# Exact tests for two-way contingency tables with structural zeros

**Luke J. West**
National Oceanography Centre, Southampton

**Robin K. S. Hankin**
University of Cambridge

**Abstract**

Fisher's exact test, named for Sir Ronald Aylmer Fisher, tests contingency tables for homogeneity of proportion. This paper discusses a generalization of Fisher's exact test for the case where some of the table entries are constrained to be zero. The resulting test is useful for assessing cases where the null hypothesis of conditional multinomial distribution is suspected to be false. The test is implemented in the form of a new R package, **aylmer**.

*Keywords*: Fisher's exact test, computational combinatorics, R, multinomial distribution, rock-paper-scissors, transitivity, pairwise comparison, odds ratio, one-sided tests, structural zeros.

## 1. Introduction

Fisher's exact test (Fisher 1954, page 96) tests 2 × 2 contingency tables for equality of proportion: under the null, the two rows are repeated Bernoulli trials with the same probability of success and failure. It is straightforward to generalize the test to larger tables (Freeman and Halton 1951).

Fisher's test and all the tests considered in this paper share two characteristics. Firstly, they are exact and not asymptotic: they are suitable for small cell counts (Bishop, Fienberg, and Holland 1975). Secondly, they condition on the marginal totals; a cogent discussion for the 2 × 2 case is given by Howard (1998). These tests have been discussed from a Neyman-Pearson perspective (Lehmann 1993), and criticized on the grounds that the marginal totals are informative (Berkson 1978).

The generalized tests are attractive because they are exact, and assess an interesting and plausible null hypothesis: each row comprises independent observations from the same multinomial distribution. However, consider table 1. This dataset is taken from the discipline of industrial quality control: Four machines, A-D, produce articles and the number of defectives is tabulated under various operating conditions. The first line shows a situation in which all four machines are on; the second line shows all machines on except D, and so on. We call such a dataset, with obligatory zero entries—"structural zeros"—a *board*. It is suspected that machine D is causing some sort of interference with machine A; note that machine A produces very few defects except when D is operating.

The null hypothesis is that each row is a sample from a conditional multinomial distribution, conditioned on the machines that are switched off having zero count defectives.

This article introduces software that tests such boards for homogeneity using a generalization

| machine | | | | |
|---|---|---|---|---|
| A | B | C | D | total |
| 4 | 1 | 1 | 1 | 7 |
| 0 | 1 | 2 | - | 3 |
| 2 | 1 | - | 1 | 4 |
| 3 | - | 2 | 0 | 5 |
| 3 | - | - | 2 | 5 |
| 0 | 4 | - | - | 4 |
| 0 | - | 3 | - | 3 |
| 12 | 7 | 8 | 4 | 31 |

Table 1: Industrial quality control (dataset `iqd` in the package) for a factory with four parallel machines A-D. Entries show number of defects attributable to each machine; dashes indicate machines which were switched off for that row. Note that machine A produces no defects when machine D is off

of Fisher's test. The software is written in the C++ programming language and implemented in the form of **aylmer**, an R (R Development Core Team 2008) package. The package provides `aylmer.test()`, a drop-in replacement for `fisher.test()` that can test contingency tables with structural zeros, such as the board shown in Table 1. We have not encountered the statistical test in the literature, and believe that computational implementation of this test is new.

## 2. Methodology and Algorithm

Recall that Fisher's exact test enumerates all contingency tables of a given size with the observed marginal totals; under the null, the probability of each of these is given by the hypergeometric distribution. The p-value of a table is then defined, following Freeman and Halton (1951), as the total probability of all tables more extreme than the observed table: 'more extreme' means that the probability does not exceed that of the observed table. This definition is used in `fisher.test()`.

Given a board, we define a *permissible* board as one with: the same marginal row- and column- totals; structural zeros in the same places; and no negative entries. The **aylmer** package generalizes Fisher's exact test to allow for the possibility of structural zeros; the enumeration operates over permissible boards.

The probability of the observed board occurring (event $A$), given that the board is permissible (event $B$), is determined using the conditional probability rule $\mathsf{P}(A|B) = \mathsf{P}(A \cap B)/\mathsf{P}(B)$. The package determines $\mathsf{P}(B)$ using direct enumeration (permissible boards are enumerated using the algorithm outlined in Figure 1) and the multiple hypergeometric probability mass function (Agresti 2002, p97):

$$\mathsf{P}(B) = \sum_{\substack{\text{permissible}\\ \text{boards}}} \frac{\prod_{i=1}^{r} t_i! \cdot \prod_{j=1}^{c} s_i! \Big/ N!}{\prod_{i=1}^{r} \prod_{j=1}^{c} (n_{ij})!} \qquad (1)$$

