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-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, Vrije Universiteit
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
mark.vdwiel@vumc.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 four data sets by adapting the general conceptual framework to these challenging inference problems and utilizing the **coin** add-on package in the R system for statistical computing to show what one can gain from going beyond the 'classical' test procedures.

KEY WORDS: Permutation tests; Independence; Asymptotic distribution; Software.

*This is a preprint of an article published in The American Statistician, Volume 60, Number 3, Pages 257–263. available online at http://www.amstat.org/publications/tas/.

1 INTRODUCTION

The distribution of a test statistic under the circumstances of a 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; Pesarin, 2001; 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 derive the conditional null distribution (by means of exact computations or approximations). 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 whose main ideas are sketched in Section 2.

Currently, the statistician's toolbox consists of rather specialized 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.*, 2006) in the R system for statistical computing (R Development Core Team, 2005).

2 A CONCEPTUAL LEGO SYSTEM

To fix notations, we assume that we are provided with independent and identically distributed 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

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

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 $h(\mathbf{Y}_i) = h(\mathbf{Y}_i, (\mathbf{Y}_1, \ldots, \mathbf{Y}_n))$ may depend on the full vector of responses $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, however only in a permutation symmetric way, i.e., the value of the function must not depend on the order in which $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ appear. We will give several 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 $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Tests that have been constructed by means of this conditioning principle are called *permutation tests*.

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):

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

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

where $\otimes$ denotes 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 a linear statistic $\mathbf{T} \in \mathbb{R}^{pq \times 1}$ into the real line can be of arbitrary form. Obvious choices are the maximum of the absolute values of the standardized linear statistic or a quadratic form:

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

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 tractable for small and moderate sample sizes. In principle, resampling procedures 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 of all permutations S. It is important to note that in the presence of a grouping of the observations into independent 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 a normal approximation of the conditional distribution can be computed for arbitrary choices of g and h. Strasser and Weber (1999) showed in their 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 ($pq = 1$) and by numerical algorithms in the multivariate case (Genz, 1992). For quadratic forms c_{quad} which follow a χ^2 distribution with degrees of freedom given by the rank of Σ (see Johnson and Kotz, 1970, Chapter 29), 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 comes with R code performing the appropriate conditional test using the **coin** functionality which enables the data analyst to benefit from this simple methodology in typical data analyses. All following analyses are reproducible from the **coin** package vignette; this document can be accessed via

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

directly in R.

Independent K-Samples: 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 in alcohol-dependent subjects (see Figure 1). Allele length is measured as a sum score built from additive dinucleotide repeat length and categorized into three groups: short ($0 - 4$, $n = 24$), intermediate ($5 - 9$, $n = 58$), and long ($10 - 12$, $n = 15$).

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 $\mathbf{Y}_i$ in $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Thus, the linear statistic $\mathbf{T}$ is the vector of rank sums in each of the three groups and the test statistic is a quadratic form $(\mathbf{T} - \mu)\Sigma^+(\mathbf{T} - \mu)^\top$ utilizing the conditional expectation μ and covariance matrix Σ. For computing p-values, the limiting χ^2 distribution is typically used.

In R, this specific test is readily implemented in the well established function `kruskal.test` which takes a symbolic formula description of the inference problem and a data set containing the actual observations as its main arguments. Here, the independence of expression levels (`elevel`) and allele lengths (`alength`) is formulated as `elevel ~ alength`, the associated observations are available in a data frame `alpha`:

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

        Kruskal-Wallis rank sum test
```

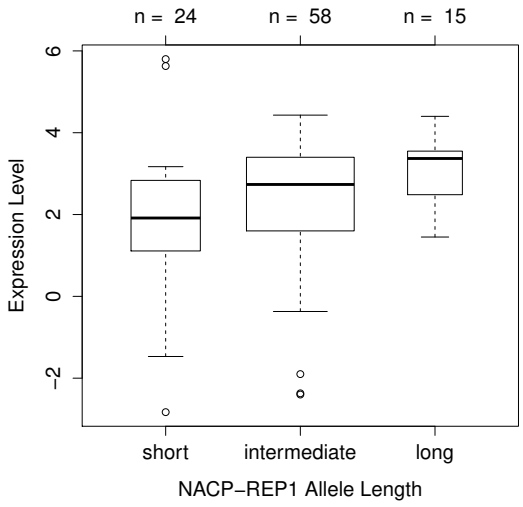

Figure 1: `alpha` data: Distribution of levels of expressed alpha synuclein mRNA in three groups defined by the *NACP*-REP1 allele lengths.

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

Alternatively, the same result can be obtained by embedding the classical Kruskal-Wallis test into the more general conditional inference framework implemented in the `independence_test` function in the **coin** package. This also takes a formula and a data frame as its main arguments and additionally allows for the specification of the transformations g and h via `xtrafo` and `ytrafo`, respectively, as well as setting `teststat` to `"maximum"` or `"quadratic"` (for $c_{\max}$ or c_{quad}, respectively) and the `distribution` to be used. Thus, for computing the Kruskal-Wallis test `ytrafo` has to be set to the function `rank_trafo` for computing ranks, in `xtrafo` dummy codings have to be used (the default for categorical variables), `teststat` is the `"quadratic"` type statistic c_{quad} and the default asymptotic `distribution` is applied:

```
R> independence_test(elevel ~ alength, data = alpha, ytrafo = rank_trafo, teststat = "quadra

        Asymptotic General Independence Test

data:  elevel by
         alength (short, intermediate, long)
chi-squared = 8.8302, df = 2, p-value = 0.01209
```

The output gives equivalent results as reported by `kruskal.test` above. So what is the advantage of using `independence_test`? Going beyond the classical functionality in `kruskal.test` would require extensive programming but

is easily possible in `independence_test`. For example, the resampling distribution instead of the asymptotic distribution could be used by setting `distribution = approximate()`. More interestingly, ignoring the ordinal structure of the allele length is suboptimal, especially when we have an ordered alternative in mind. An intuitive idea for capturing the ordinal information would be to assign numeric scores to the allele length categories in the transformation g rather than the dummy codings used above. A natural choice of scores would be the mid-points of the intervals originally used to categorize the allele lengths, i.e., $g(X_i) = 2$ for short ($\in [0, 4]$), 7 for intermediate ($\in [5, 9]$) and 11 for long ($\in [10, 12]$) alleles. In R, such a function g is easily implemented as

```
R> mpoints <- function(x) c(2, 7, 11)[unlist(x)]
```

which returns an n-vector and can then be passed as `xtrafo` argument to `independence_test`:

```
R> independence_test(elevel ~ alength, data = alpha, ytrafo = rank_trafo, xtrafo = mpoints)

        Asymptotic General Independence Test

data:  elevel by
         alength (short, intermediate, long)
Z = 2.9263, p-value = 0.00343
alternative hypothesis: two.sided
```

This p-value emphasizes the impression from Figure 1 that the expression levels increase with increasing allele lengths. Note that due to usage of scalar transformations g and h, the $c_{\max}$- and c_{quad}-type test statistics are equivalent and hence `teststat` is not set (defaulting to `"maximum"`). Furthermore, it should be pointed out that a test based on such a numerical transformation for ordinal variables is equivalent to linear-by-linear association tests (Agresti, 2002) for which further convenience infrastructure is available in the `independence_test` function via the `scores` argument.

Contingency Tables: 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 female and male Alzheimer patients and 340 controls. The data shown in Table 1 have been re-constructed from Table 4 in Salib and Hillier (1997) and are depicted in Figure 2. The authors conclude that 'cigarette smoking is less frequent in men with Alzheimer's disease.'

We are interested to assess whether there is any association between smoking and Alzheimer's (or other dementia) diseases and, in a second step, how a potential association can be described. First, the global null hypothesis of independence between smoking behavior and disease status for both females and males, i.e., treating gender as a block factor, can be tested with a c_{quad}-type test statistic, i.e., the Cochran-Mantel-Haenszel test:

```
R> it_alz <- independence_test(disease ~ smoking | gender, data = alzheimer,
                               teststat = "quadratic")
R> it_alz
```

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

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

```
        Asymptotic General Independence Test

data:  disease by
         smoking (None, <10, 10-20, >20)
         stratified by gender
chi-squared = 23.316, df = 6, p-value = 0.0006972
```

which suggests that there is a clear deviation from independence. By default, the influence function h (the `ytrafo` argument) and the transformation g (the `xtrafo` argument) are dummy codings of the factors disease status $\mathbf{Y}$ and 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 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.

Given that there is significant departure from independence, we further investigate the structure of association between smoking and Alzheimer's disease. First we assess for which gender the violation of independence occured, i.e., perform independence tests for female and male subjects separately

