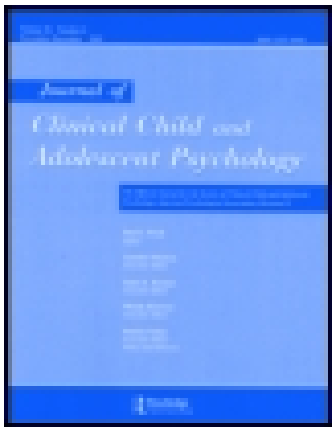

Journal of Clinical Child & Adolescent Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hcap20

Introduction to Permutation and Resampling-Based Hypothesis Tests*

Bonnie J. LaFleur [a] & Robert A. Greevy [b]

[a] Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health , University of Arizona ,

[b] Department of Biostatistics , Vanderbilt University Medical Center ,

Published online: 12 Mar 2009.

To cite this article: Bonnie J. LaFleur & Robert A. Greevy (2009) Introduction to Permutation and Resampling-Based Hypothesis Tests*, Journal of Clinical Child & Adolescent Psychology, 38:2, 286-294, DOI: 10.1080/15374410902740411

To link to this article: http://dx.doi.org/10.1080/15374410902740411

Journal of Clinical Child & Adolescent Psychology, 38(2), 286–294, 2009

ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374410902740411

Routledge
Taylor & Francis Group

METHODOLOGICAL ARTICLE

Introduction to Permutation and Resampling-Based Hypothesis Tests

Bonnie J. LaFleur
Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona

Robert A. Greevy
Department of Biostatistics, Vanderbilt University Medical Center

A resampling-based method of inference—permutation tests—is often used when distributional assumptions are questionable or unmet. Not only are these methods useful for obvious departures from parametric assumptions (e.g., normality) and small sample sizes, but they are also more robust than their parametric counterparts in the presences of outliers and missing data, problems that are often found in clinical child and adolescent psychology research. These methods are increasingly found in statistical software programs, making their use more feasible. In this article, we use an application-based approach to provide a brief tutorial on permutation testing. We present some historical perspectives, describe how the tests are formulated, and provide examples of common and specific situations under which the methods are most useful. Finally, we demonstrate the utility of these methods to clinical and adolescent psychology by examining four recent articles employing these methods.

HISTORY AND BACKGROUND

There is a renewed interest in using distribution-free methods for making parametric inferences based solely on the principle of permutation (sometimes called randomization or re-randomization tests). This methodology, examined early by Fisher (1936), Pitman (1937a–c), and Kempthorne (1952), is used in multiple ways. For example, it is frequently used to support the validity of normal theory results (e.g., Fisher's argument in the 1930's was one in support of the Student's t-test). Historically, applied statisticians have revisited and extended these methods in many contexts. Zerbe (1979) and Raz (1989) extended Kempthorne's work for growth curve analysis. Draper and Stoneman (1966) employed the randomization method for the special case of a multiple linear regression. In the context of multiple regression, Kennedy (1995) and Kennedy and Cade (1996) gave thorough summaries of several methods that may be used to conduct randomized tests.

One of the most compelling reasons to use permutation methods for inference is the robust nature of these methods. Not only do permutation test statistics have relatively weak assumptions they are also more robust than their parametric counterparts when faced with typical challenges of experimental data (e.g., outliers or extreme distributions). Statistical tests that rely on summary statistics (e.g., the mean) can be unduly influenced by the presence of outliers. Since permutation tests are based on test statistics obtained for the observed data relative to test statistics of permutations of the data, the influence of extreme data points is mitigated. Parametric statistical inference relies on assumptions that are justified by taking a random sample from an infinite population. Violating this principle invalidates the use of parametric test statistics and their subsequent inference although the permutation test can

Correspondence should be addressed to Bonnie J. LaFleur, 1295 N. Martin Avenue, PO Box 245211, Tuscon, AZ 85724-5163. E-mail: blafleur@email.arizona.edu

still be applied. Most research in child and adolescent psychology is not done on true random samples from an infinite population. Instead, most samples are drawn from clinics, schools or other populations that are assessable to investigators. Frequently, assumptions required for parametric hypothesis testing are unmet or questionable, and while there are many reasons that permutation tests are attractive, flexible assumptions and their robust nature in these circumstances are the most appealing.