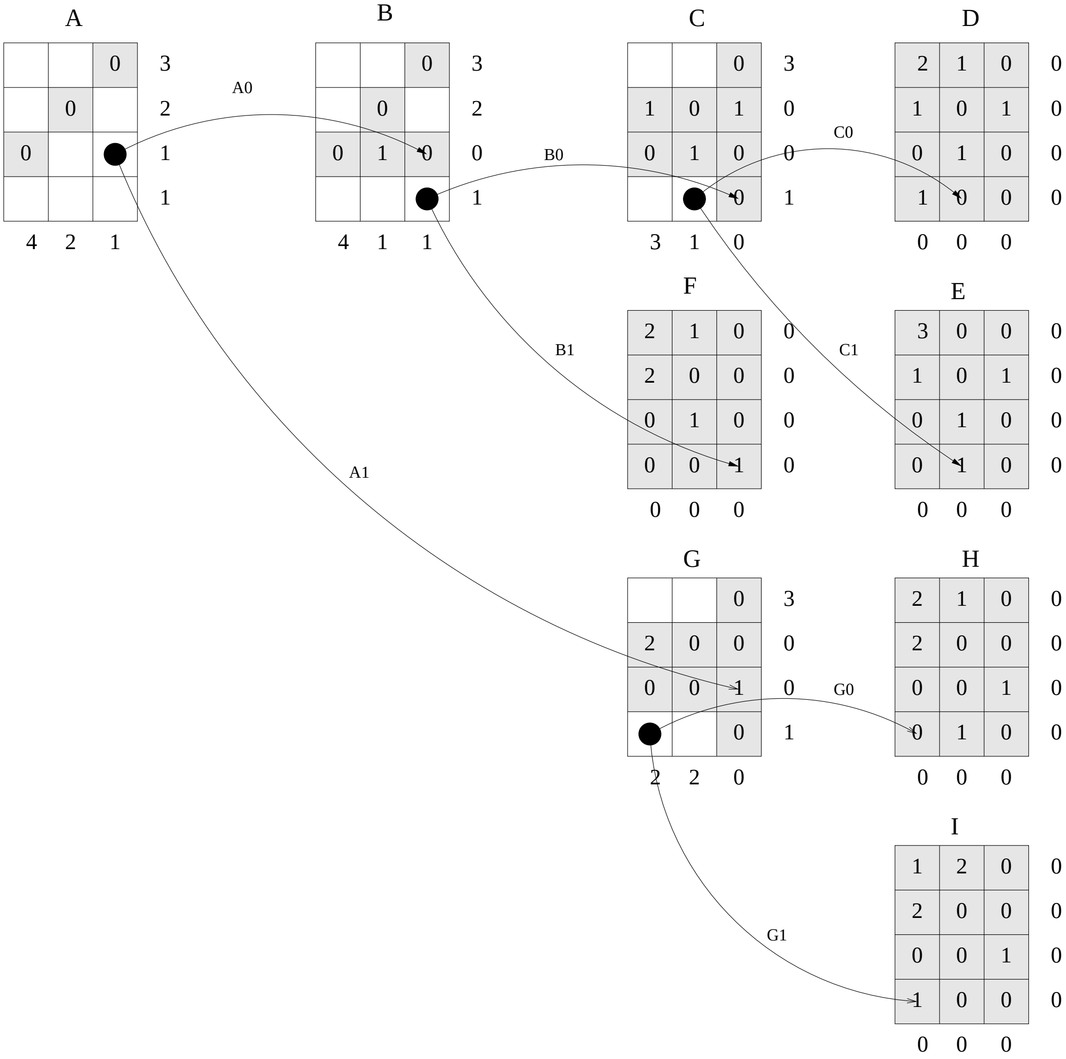

Figure 1: A pictorial description of the algorithm used in function `allboards()` to enumerate all permissible boards. Each table is assigned a letter from A to I. Table A shows the initial configuration; marginal totals are shown and "fixed" entries are shown in grey; in the case of table A these are specified by the user. The algorithm terminates when all entries are fixed and the table becomes totally grayed out. The "pivot" position of a table is shown as a black circle; this is chosen as the square with the lowest variability: the row or column with the smallest marginal is chosen, then the pivot square is the one with the smallest cross-marginal. This indicates position $(3, 3)$ of table A as both marginals of this square are 1. The curved arrows indicate the possible choices for the pivot square and are labelled according to the origin of the table and pivot choice; thus arrow A0 connects A to B (and `B[3,3]=0`), and arrow A1 connects A to G (and `G[3,3]=1`). Filling in the pivot element with 0 in table B allows one to deduce that `B[3,2]=1` and this element appears shaded because it has become fixed; note that the second marginal column sum and third marginal row sum of table B have been reduced by one: the marginal figures represent the marginal sum of the *non-fixed* squares. The algorithm terminates when the entire table is gray or, equivalently, when the residual marginal totals are all zero. Thus tables D,E,F,H,I enumerate the possible tables having the specified marginal totals and specified zero entries

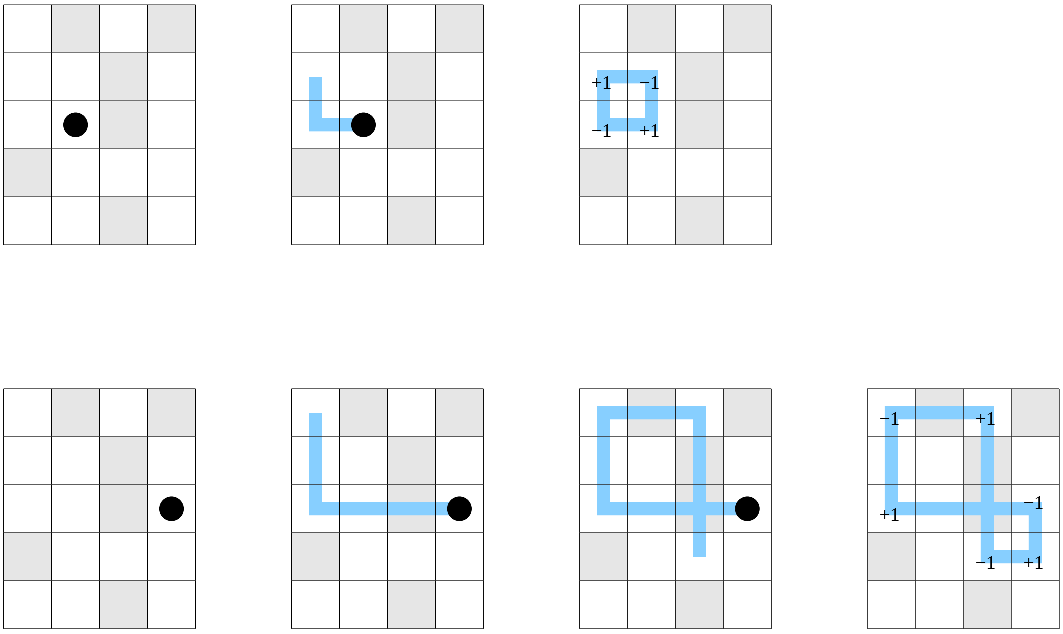

Figure 2: A pictorial description of the algorithm used in function `randomboards()` to generate a random table with given marginal totals. This graphic shows how a table is used to generate, randomly, another table by adding a perturbation called a 'df1 loop' by Aoki and Takemura (2005). The table so generated is called a "candidate" and is either accepted or rejected according to the standard Metropolis-Hastings algorithm (Metropolis *et al.* 1953): the frequencies of tables in the resulting Markov chain are (asymptotically) proportional to their probabilities given by Equation 1.

Two examples, one per row, are shown; structural zeros are shown in gray. The algorithm generates a random closed path of alternating horizontal and vertical lines, shown in light blue. A table generates a candidate table by alternately incrementing and decrementing squares on the corners of the path, thereby preserving the marginal totals.