```
R> females <- alzheimer$gender == "Female"
R> males <- alzheimer$gender == "Male"
R> pvalue(independence_test(disease ~ smoking, data = alzheimer,
          subset = females, teststat = "quadratic"))
```

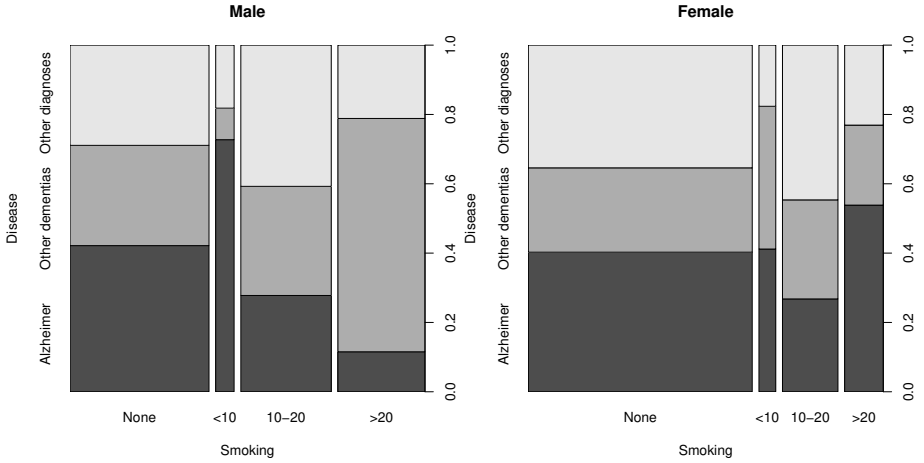

Figure 2: `alzheimer` data: Association of smoking behavior and disease status stratified by gender.

```
[1] 0.09060652
```

```
R> pvalue(independence_test(disease ~ smoking, data = alzheimer,
          subset = males, teststat = "quadratic"))
```

```
[1] 3.169418e-06
```

where it turns out that the association is due to the male patients only (see also Figure 2) and we therefore focus on the male patients in the following. A standardized contingency table is useful for gaining insight into the association structure of contingency tables. Thus, a test statistic based on the standardized linear statistic $\mathbf{T}$ (and thus the standardized contingency table) would be more useful than a c_{quad}-type test statistic where the contributions of all cells are collapsed in such a quadratic form. Therefore, we choose the maximum of the standardized contingency table as $c_{\max}$ test statistic via

```
R> it_alzmax <- independence_test(disease ~ smoking, data = alzheimer,
          subset = males, teststat = "maximum")
R> it_alzmax

        Asymptotic General Independence Test

data:  disease by smoking (None, <10, 10-20, >20)
maxT = 4.9504, p-value = 6.781e-06
alternative hypothesis: two.sided
```

where the underlying standardized contingency table highlights the cells with deviations from independence

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

       Alzheimer 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
```

This leads to the impression that heavy smokers suffer less frequently from Alzheimer's disease but more frequently from other dementias than expected under independence. However, interpreting the standardized contingency table requires knowledge about the distribution of the standardized statistics. An approximation of the joint distribution of all elements of the standardized contingency table can be obtained from the 12-dimensional multivariate limiting normal distribution of the linear statistic $\mathbf{T}$. Most useful is an approximation of the 95% quantile of the permutation null distribution which is available from

```
R> qperm(it_alzmax, 0.95)

[1] 2.814126
```

Alternatively, and more conveniently, one can switch to p-values adjusted for multiple testing by a single-step max-T multiple testing approach:

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

       Alzheimer Other dementias Other diagnoses
None  0.09269080    1.707170e-01       0.9999984
<10   0.03160184    3.072381e-01       0.9719564
10-20 0.98165407    8.418199e-01       0.2751902
>20   0.00271631    7.906569e-06       0.6622096
```

These results support the conclusion that the rejection of the null hypothesis of independence is due to a large number of patients with other dementias and a small number with Alzheimer's disease in the heavy smoking group. In addition, there is some evidence that, for the small group of men smoking less than ten cigarettes per day, the reverse association is true.

Multivariate Response: 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 hypothesis of no treatment effect with respect to any of the three responses 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.

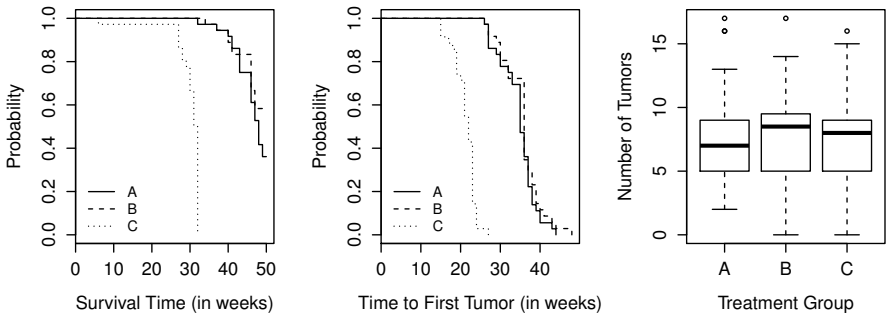

Figure 3: `photocar` data: Kaplan-Meier estimates of time to death and time to first tumor as well as boxplots of the total number of tumors in three treatment groups.

Molefe *et al.* (2005) report data of an experiment where 108 female mice were exposed to different levels of UVR (group A: $n = 36$ with topical vehicle and 600 Robertson–Berger units of UVR, group B: $n = 36$ without topical vehicle and 600 Robertson–Berger units of UVR and group C: $n = 36$ without topical vehicle and 1200 Robertson–Berger units of UVR). The data are taken from Tables 1–3 in Molefe *et al.* (2005), where a parametric test procedure is proposed. Figure 3 depicts the group effects for all three response variables.

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 group (A, B, C)
maxT = 7.0777, p-value = 6.259e-12
alternative hypothesis: two.sided
```

Here, the influence function h consists of the logrank scores (the default `ytrafo` argument for censored observations) 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 linear statistic $\mathbf{T}$ is the sum of each of the three components 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. 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 comparing the standardized linear statistic $\mathbf{T}$ to its critical value 2.715 (which can be obtained by `qperm(it_ph, 0.95)`)

```
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 again by means of the corresponding adjusted p-values

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