Permutation tests do have assumptions. The primary assumption underlying permutation tests is "exchangeability" of errors. Exchangeability assumes that, under the null hypothesis, the labels in an experiment (e.g., subject identification with respect to experimental condition) do not influence the outcome of the experiment. This means that if the subject identification labels, for example, were randomly placed on the observed data, the results would not change. Additionally, permutation tests do assume that the underlying distributions are symmetric and primarily are designed to test shifts (e.g., difference in means).

Critics of permutation or randomization tests state that one of the drawbacks is that the permutation distribution is the sampling distribution and inference can only be made about the sample at hand, not generalized to a larger population (Koch, 1988). Many proponents of permutation or randomization tests, including Manly (1997), argue that realistically this is a drawback of all statistical methodology that relies on the existence of a larger, perhaps unknown, sampled distribution. Additionally, there are many who believe these tests should only be used with randomized experiments. Kempthorne's (1952) work dealt mainly with analysis of variance and focused on permuting subjects to positions based on treatment randomization. Permutation tests are sometimes more conservative than their parametric equivalent test statistics. This is in part due to the discrete nature of the permutational p-values. While it is always possible that the permutation test is not the most powerful test, in which case the most powerful test will be preferred, this is not specific only to permutation tests but all statistical inferential procedures. The ideas generated by Fisher (1935) and described by Pitman (1937a–c) continue to be source of theoretical discussion. Interested readers can find thorough and understandable clarifications in books by Edgington (1995), Manly (1997), and Lunneborg (2000). Berger (2000) has a very nice discussion on the use of permutation tests specific to clinical trials. Good's (2004) book also contains excellent descriptions and details of permutation tests, as well as a detailed bibliography.

Our tutorial in this article focuses on methods of inference that are available using standard software and provides examples of the best instances to use these types of hypothesis testing. We do not delve into theoretic underpinnings unless necessitated by the examples we provide. There is a relationship to methods employing permutations and rank-based methods, such as the rank sum test statistic. However, since they are not based on random permutations of sample data, they are not described here.

DEFINITIONS

Permutation tests are considered a special case of nonparametric tests. Nonparametric test statistics do not rely on a specific probability distribution (e.g., normal, chi-square, binomial) that describes the underlying population. However, permutation tests are not quite "distribution free." Some underlying assumptions are required with respect to the samples (e.g., exchangeability). Permutation tests are sometimes called randomization (or rerandomization) tests and may be used interchangeably by some. Kempthorne (1986) states that the fundamental difference between the two is that permutation tests are based on random sampling. Randomization, or rerandomization, tests are based on a sample that has been randomized *a priori* (before data collection). Edgington (1995) discusses randomization tests as special types of permutation tests and notes that the rationale is different. These discussions also are found throughout Kempthorne's work (1955, 1966, 1972, 1975). We have chosen to use the terms loosely in this tutorial, although the examples may or may not be from a distribution that has been randomized *a priori*.

Permutation tests also are called resampling tests, a subset of nonparametric statistics. Statistical inference depends upon examining random samples of observations from a particular population. Resampling-based methods work within the principle of resampling from a sample that may or may not be a random sample. Resamples are used as the "data" for inference. Generating p-values from a permutation test is easy to implement. Data permutation is one of the most complicated processes, along with ensuring randomness when a random sample of permutations is used.