A non-gray square ("START") is chosen at random; this is marked with a black circle. The path starts by choosing a random non-gray square in the same row as the start; from this square the path continues vertically to another non-gray square. If the loop may be closed (that is, if the square on the same column as START and the same row as the free end is non-gray) then the path is closed and the algorithm terminates. If the path may not be closed (because the relevant square is gray), then the path is extended in a similar fashion; excepting that columns and rows with only a single remaining unshaded square are disallowed. The resulting path then has the property that any row of the board has an even number of path corners: and no path corner lies on a gray square. Thus taking a permissible board and alternately incrementing and decrementing squares along such a path will result in a permissible board: structural zeros (gray squares) remain zeros; and the marginal totals remain unaltered. Modifying a board in this manner yields the candidate sample for the Metropolis-Hastings algorithm.

In the first row, the path proceeds from the START to the square immediately to the left; thence to the square immediately above. The path may be closed because the square immediately above START is non-gray. It is clear that if the board has no non-gray squares, a path of this type (viz: a simple square) is always chosen and the method reduces to that of Raymond and Rousset (1995).

The second line shows a more involved example in which the path needs a second leg to be closed; note how the path crosses itself (not forbidden). Observe that the algorithm applied to boards with an ordered sample space—such as Table 5 or 6—would result in every non-gray square being either incremented or decremented where $s_i = \sum_{j=1}^{c} n_{ij}$ and $t_j = \sum_{i=1}^{r} n_{ij}$ are the row- and column- sums respectively, and $N = \sum s_i = \sum t_j$ is the total board count. The numerator in equation 1 is constant for all permissible boards so its evaluation is not necessary.

Thus the null hypothesis that a particular table with specified zero elements (a board) is in fact drawn at random from all permissible boards is then tested just as in Fisher's test: the p-value is the sum of the probabilities of all permissible boards with a probability less than or equal to that of the observed board.

## 2.1. The number of distinct contingency tables with given marginal totals

There are a number of ways of assessing $M = M(s_1, \ldots, s_r, t_1, \ldots, t_c)$, the number of distinct contingency tables with specified totals. Most results are asymptotic; no simple exact formula for tables as small as $r = s = 3$ is known.

The appropriate generating function for contingency tables is

$$\prod_{i=1}^{r}\prod_{j=1}^{c}\frac{1}{1-x_iy_j}$$

[the number of boards is given by the coefficient of $x_1^{s_1}\cdots x_r^{s_r}y_1^{t_1}\cdots y_c^{t_c}$]. Generalizing this to a board is straightforward; the generating function is

$$\prod_{\substack{1\leqslant i\leqslant r\\ 1\leqslant j\leqslant c\\ (i,j)\in\{1,\ldots,r\}\times\{1,\ldots,c\}\setminus Z}}\frac{1}{1-x_iy_j}$$

where $Z$ is the set of structural zeros. Good (1976) presents arguments that suggest

$$\frac{\prod_{i=1}^{r}\binom{s_i+r-1}{s_i}\prod_{j=1}^{c}\binom{t_j+c-1}{t_j}}{\binom{N+rc-1}{N}}$$

[function `good()` in the package] is asymptotic to $M$. If the number of permissible boards is large, as in the `frogs` or `icons` examples discussed below in section 3.4, the Monte Carlo techniques of Aoki and Takemura (2005) are used; our computational algorithm is outlined in Figure 2. Random permissible boards are generated and the p-value reported is as above except that instead of a complete enumeration, an ensemble of randomly generated boards is used.

# 3. Package aylmer in use

This section illustrates the functionality of the package with examples taken from industrial quality control, sociology of climate change, and behavioural neuropsychology. The special case of pairwise comparisons is then discussed using examples from sports, aviation quality control, and animal behaviour.

### 3.1. Industrial quality control

Table 1 (dataset `iqd` in the package) may be tested straightforwardly using the **aylmer** package:

```
        Aylmer test for count data

data:  iqd
p-value = 0.1297
alternative hypothesis: two.sided
```

Thus the p-value would indicate failure to reject the null hypothesis at the 5% level.

Further investigation suggests instead that machine A produces more defects than expected when all other machines are switched on (row 1). In this case, the appropriate diagnostic would be that `iqd[1,1]` is larger than expected. Following Silvapulle and Sen (2005, page 326) one would define a test statistic `f(x)=x[1,1]` and sum the probabilities of all permissible boards with `f(x)` $\geqslant$ `f(iqd)`. In package **aylmer**, the R idiom is to pass the test statistic, in the form of a function whose domain is the set of permissible boards, to argument `alternative` of `aylmer.test()`:

```
    > f <- function(x){x[1,1]}
    > aylmer.test(iqd,alternative = f)

        Aylmer functional test for count data

data:  iqd
p-value = 0.04404
alternative hypothesis: test function exceeds observed
```

Thus there is sufficient evidence to reject this null at the 5% level.

Dataset `glass` in the **aylmer** package includes an illustration of this technique applied to ordered categorical factors (type “`?glass`” at the R prompt).

*Incomplete cases*

One might consider instead the effect of changing personnel. Suppose now that there are three machines A, B, C and three supervisors S1, S2, S3. It is suspected that the supervisors use slightly differing work practices and that these differences change the ratio of defects produced by the three machines. However, on S1’s shift, defects produced by machine C cannot be detected (perhaps C’s entire output for that day was discarded for some other, unrelated, reason). The null hypothesis would be that the proportions of defects made by the three machines are independent of supervisor—or, that the proportions of defects produced during the shifts of the three supervisors are independent of the machine:

```
> shifts
```

```
        machine
operator A B  C
      S1 9 1 NA
      S2 2 3  8
      S3 3 3  2

> aylmer.test(shifts)

        Aylmer test for count data

data:  shifts
p-value = 0.04752
alternative hypothesis: two.sided
```

showing that one may reject the null hypothesis at the 5% level. Note that in this case the Fisher test is appropriate but may only be used on complete cases:

```
> fisher.test(shifts[-1, ])$p.value

[1] 0.3065015

> fisher.test(shifts[, -3])$p.value

[1] 0.09842621
```

(the first test considering only supervisors S2 and S3; and the second only machines A and B). Thus there is insufficient evidence in either complete case to reject the null hypothesis at the 5% level; compare the aylmer test where all relevant information was retained and the null rejected.

## 3.2. Public perception of climate change

Lay perception of climate change is a complex and interesting process (Lorenzoni and Pidgeon 2005); the issue of immediate practical import is the engagement of non-experts by the use of "icons"[1] that illustrate different impacts of climate change.

