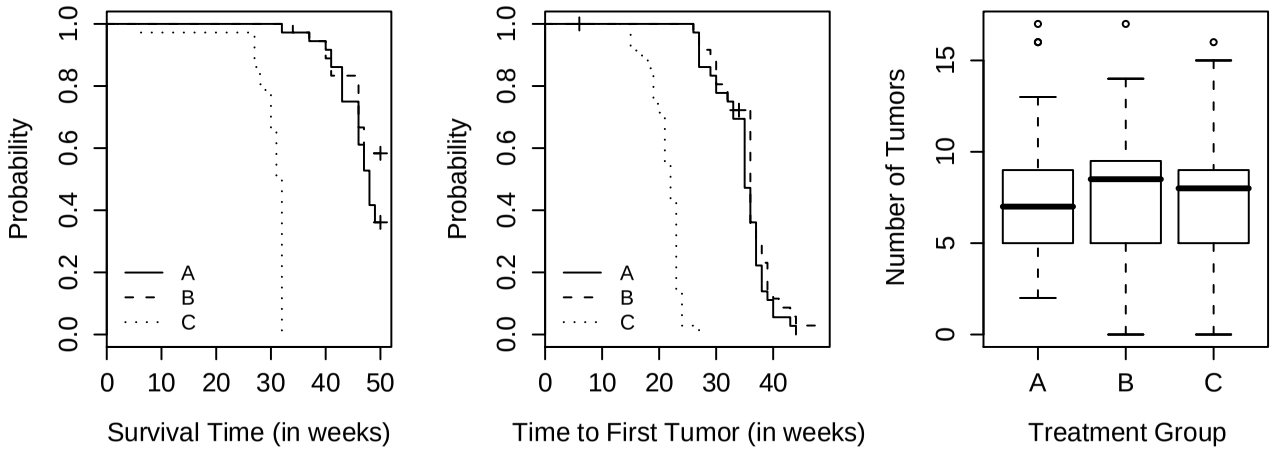

Figure 2: `photocar` data: Kaplan-Meier estimates of time to death and time to first tumor as well as boxplots of the total number of tumors in three treatment groups.

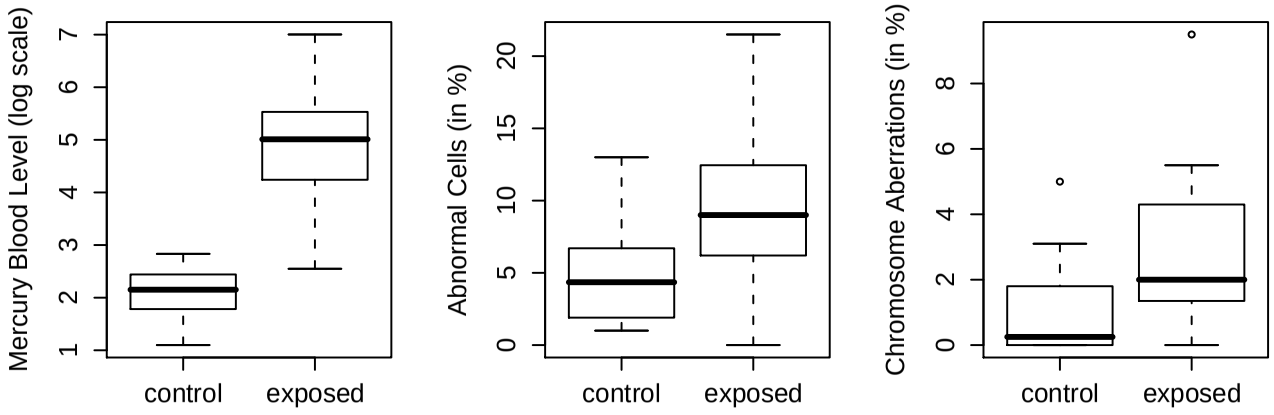

Figure 3: `mercuryfish` data: Distribution of all three response variables in the exposed group and control group.

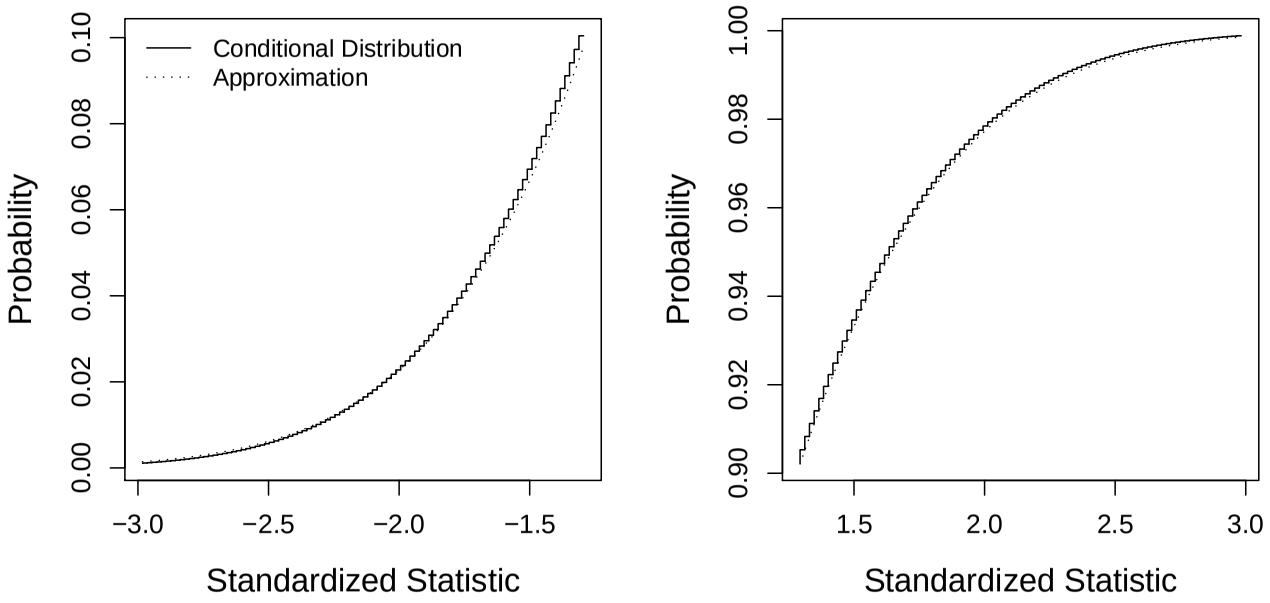

Figure 4: `mercuryfish` data: Conditional distribution and asymptotic normal approximation for the POSET test.

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