Permutation tests proceed as follows: (1) data from an experiment are tested using some pre-specified test statistic, (2) the test statistic is generated for the original sample of the data (sometimes called the observed permutation), and (3) the results are saved. Data permutations are then enumerated. The permutations often are referred to as *the physical act of permuting subjects to labels*. The permutations either can be all n! permutations, the number of conditional permutations based on experimental design, or a random sample of all possible permutations. A test statistic is then calculated for each of the permutations and compared against the test statistic based on the original data. The permutational p-value is calculated by the following: the number of times the test statistics from the permuted data are equal

to or more extreme (larger) than the original test statistic divided by the total number of permutations examined. These test statistics can be based on distributions (e.g., a t-test statistic in the case of a two-sample test with continuous measurements) or based on some other defined statistic (e.g., the value of 1,1 cell in a 2 by 2 table is often used in Fisher's exact test).

Even with the immense power available in today's personal computers, enumerating all possible permutations for even a modest size dataset remains a daunting task. Edgington's (1980) rationale proves that employing a random sample of possible permutations is valid. Based on the work by Dwass (1957), Manly (1997) notes that permutation tests based on a random sample of permutations is still "exact." However, random permutations will be less powerful than all possible permutations, and increasing the number of random permutations will increase power. Manley suggests that a minimum of 1,000 permutations are desired for tests to result in a 5% level of significance. As few as 200 permutations may be sufficient as demonstrated in a recent paper by Fitzmaurice and Lipsitz (2007). In our experience, 10,000 random permutations give reliable results while reducing computing time considerably compared to evaluation of all data permutations. This is particularly true for complicated models (i.e., many predictors), as these models can still require nontrivial computing time to process.

Resampling statistics also include the bootstrap (sampling with replacement) and the jackknife (leave-one-out) methods. Traditionally, the jackknife has been used to reduce bias in small samples, calculate confidence intervals around parameter estimates, and to test hypotheses (Manly, 1997; Tukey, 1958). Bootstrap methods have long been used to estimate standard errors in cases where the distribution of the data is not known, and are often used to construct confidence intervals around parameter estimates. Efron's and Tibshirani's (1993) text describes bootstrap resampling. Other reviews can be found in Manly (1993) and Davison and Hinkely (2003). In most cases, permutation testing is more powerful than the bootstrap approach (and perhaps the jackknife), although Good (2000) considers some conditions under which the bootstrap may be more powerful. Westfall and Young (1993, Chapter 5) show that the difference in reported p-values between using bootstrap resampling and permutation resampling is quite small in most examples. Bootstrap and permutation resampling almost always result in the same inferential interpretation (i.e., reject or not reject the null hypothesis). The bootstrap approach, using confidence intervals for hypothesis testing, will work in some situations where the permutation testing approach will not. For example, neither the parametric nor the permutation-based tests are estimable in some unbalanced ANOVA designs. However, it is possible to calculate bootstrap confidence intervals of the interaction of interest.

WHEN ARE THESE METHODS USED?

Primarily, these tests are used when assumptions for parametric tests cannot be met, experiments with small sample sizes, or when an exact test is desired. Exact tests are those where the significance level of the test is equal to the false rejection rate. If all distributional assumptions are met, the parametric tests are exact. Permutation tests always calculate exact significance levels when looking at all data permutations. Deviations from the exact significance level will occur when the exchangeability assumption is not met. Generally, significance levels (using bootstrap or jackknife sampling) are not exact. Permutation tests also are as powerful as the unbiased parametric test for small sample sizes (Good, 2000).

Another advantage of permutation tests is that inference can be made in cases when analysis is hampered by computational difficulties. For example, when there is collinearity in the data that results in separation of data points or structural zeros, and also in sparse datasets. This will be discussed further in the logistic regression example. It is debatable whether these methods help in cases where the assumption of unequal variances is questioned or violated. For instance, if the study is a comparison of groups from a generalizable random sample and the question is whether or not these groups have different means, the permutation tests are as vulnerable to unequal variances as their normal theory counterparts. However, if the sample is a randomized, controlled trial and inference is limited to the randomized sample and exchangeability is the only required assumption; unequal variances are not an issue. Viewed in this light, permutation tests can be tailored to include many sampling schemes (random or not) and the statistical tests will be viable, and perhaps exact, for most practical problems.