In one study (O'Neill 2008), subjects are presented with a set of icons of climate change and asked to identify which of them they find most concerning. Six icons were used: PB [polar bears, which face extinction through loss of ice floe hunting grounds], NB [the Norfolk Broads, which flood due to intense rainfall events], L [London flooding, as a result of sea level rise], THC [the thermo-haline circulation, which may slow or stop as a result of anthropogenic modification of the water cycle], OA [oceanic acidification as a result of anthropogenic emissions of $CO_2$], and WAIS [the West Antarctic Ice Sheet, which is rapidly calving as a result of climate change].

Methodological constraints dictated that each respondent could be presented with a maximum of four icons. Table 2 (dataset `icons` in the package) shows the experimental results.

[1] This word is standard in this context. An icon is a "representative symbol".

| icon | | | | | | |
|---|---|---|---|---|---|---|
| NB | L | PB | THC | OA | WAIS | total |
| 5 | 3 | - | 4 | - | 3 | 15 |
| 3 | - | 5 | 8 | - | 2 | 18 |
| - | 4 | 9 | 2 | - | 1 | 16 |
| 1 | 3 | - | 3 | 4 | - | 11 |
| 4 | - | 5 | 6 | 3 | - | 18 |
| - | 4 | 3 | 1 | 3 | - | 11 |
| 5 | 1 | - | - | 1 | 2 | 9 |
| 5 | - | 1 | - | 1 | 1 | 8 |
| - | 9 | 7 | - | 2 | 0 | 18 |
| 23 | 24 | 30 | 24 | 14 | 9 | 124 |

Table 2: Experimental results from O'Neill (2008) (dataset `icons` in the package): respondents' choice of 'most concerning' icon of those presented. Thus the first row shows results from respondents presented with icons NB, L, THC, and WAIS; of the 15 respondents, 5 chose NB as the most concerning (see text for a key to the acronyms). Note the "0" in row 6, column 9: this option was available to the 18 respondents of that row, but none of them actually chose WAIS

One natural null hypothesis $H_0$ is that there exist $p_1, \ldots, p_6$ with $\sum p_i = 1$ such that the probability of choosing icon $i$ is proportional to $p_i$. Alternative hypotheses are necessarily vague, but it is interesting to note that $H_0$ is "uncritically" adopted in the literature (O'Neill 2008). However, the exact Aylmer test discussed above cannot be used here as the sample space is too large; the number of possible boards consistent with the marginal totals is astronomical:

```
> data("icons")
> good(icons)

[1] 2.043647e+29
```

Setting the `simulate.p.value` flag forces the package to use Monte-Carlo simulation techniques:

```
> aylmer.test(icons, simulate.p.value=TRUE)

        Aylmer test for count data with simulated p-value (based on 2000
        replicates)

data:  icons
p-value = 0.1579
alternative hypothesis: two.sided
```

Thus there is insufficient evidence to reject the null hypothesis[2] and on the assumption that a set of $p_i$ exists, Hankin (2008a) presents software that calculates their values numerically.

[2] Such estimates are necessarily random variables; using a batch method, following Aoki and Takemura (2005), we estimate the true p-value to be $0.164 \pm 0.02$.

| | Husband's sib | | | | | |
|---|---|---|---|---|---|---|
| wife's sib | Marrim | Makan | Parpa | Thao | Kheyang | total |
| Marrim | - | 5 | 17 | - | 6 | 28 |
| Makan | 5 | - | 0 | 16 | 2 | 23 |
| Parpa | - | 2 | - | 10 | 11 | 23 |
| Thao | 10 | - | - | - | 9 | 19 |
| Kheyang | 6 | 20 | 8 | 0 | 1 | 35 |
| total | 21 | 27 | 25 | 26 | 29 | 128 |

Table 3: Data for 128 Purum marriages (dataset `purum` in the package). The Purums are an isolated tribe of India, divided into five sibs. White (1963) argues that the Purum sib is exogamous (that is, within-sib marriages are disallowed; the single Kheyang-Kheyang marriage was a special case) and that males and females could marry only in selected sibs. In the table, a dash denotes combinations forbidden by Purum tradition. Note the lack of symmetry in the structural zeros, which implies a gender asymmetry: thus a male Parpa may marry a female Marrim, but a male Marrim may not marry a female Parpa

It is interesting to note that there exist permissible boards with a probability, according to Equation 1, of over 20000 times that of the `icons` dataset.

The default value of the number of random samples to use in the Monte-Carlo case—argument `B` of `aylmer.test()`—is 2000, following `fisher.test()`. Figure 3 shows an example that illustrates graphically whether a given value is sufficient.

## 3.3. Social anthropology

Table 3 shows an example taken from social anthropology, detailing 128 marriages. The standard null is rejected by the Aylmer test (see online documentation), in agreement with Bishop *et al.* (1975): within the prescriptive framework, preference plays a part. We wish to make inferences about gender asymmetry in the preferential component of the dataset.

Given a pair of sibs, the marriage restrictions imply that at least one is a wife-giver, and at least one is a wife-taker: For example, in the case of Parpa-Marrim marriages, the Marrim are wife-givers and the Parpa are wife-takers.

There are five pairs of sibs that may act as both wife-givers *and* wife-takers. Amongst these pairs, is there evidence to suggest that the preferences are gender asymmetric?

An appropriate test function would be the maximum absolute difference between the number of M-F marriages and F-M marriages, amongst (ordered) pairs of sibs that allow both types of marriages:

```
> g <- function(x) max(abs(x - t(x)), na.rm = TRUE)
```

One would expect `g(.)` to return small values if the sibs' behaviour is indeed gender neutral. This hypothesis may be tested straightforwardly by sampling from permissible boards and reporting the fraction of boards with `g(.)` exceeding that of our observation:

```
> aylmer.test(purum, alternative=g, simulate.p.value=TRUE, B=2000)
```

```
        Aylmer functional test for count data with simulated p-value (based on
        2000 replicates)

data:  purum
p-value = 0.0004998
alternative hypothesis: test function exceeds observed
```

Thus there is strong evidence that Purum marriage preferences are not gender neutral, even after accounting for the incest prohibitions marked by structural zeros.