```
  Surv(time, event) Surv(dmin, tumor)  ntumor
A           0.13591           0.18946 0.99989
B           0.00001           0.00034 1.00000
C           0.00000           0.00000 0.99859
```

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.

Independent Two-Samples: Contaminated Fish Consumption. In the former three applications, pre-fabricated Lego bricks—i.e., standard transformations for g and h such as dummy codings, ranks and logrank scores—have been employed. In the fourth application, we will show how the Lego system can be used to construct new bricks and implement a newly invented test procedure.

Rosenbaum (1994) proposed to compare groups by means of a *coherence criterion* and studied a data set of subjects who ate contaminated fish for more than three years in the 'exposed' group ($n = 23$) and a control group ($n = 16$). 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 4). The coherence criterion defines a partial ordering: 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 sketched partial ordering) than observation i minus the number of observations that are smaller. The distribution of the rank scores in both groups is to be compared

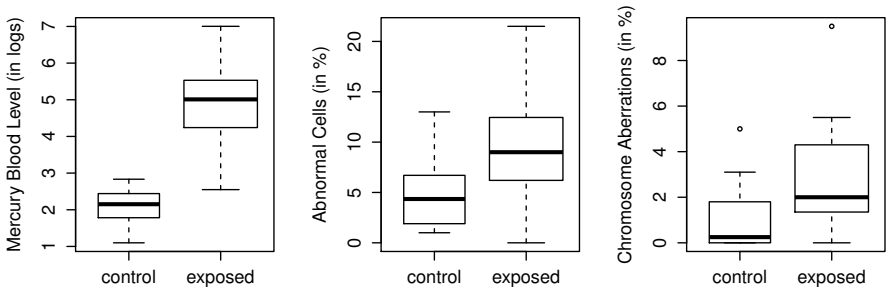

Figure 4: `mercuryfish` data: Distribution of all three response variables in the exposed group and control group.

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.

The coherence criterion can be formulated in a simple function by utilizing column-wise sums of indicator functions applied to all individuals

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

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 (the default 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 5. It turns out that the quality of the normal approximation is excellent for this particular problem and using the normal approximation would be sufficient for all practical purposes in this application.

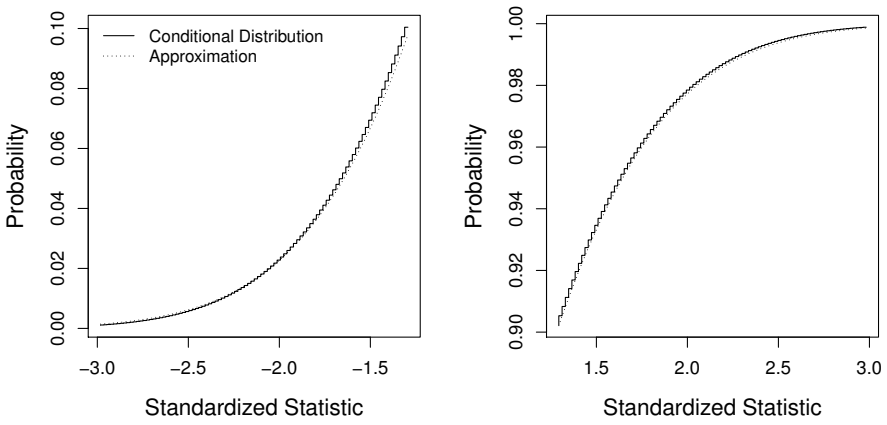

Figure 5: `mercuryfish` data: Conditional distribution and asymptotic normal approximation for the POSET test.

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 setups, 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 preserving the simplicity and flexibility of the theory as closely as possible. With this package, the specialized 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 the S language and `NPAR1WAY` for linear rank statistics in SAS as well as the tools

implemented in StatXact, LogXact, Stata, and Testimate (see Oster, 2002, 2003, for an overview), 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 defined, but the burden of implementing a resampling procedure, or even deriving the limiting distribution of a newly invented test statistic, 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 and new inference problems, is only a matter of putting together adequate Lego bricks.

REFERENCES

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

Balkin SD, Mallows CL (2001). "An Adjusted, Asymmetric Two-Sample t Test." *The American Statistician*, **55**(3), 203–206.

Berger VW (2000). "Pros and Cons of Permutation Tests in Clinical Trials." *Statistics in Medicine*, **19**(10), 1319–1328.

Bönsch D, Lederer T, Reulbach U, Hothorn T, Kornhuber J, Bleich S (2005). "Joint Analysis of the NACP-REP1 Marker Within the Alpha Synuclein Gene Concludes Association with Alcohol Dependence." *Human Molecular Genetics*, **14**(7), 967–971.

Edgington ES (1987). *Randomization Tests*. Marcel Dekker, New York, USA.

Ernst MD (2004). "Permutation Methods: A Basis for Exact Inference." *Statistical Science*, **19**(4), 676–685.

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

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–149.

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

Hothorn T, Hornik K, van de Wiel M, Zeileis A (2006). *coin: Conditional Inference Procedures in a Permutation Test Framework*. R package version 0.4-5, http://CRAN.R-project.org/.

Janssen A, Pauls T (2003). "How Do Bootstrap and Permutation Tests Work?" *The Annals of Statistics*, **31**(3), 768–806.

Johnson NL, Kotz S (1970). *Distributions in Statistics: Continuous Univariate Distributions 2*. John Wiley & Sons, New York.

Ludbrook J, Dudley H (1998). "Why Permutation Tests are Superior to t and F Tests in Biomedical Research." *The American Statistician*, **52**(2), 127–132.

Molefe DF, Chen JJ, Howard PC, Miller BJ, Sambuco CP, Forbes PD, Kodell RL (2005). "Tests for Effects on Tumor Frequency and Latency in Multiple Dosing Photococarcinogenicity Experiments." *Journal of Statistical Planning and Inference*, **129**, 39–58.

Oster RA (2002). "An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods." *The American Statistician*, **56**(3), 235–246.

Oster RA (2003). "An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods–Part II." *The American Statistician*, **57**(3), 201–213.

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

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, URL http://www.R-project.org.

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

Salib E, Hillier V (1997). "A Case-Control Study of Smoking and Alzheimer's Disease." *International Journal of Geriatric Psychiatry*, **12**, 295–300.

Shuster JJ (2005). "Diagnostics for Assumptions in Moderate to Large Simple Clinical Trials: Do They Really Help?" *Statistics in Medicine*, **24**(16), 2431–2438.

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250. Preprint available from http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_94c.