EXAMPLES

Two Group Example Using Student's t-Test Statistic

For a simple, illustrative example imagine a random sample of fifth grade girls, three from a traditional coed school and three from an experimental school that offers single-gender classes. The unrealistically small sample size of six girls is used so the reader can see the permutation test completely illustrated in Table 1.

Each of the girls is given the Harter Self-Perception Profile for Children (Harter, 1985) this 36 item profile measures of six domains of self-perception, but for simplicity will be presented as a summary score ranging from 0–36.

TABLE 1
All Possible Permutations of Six Students' Self-perception Scores

Permutation Number	*Traditional School (T)*			*Experimental School (E)*			*Mean T*	*Mean E*	*Difference (E-T)*
1–36	10	11	12	13	14	28	11.0	18.3	7.3
37–72	10	11	13	12	14	28	11.3	18.0	6.7
73–108	10	11	14	12	13	28	11.7	17.7	6.0
109–144	10	12	13	11	14	28	11.7	17.7	6.0
145–180	10	12	14	11	13	28	12.0	17.3	5.3
181–216	11	12	13	10	14	28	12.0	17.3	5.3
217–252	11	12	14	10	13	28	12.3	17.0	4.7
253–288	10	13	14	11	12	28	12.3	17.0	4.7
289–324	11	13	14	10	12	28	12.7	16.7	4.0
325–360	12	13	14	10	11	28	13.0	16.3	3.3
361–396	10	11	28	12	13	14	16.3	13.0	−3.3
397–432	10	12	28	11	13	14	16.7	12.7	−4.0
433–468	10	13	28	11	12	14	17.0	12.3	−4.7
469–504	11	12	28	10	13	14	17.0	12.3	−4.7
505–540	11	13	28	10	12	14	17.3	12.0	−5.3
541–576	10	14	28	11	12	13	17.3	12.0	−5.3
577–612	12	13	28	10	11	14	17.7	11.7	−6.0
613–648	11	14	28	10	12	13	17.7	11.7	−6.0
649–684	12	14	28	10	11	13	18.0	11.3	−6.7
685–720	13	14	28	10	11	12	18.3	11.0	−7.3

The self-perception scores for the girls in the traditional (T) co-ed school were as follows: child 1 = 12, child 2 = 10, child 3 = 11; and for the experimental (E) single-gender school: child 1 = 14, child 2 = 13, and child 3 = 28. The mean self-perception score for the experimental school is 18.3 and for the traditional school is 11.0, yielding a difference in means of 7.3. For a parametric analysis, a standard two-sample t-test gives a one-sided p-value of 0.10. Thus, we would not reject the null hypothesis of no difference in self-perception between the schools at a 0.05 significance level. The standard two-sample t-test (ANOVA) assumes the two populations from which the girls were drawn both had normally distributed self-perception scores and that the variances of those two populations are the same. While these assumptions can never be proven, one typically uses the sample data to comment on the plausibility of those assumptions. With small sample sizes, the plausibility can be difficult to determine.

Now consider doing a permutation test instead. There are 6!, i.e., $6 * 5 * 4 * 3 * 2 * 1 = 720$, ways to permute the six students' self-perception scores and 20 unique permutations of the data (6!/3!3!, since there are 3 in each of the 2 groups). The numerator is based on the total sample size and the denominator is from the number of ways that the data can be partitioned into the number of groups or categories in the study. This example has 2 groups of 3 subjects in each group. A discussion about permutations and combinations of data can be found in introductory mathematical statistics textbooks, such as Bain and Engelhardt (1987). Based on these partitions, there are 36 ways we can find a difference between the means of 7.3 (simply rearrange the observed scores within each group and take the difference). Table 1 lists the 20 possible differences in means for the 720 different permutations. Permutation numbers are groups of permutations that result in the same data (e.g., permutation numbers 1–36 all had self-perception scores of 10, 11, and 12 for traditional school and scores of 13, 14, and 28 for the experimental school). In this example we are using the difference in means as the "test statistic," we could have used the calculated t- or F-test statistic as our test statistic with the same conclusion and resultant p-value.