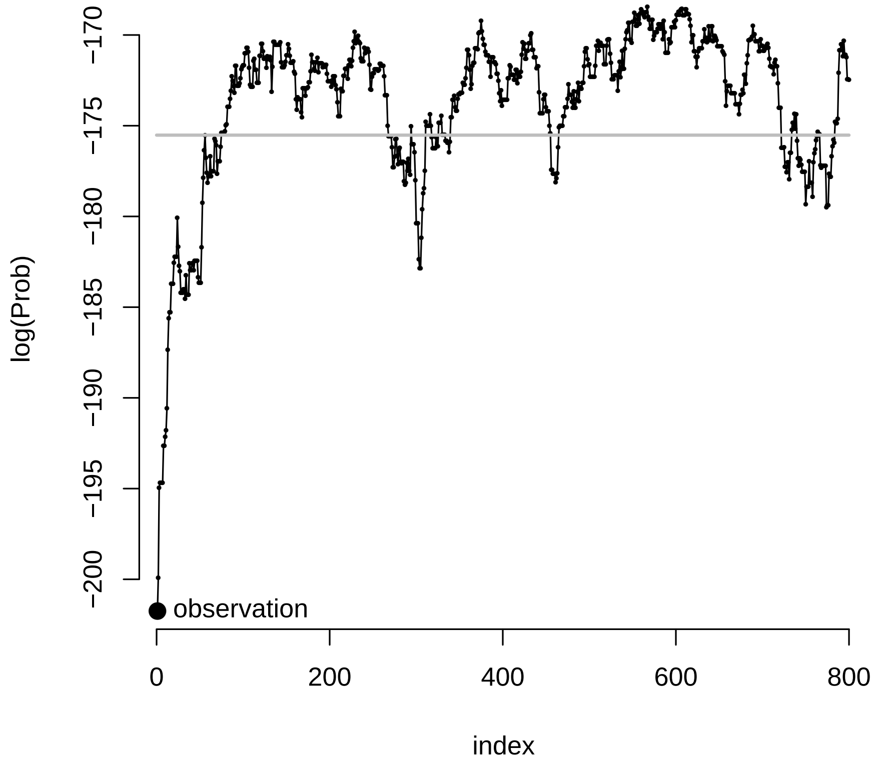

Figure 3: Probabilities of sequential boards in a Markov chain of boards permissible to the *Purum* dataset shown in Table 3. To within a constant, the ordinate is the natural logarithm of the probability of the Markov boards: the gray horizontal line marks the critical region for an Aylmer test of size 5% (any board below this level is rejected). The observation, being the first member of the Markov chain, is clearly in the critical region and the null may be rejected

### 3.4. Pairwise comparison

Although each row of a board is in general a multinomial distribution, by far the most commonly occurring case is when all but two possibilities in each row are disallowed: the entries are then drawn from a binomial distribution, if the null hypothesis is correct. Davidson and Farquhar (1976) give an extensive bibliography of this case.

Many examples exist of repeated pairwise comparisons between two of a larger number of "players". Examples abound in the sporting world (Jech 1983), although in sport the possibility of a draw must sometimes be considered. Non-sporting examples would include forced-choice discrimination (Bradley and Terry 1952): in the field of, say, olfactory research, a subject is repeatedly presented with two odours and asked to report which is preferable (or stronger, or whatever).

The canonical null hypothesis, introduced by Zermelo (1929), is that there exist numbers $\pi_1, \ldots, \pi_n$ ("skills") with $\sum_{i=1}^{n} \pi_i = 1$; a match between player $i$ and $j$ is then a Bernoulli trial with probability $\pi_i/(\pi_i + \pi_j)$; Connor and Grant (2000) give an historical overview. Note that Zermelo's model readily generalizes to situations in which more than two players compete.

Consider the `frogs` dataset, provided with the package and shown in Table 4. This shows the result of repeated forced-choice experiments taken with the intention of investigating intransitive preferences.

In this context, intransitivity is defined as the existence of stimuli $s_1, \ldots, s_n$ with $s_i \longrightarrow s_{i+1}$ for $1 \leqslant i \leqslant n-1$ and $s_n \longrightarrow s_1$, where "$a \longrightarrow b$" means "$a$ was preferred to $b$ with a probability exceeding 0.5 in a forced-choice between $a$ and $b$". Such intransitive preferences are of great interest in the field of animal behaviour as they are readily observable and elucidate the neural algorithms underlying choice; explanation of non-transitive choice is a "challenging problem" (Colgan and Smith 1985) and is "the focus of considerable contemporary research" (Waite 2001). Note that Zermelo's null precludes intransitivity.

However, the Aylmer test discussed above cannot be used here[3] so the `simulate.p.value` flag is again set:

```
> data("frogs")
> aylmer.test(frogs, simulate.p.value=TRUE)

        Aylmer test for count data with simulated p-value (based on 2000
        replicates)

data:  frogs
p-value = 0.06847
alternative hypothesis: two.sided
```

thus the null hypothesis may be rejected, and some form of non-transitive mechanism is required to explain the frogs' choices. Aoki and Takemura's batch method gives $0.016 \pm 0.004$.

It is interesting to compare the approach adopted here with that of Kendall and Babington Smith (1940), who considered pairwise comparison matrices of the form of `frogs.matrix`, also provided with the package:

[3] Function `good()` is not useful in this case because of the large number of `NA` entries. The relevant combinatorics are involved; an example is given in (Hankin 2008b). But it is interesting to consider just the first column (`Sc`). The **partitions** package (Hankin 2007b) can be used to show that this column alone has `S(rep(20,7),110)`=1912757 combinations; it accounts for only 7 of the 28 degrees of freedom available. Also note the large magnitude of the numbers involved; the denominator of Equation 1 is $\simeq 4.6 \times 10^{501}$, necessitating use of the **Brobdingnag** package (Hankin 2007c).

| stimulus | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Sc | Sb | Ob | Oa | Oc | Sa | Sd | Od | M | |
| 10 | 10 | - | - | - | - | - | - | - | 20 |
| 12 | - | 8 | - | - | - | - | - | - | 20 |
| 13 | - | - | 7 | - | - | - | - | - | 20 |
| 14 | - | - | - | 6 | - | - | - | - | 20 |
| 13 | - | - | - | - | 7 | - | - | - | 20 |
| 16 | - | - | - | - | - | 4 | - | - | 20 |
| 15 | - | - | - | - | - | - | 5 | - | 20 |
| 17 | - | - | - | - | - | - | - | 3 | 20 |
| - | 13 | 7 | - | - | - | - | - | - | 20 |
| - | 8 | - | 12 | - | - | - | - | - | 20 |
| - | 12 | - | - | 8 | - | - | - | - | 20 |
| - | 16 | - | - | - | 4 | - | - | - | 20 |
| - | 19 | - | - | - | - | 1 | - | - | 20 |
| - | 15 | - | - | - | - | - | 5 | - | 20 |
| - | 16 | - | - | - | - | - | - | 4 | 20 |
| - | - | 12 | 8 | - | - | - | - | - | 20 |
| - | - | 10 | - | 10 | - | - | - | - | 20 |
| - | - | 14 | - | - | 6 | - | - | - | 20 |
| - | - | 12 | - | - | - | 8 | - | - | 20 |
| - | - | 12 | - | - | - | - | 8 | - | 20 |
| - | - | 18 | - | - | - | - | - | 2 | 20 |
| - | - | - | 10 | 10 | - | - | - | - | 20 |
| - | - | - | 10 | - | 10 | - | - | - | 20 |
| - | - | - | 16 | - | - | 4 | - | - | 20 |
| - | - | - | 16 | - | - | - | 4 | - | 20 |
| - | - | - | 11 | - | - | - | - | 9 | 20 |
| - | - | - | - | 5 | 15 | - | - | - | 20 |
| - | - | - | - | 10 | - | 10 | - | - | 20 |
| - | - | - | - | 12 | - | - | 8 | - | 20 |
| - | - | - | - | 18 | - | - | - | 2 | 20 |
| - | - | - | - | - | 14 | 6 | - | - | 20 |
| - | - | - | - | - | 9 | - | 11 | - | 20 |
| - | - | - | - | - | 11 | - | - | 9 | 20 |
| - | - | - | - | - | - | 15 | 5 | - | 20 |
| - | - | - | - | - | - | 12 | - | 8 | 20 |
| - | - | - | - | - | - | - | 7 | 13 | 20 |
| 110 | 109 | 93 | 90 | 79 | 76 | 60 | 53 | 50 | 720 |

Table 4: Experimental results of Kirkpatrick *et al.* (2006), included as the `frogs` dataset in the package. Each row corresponds to a series of forced-choice experiments in which a female túngara frog was exposed to two stimuli (mating calls of male frogs). The entries show the results; thus the first row shows that, when given a choice between stimulus `Sc` and stimulus `Sb`, each was chosen 10 times. Full details are given by Kirkpatrick *et al.* (2006) and Ryan and Rand (2003).

```
> frogs.matrix

   Sc Sb Ob Oa Oc Sa Sd Od  M
Sc NA 10  8  7  6  7  4  5  3
Sb 10 NA  7 12  8  4  1  5  4
Ob 12 13 NA  8 10  6  8  8  2
Oa 13  8 12 NA 10 10  4  4  9
Oc 14 12 10 10 NA 15 10  8  2
Sa 13 16 14 10  5 NA  6 11  9
Sd 16 19 12 16 10 14 NA  5  8
Od 15 15 12 16 12  9 15 NA 13
M  17 16 18 11 18 11 12  7 NA
```

This matrix contains the same data as the `frogs` dataset shown in Table 4 in a more compact form (the first line of `frogs` appears as elements `[1,2]` and `[2,1]`). Kendall and Babington Smith (1940) considered the special case of such matrices where each entry was 0 or 1, thus corresponding to the case where the female frog was presented with each pairwise choice exactly once. Their test counts the number of circular triads[4] appearing in the table; the asymptotic distribution of this statistic is known under the null which gives a critical region. Knezek, Wallace, and Dunn-Rankin (1998) noted that the test was "computationally intense"—the complexity rising as $\mathcal{O}(2^{k!})$—and presented an asymptotic approximation.

We suggest that our test is not directly comparable to that of Kendall and Babington Smith (it is clear that our test fails to reject *any* board whose elements are all zero or one) but further work would be required to explore any relationship.

### *One-tailed and two-tailed tests in pairwise comparison*

The general problem of comparing $n$ players $p_1, \ldots, p_n$ potentially has $n(n-1)/2$ pairwise comparisons. The system of players and possible comparisons may be represented as a graph (Bollobás 1979); two nodes (players) are connected by an edge if and only if they compete against one another.

If the only comparisons that may be made are between $p_i$ and $p_{i+1}$ for $1 \leqslant i \leqslant n-1$ (and between $p_1$ and $p_n$), then the competition graph becomes cyclic. The sample space possesses a natural ordering, because the board has only a single degree of freedom, and one-sided tests become possible. A succinct overview of one-sided and two-tailed tests in the context of two by two contingency tables is given by Ghent (1972).

When considering $2 \times 2$ contingency tables (Agresti 2002), one often considers the *odds ratio* $\theta$ defined as

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

where $\pi_1$ and $\pi_2$ are the binomial probabilities of the first and second rows respectively [the *odds* of an event with probability $\pi$ are defined to be $\pi/(1-\pi)$] . The maximum likelihood estimate for the odds ratio is given by $\hat{\theta} = \frac{ad}{bc}$.

[4] Following Alway (1962), a circular triad is a triple of stimuli $A$, $B$, $C$ with either $A \to B \to C \to A$ or $A \to C \to B \to A$.

| Topalov | Anand | Karpov | total |
|---|---|---|---|
| 22 | 13 | - | 35 |
| - | 23 | 12 | 34 |
| 8 | - | 10 | 18 |
| 30 | 36 | 22 | 87 |

Table 5: Intransitive example of chess players (dataset `chess` in the package); entries show number of games won up to 2001 (draws are discarded). Topalov beats Anand 22-13; Anand beats Karpov 23-12; and Karpov beats Topalov 10-8. Games between these three players thus resemble a noisy version of iterated "rock-paper-scissors"

Tables 5, 6 and 7 immediately suggest a generalization of the odds ratio, which is the product of the odds of each edge in the competition graph [`odds.ratio()` in the package]: if the data is organized as in these boards, the generalized odds ratio is given by the product of the elements on the leading diagonal, divided by the product of the off-diagonal elements. In the case of Table 5, the maximum likelihood estimate for the generalized odds ratio would be $\frac{22\cdot 23\cdot 10}{13\cdot 12\cdot 8} \simeq 4.04$, and in Table 7 it is $\simeq 0.00638$.

The generalized odds ratio thus furnishes a natural ordering of a sample space: simply order the sample space from lowest generalized odds ratio to largest; Table 6 enumerates a small sample space and illustrates how the ordering works.

The simplest nontrivial example of pairwise comparison would be to consider three players A, B, and C who compete in pairs. This case was considered by Bradley (1954), although the test presented was asymptotic, and not exact. Triads of players with Player A beating B, player B beating C *and* player C beating A certainly exist (Table 5 shows a real example, taken from the chess world). Such players form a circular triad in the sense of Knezek *et al.* (1998) but here we allow *repeated* comparisons (matches).