The set of permutations that include the observed data are shown in the first row of Table 1. The p-value for a permutation test is the proportion of permutations that yield a value as extreme (equal to) or more extreme (greater than) compared to the observed value. For these differences, "extreme" means in the direction of, or strongly favoring, the alternative hypothesis. The null hypothesis is that there is no difference in self-perception between schools, and the alternative hypothesis is that the self-perception is higher for children in the experimental school (i.e., the difference of mean E–mean T is greater than zero). For the observed permutation the difference in means is 7.3. There are 36 permutations as extreme as this and no permutations, which yield values more extreme, yielding a p-value for the permutation test of $36/720 = 0.05$. Thus, the permutation test would reject the null hypothesis at a 0.05 significance level in favor of the alternative that girls attending the experimental school have higher self-perception scores. This small, illustrative dataset is useful in allowing the reader to see all the possible permutations of the data in a simplified format. However, it is worth noting that

an equally divided sample of size six could never detect a significant difference in a two-sided test on the difference in means. For a two-sided hypothesis of no difference between the self-perception scores the smallest possible p-value we could obtain would be $72/720 = 0.10$.

Creating Confidence Intervals

The previous example tested whether girls in a traditional school had the same mean self-perception score as girls in an experimental single-gendered school (i.e., it tested the null hypothesis *Ho*: $\mu_E - \mu_T = 0$–. What about testing the hypothesis that on average girls in the experimental school score one and half points higher (i.e., *Ho*: $\mu_E - \mu_T = 1.5$)? What about calculating a plausible range of values for the difference in the mean scores, μ_E-μ_T? In this example, these two questions are related. Using the previous example, let's suppose we want to create a 90% confidence interval (CI) for the difference in the mean self-perception scores between girls in the experimental and traditional schools. We are using a 90% CI instead of the more common 95% because the example dataset is so small. The simplest way to create a confidence interval using a permutation test is to imagine doing many two-sided tests at some significance level to test many null hypotheses (e.g., the true difference in mean scores is 1, . . . is 1.5, . . . is 2.1, etc.) Then use the range of values that were not rejected in the hypothesis test as the confidence interval. You can call this CI the *acceptance region of your test*. To create a 90% CI, we will use a 0.10 significance level, as opposed to a 0.05 level used for a 95% CI.

Let's consider testing the null hypothesis that the true difference between the mean scores is 1.5 (i.e., the self-perception scores for girls in the experimental school are on average 1.5 points higher). We test this by subtracting 1.5 from each of the observed scores for the girls in the experimental schools and apply our permutation test to the shifted dataset. Table 2 shows the resulting data after this shift.

The set of permutations that include the observed data are labeled permutations 1–36 and shown in bold. The shifted observed data have a difference in means of 5.8. There are 144 permutations rated as extreme or more extreme (± 5.8 and ± 6.2). The p-value for the permutation test on the shifted data is $144/720 = 0.20$ and would not be rejected at a 0.10 level. Thus, the hypothesized difference in mean scores of 1.5 belongs in the interval. An examination of Table 1 will reveal that values down to 1.00 also would be included in the interval, and values of 0.99 or less would not. Thus, 1.00 would be the lower bound of the CI. Similarly, 18.00 would be included in the interval, but 18.01 would not. Hence, the 90% CI is [1.00, 18.00]. Although the construction of these acceptance regions may be more complex with larger datasets and different data types, advanced software packages have sophisticated mechanisms to create the confidence intervals.

Three Group Example Using Anova

The previous example looked at the simplest case of comparing the means of two groups. What if there are three groups? The following example is taken from the

TABLE 2
Shifted Behavior Scores for Testing the Hypothesis Ho: $\mu_E - \mu_T = 1.5$