Further examples are found in biology: male side-blotched lizards are territorial and possess three variants (yellow, orange, blue). Territory held by Y is lost to O, territory held by O is lost to B, and territory held by B is lost to Y (Sinervo and Lively 1996). Competition between these three morphs is thus a noisy version of "rock-paper-scissors" (Wikipedia 2007), a system encountered in diverse scientific contexts including population ecology (Frean and Abraham 2001), game theory (Szabó and Fáth 2007), and sociology (Semmann, Krambeck, and Milinski 2003).

In these examples, non-transitivity often has a plausible mechanism, whose existence serves as an alternative hypothesis and indicates a one-tailed test; this would be a generalization of the one-tailed Fisher's exact test for the $2 \times 2$ case. In the case of the side-blotched lizard, O beats Y through aggression, B beats O through concentrating on defending only a small territory, and Y beats B through stealth.

In many branches of engineering, one encounters systems which comprise components arranged in a circular configuration. Each component may be compared only against the two adjacent components (Hankin 2007a). Commonly occurring examples include turbine blades, ball bearings, and gear teeth. The comparisons might involve objective measurements—such as turbine blade lengths—or subjective quantities, such as amount of wear. It is desired to determine whether the measurement system possesses a 'handedness', in that (for example), the clockwise blade is judged to be longer more frequently than reasonable. Table 7 shows an

| A | B | C | D |
|---|---|---|---|
| 0 | 3 | - | - |
| - | 3 | 9 | - |
| - | - | 1 | 4 |
| 4 | - | - | 3 |

| A | B | C | D |
|---|---|---|---|
| 1 | 2 | - | - |
| - | 4 | 8 | - |
| - | - | 2 | 3 |
| 3 | - | - | 4 |

| A | B | C | D |
|---|---|---|---|
| 2 | 1 | - | - |
| - | 5 | 7 | - |
| - | - | 3 | 2 |
| 2 | - | - | 5 |

| A | B | C | D |
|---|---|---|---|
| 3 | 0 | - | - |
| - | 6 | 6 | - |
| - | - | 4 | 1 |
| 1 | - | - | 6 |

Table 6: An ordered sample space. Rows show the result of repeated pairwise comparisons of four players, A-B, B-C, C-D, D-A. Marginal totals are held constant. From left to right, the generalized odds ratios are $0, \frac{2}{9}, \frac{75}{14}, \infty$. Suppose the first board were the observation and the null hypothesis is to be tested against the (one-sided) alternative hypothesis that the odds ratio is smaller than that observed: in practice, this would be conceptualized as $A \to B \to C \to D \to A$, where "$X \to Y$" means that the probability of $X$ beating $Y$ exceeds 0.5. Note that all four scorelines are consistent with the alternative hypothesis. Then the one-sided p-value would be $\left(\Sigma \cdot 3!^3 4!^2 9!\right)^{-1} \simeq 0.0353$ where $\Sigma = \left(3!^3 4!^2 9!\right)^{-1} + \left(2!^2 3!4!^2 8!\right)^{-1} + \left(2!^3 3!5!^2 7!\right)^{-1} + \left(3!4!6!^2\right)^{-1}$. The two-sided p-value would be $\frac{1}{\Sigma}\left[\left(3!^3 4!^2 9!\right)^{-1} + \left(3!4!6!^2\right)^{-1}\right] \simeq 0.065$

example taken from the field of aviation quality control; it is given in the **aylmer** package as the `gear` dataset:

```
> data(gear)
> aylmer.test(gear)

        Aylmer test for count data

data:  gear
p-value = 0.05094
alternative hypothesis: two.sided
```

showing that a two sided test is not significant at the 5% level, although it is interesting to observe that the one-sided test has a p-value of about $6.651 \times 10^{-5}$. Note the natural one-sidedness of any significance test of this type: the preference may be clockwise or anticlockwise, corresponding to high or low values of the odds ratio.

## 4. Conclusions

Fisher's test is attractive because it is exact, and tests an interesting and plausible null hypothesis: each row comprises independent observations from the same multinomial distribution. In this paper, we present a generalization of Fisher's exact test, with the same null except that the rows comprise independent *conditional* observations from the same multinomial distribution. The natural null hypothesis is an interesting and useful construction in a variety of scientific, industrial, and sociological contexts.

Throughout this paper, the ensemble considered is that of permissible boards. By default, the critical set includes all permissible boards with conditional probabilities not exceeding that of the observation; the size of the test is the probability of observing a board in the critical set. However, it is possible to generalize the above test by defining a test statistic $t(\cdot)$ defined

| tooth | | | | | | | |
|---|---|---|---|---|---|---|---|
| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ | total |
| 1 | 5 | - | - | - | - | - | 6 |
| - | 2 | 4 | - | - | - | - | 6 |
| - | - | 3 | 8 | - | - | - | 11 |
| - | - | - | 3 | 7 | - | - | 10 |
| - | - | - | - | 5 | 6 | - | 11 |
| - | - | - | - | - | 5 | 7 | 12 |
| 6 | - | - | - | - | - | 4 | 10 |
| 7 | 7 | 7 | 11 | 12 | 11 | 11 | 68 |

Table 7: Engineering quality control results (simplified) for a gear with seven teeth; dataset `gear` in the package. Each tooth may be compared subjectively with the two adjacent teeth and the numbers indicate the number of times each one is judged to be the more heavily worn. With fixed row and column totals, the board possesses one degree of freedom, although in this case a two-sided test is appropriate because there is no prior reason to favour a clockwise bias over an anticlockwise bias

on permissible boards, and considering instead a critical set comprising permissible boards $x$ with $t(x)$ not exceeding that of the observation: $\{x : t(x) \geqslant t\left(x_{\mathrm{obs}}\right)\}$. This approach leads naturally to a number of interesting and useful tests on tables with structural zeros.

The special case of a cyclic competition graph occurs naturally in a variety of contexts; this allows one-sided tests, and the form of the board immediately suggests a generalization of the odds ratio, which has a straightforward maximum likelihood estimate.

We provide software for carrying out these statistical tests in the form of **aylmer**, an R package that includes `aylmer.test()`, a drop-in replacement for the `fisher.test()` function that can accommodate `NA` entries representing structural zeros.

## Acknowledgements

We acknowledge the many stimulating and helpful comments made by the R-help list; and we thank A. M. Hankin for a sequence of ever-more demanding test cases which repeatedly uncovered bugs in earlier versions of our software.

Function `aylmer.function()` is the result of a suggestion by an anonymous JSS referee.

# References

Agresti A (2002). *Categorical Data Analysis.* Wiley, second edition.

Alway GG (1962). "The Distribution of the Number of Circular Triads in Paired Comparisons." *Biometrika*, **49**(1/2), 265–269.

Aoki S, Takemura A (2005). "Markov Chain Monte Carlo Exact Tests for Incomplete Two-Way Contingency Tables." *Journal of Statistical Computation and Simulation*, **75**(10), 787–812.

Berkson J (1978). "In Dispraise of Fisher's Exact Test: Do the Marginal Totals of the 2 × 2 Table Contain Relevant Information Respecting the Table Proportions?" *Journal of Statistical Planning and Inference*, **2**, 27–42.

Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bollobás B (1979). *Graph Theory: An Introductory Course*. Springer.

Bradley RA (1954). "Incomplete Block Rank Analysis: On the Appropriateness of the Model for a Method of Paired Comparisons." *Biometrics*, **10**(3), 375–390.

Bradley RA, Terry ME (1952). "The Rank Analysis of Incomplete Block Designs I. The Method of Paired Comparisons." *Biometrika*, **39**, 324–345.

Colgan PW, Smith JT (1985). "Experimental Analysis of Food Preference Transitivity in Fish." *Biometrics*, **41**, 227–236.

Connor GR, Grant CP (2000). "An Extension of Zermelo's Model for Ranking by Paired Comparisons." *European Journal of Applied Mathematics*, **11**, 225–247.

Davidson RR, Farquhar PH (1976). "A Bibliography on the Method of Paired Comparisons." *Biometrics*, **32**(2), 241–252.

Fisher RA (1954). *Statistical Methods for Research Workers*. Oliver and Boyd.

Frean M, Abraham ER (2001). "Rock-Scissors-Paper and the Survival of the Weakest." *Biological Sciences*, **268**(1474), 1323–1327.

Freeman GH, Halton JH (1951). "Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance." *Biometrika*, **38**(1-2), 141–149.

Ghent AW (1972). "A Method for Exact Testing of 2 × 2, 2 × 3, 3 × 3, and Other Contingency Tables, Employing Binomial Coefficients." *American Midland Naturalist*, **88**(1), 15–27.

Good IJ (1976). "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables." *The Annals of Statistics*, **4**(6), 1159–1189.

Hankin AGS (2007a). Personal Communication.

Hankin RKS (2007b). "Urn Sampling Without Replacement: Enumerative Combinatorics in R." *Journal of Statistical Software, Code Snippets*, **17**(1).

Hankin RKS (2007c). "Very Large Numbers in R: Introducing Package **Brobdingnag**." *R News*, **3**(3), 15–16. URL http://CRAN.R-project.org/doc/Rnews/.

Hankin RKS (2008a). ***Hyperdirichlet****: A Generalization of the Dirichlet Distribution*. R package version 1.0-2; paper submitted to Journal of Statistical Software and currently under review.

Hankin RKS (2008b). "Programmers' Niche: Multivariate Polynomials in R." *R News*, **8**(1), 41–45.

Howard JV (1998). "The 2 × 2 Table: A Discussion from a Bayesian Viewpoint." *Statistical Science*, **13**(4), 351–367.

Jech T (1983). "The Ranking of Incomplete Tournaments: A Mathematician's Guide to Popular Sports." *The American Mathematical Monthly*, **90**(4), 246–266.

Kendall MG, Babington Smith B (1940). "On the Method of Paired Comparisons." *Biometrika*, **31**(3–4), 324–345.

Kirkpatrick M, Rand AS, Ryan MJ (2006). "Mate Choice Rules in Animals." *Animal Behaviour*, **71**, 1215–1225.

Knezek G, Wallace S, Dunn-Rankin P (1998). "Accuracy of Kendall's Chi-Square Approximation to Circular Triad Distributions." *Psychometrika*, **63**(1), 23–34.

Lehmann EL (1993). "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, **88**(424), 1242–1249.

Lorenzoni I, Pidgeon N (2005). "Defining Dangers of Climate Change and Individual Behaviour: Closing the Gap." In "Avoiding Dangerous Climate Change," UK Met Office. Exeter, 1-3 February.

Metropolis NA, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953). "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics*, **21**, 1087–1092.

O'Neill S (2008). *An Iconic Approach to Communicating Climate Change*. Ph.D. thesis, School of Environmental Science, University of East Anglia.

Raymond M, Rousset F (1995). "An Exact Test for Population Differentiation." *Evolution*, **49**(6), 1280–1283.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ryan MJ, Rand AS (2003). "Sexual Selection in Female Perceptual Space: How Female Túngara Frogs Perceive and Respond to Complex Population Variation in Acoustic Mating Signals." *Evolution*, **57**(11), 2608–2618.

Semmann D, Krambeck HJ, Milinski M (2003). "Volunteering Leads to Rock-Paper-Scissors Dynamics in a Public Goods Game." *Nature*, **425**, 390–92.

Silvapulle MJ, Sen PK (2005). *Constrained Statistical Inference*. Wiley.

Sinervo B, Lively CM (1996). "The Rock-Paper-Scissors Game and the Evolution of Alternative Male Strategies." *Nature*, **380**, 240–243.

Szabó G, Fáth G (2007). "Evolutionary Games on Graphs." *Physics Reports*, **446**, 97–216.

Waite TA (2001). "Intransitive Preferences in Hoarding Gray Jays (*Perisoreus Canadensis*)." *Behavioral Ecology and Sociobiology*, **50**, 116–121.

White HC (1963). *An Anatomy of Kinship*. Prentice-Hall.

Wikipedia (2007). “Rock, Paper, Scissors — Wikipedia, The Free Encyclopedia.” [Online; accessed 14-September-2007], URL `http://en.wikipedia.org/w/index.php?title=Rock%2C_Paper%2C_Scissors&oldid=157766515`.

Zermelo E (1929). “Die Berechnung der Turnier-Ergebnisse als ein Maximum-problem der Wahrscheinlichkeitsrechnung.” *Math Z*, **29**, 436–460.

## Affiliation:

Luke J. West Robin K. S. Hankin
Cambridge Centre for Climate Change Mitigation
University of Cambridge
19 Silver Street
Cambridge CB3 9EP
United Kingdom
E-mail: `rksh1@cam.ac.uk`
URL: `http://www.landecon.cam.ac.uk/staff/profiles/rhankin.htm`