Permutation Number	*Traditional School (T)*			*Experimental School (E)*			*Mean T*	*Mean E*	*Difference (E−T)*
1–36	10	11	12	11.5	12.5	26.5	11.0	16.8	5.8
37–72	10	11	11.5	12	12.5	26.5	10.8	17.0	6.2
73–108	10	11	12.5	12	11.5	26.5	11.2	16.7	5.5
109–144	10	12	11.5	11	12.5	26.5	11.2	16.7	5.5
145–180	10	12	12.5	11	11.5	26.5	11.5	16.3	4.8
181–216	11	12	11.5	10	12.5	26.5	11.5	16.3	4.8
217–252	11	12	12.5	10	11.5	26.5	11.8	16.0	4.2
253–288	10	11.5	12.5	11	12	26.5	11.3	16.5	5.2
289–324	11	11.5	12.5	10	12	26.5	11.7	16.2	4.5
325–360	12	11.5	12.5	10	11	26.5	12.0	15.8	3.8
361–396	10	11	26.5	12	11.5	12.5	15.8	12.0	−3.8
397–432	10	12	26.5	11	11.5	12.5	16.2	11.7	−4.5
433–468	10	11.5	26.5	11	12	12.5	16.0	11.8	−4.2
469–504	11	12	26.5	10	11.5	12.5	16.5	11.3	−5.2
505–540	11	11.5	26.5	10	12	12.5	16.3	11.5	−4.8
541–576	10	12.5	26.5	11	12	11.5	16.3	11.5	−4.8
577–612	12	11.5	26.5	10	11	12.5	16.7	11.2	−5.5
613–648	11	12.5	26.5	10	12	11.5	16.7	11.2	−5.5
649–684	12	12.5	26.5	10	11	11.5	17.0	10.8	−6.2
685–720	11.5	12.5	26.5	10	11	12	16.8	11.0	−5.8

TABLE 3
Authoritarianism Scores of Three Groups of Educators (Artificial Data)

Teaching-oriented teachers (Group 1)	*Administration-oriented Teachers (Group 2)*	*Administrators (Group 3)*
96	82	115
128	124	149
83	132	166
61	135	147
101	109	

textbook by Siegel (1956, Table 8.5). It is a hypothetical dataset that examines the hypothesis–school administrators are more authoritarian than classroom teachers. The response interest variable is the score on an F scale and measures the degree of authoritarianism. There are 14 subjects divided into 3 groups: (1) teaching-oriented teachers (teachers currently in the classroom who wish to stay in the classroom), (2) administration-oriented teachers (teachers currently in the classroom but aspire to administrative positions), and (3) administrators. Table 3 shows the fictional scores of these groups.

There are 252,252 (14!/(5!5!4!)) possible permutations of these data. Using a random sample of 10,000 permutations, Table 4 shows the permutational p-values for the various hypothesis tests of interest as well as the normal-theory results from an ANOVA.

In Table 4, the population means for the three groups in Table 3 are labeled μ_1, μ_2, and μ_3, respectively. The exact results from the permutation tests are similar to the approximate results from the test that assumes normality. These methods can be applied to more complicated, multifactor experiments. Although the computational needs increase greatly with the number of subjects, as well as the number of parameters of interest, Manly (1997) and Pesario (2001) discuss multivariate permutation models and complicated design structures.

Logistic Regression Example

The primary utility for using permutation based p-values in logistic regression is the same as that for the linear models described above (i.e., parametric based methods can result in incorrect or biased inference because of assumption violations, particularly when there are small sample sizes). Additionally, algorithms that are used to fit nonlinear models are complicated, and can have convergence problems, adding yet another level of utility for permutation methods. Permutation-based tests for these models employ algorithms that work in situations where parametric-based methods do not.

The most well known approach to exact methods for logistic regression is implemented in a commercial software package called LogExact (Cytel, Inc., Cambridge, MA) and has recently been implemented in SAS (SAS Institute Inc., Carey, NC). These software packages implement an algorithm developed by Hirji, Mehta, and Patel (1987) and are based on what is known as exact conditional inference. The comparison between conditional and unconditional inference has been a source of controversy since the introduction of Fisher's exact test. A well-written survey about the use of conditional permutation methods for contingency tables can be found in Agresti (1992). He discusses the 2×2 table where test statistics are constructed for data conditioning on marginal totals. The extension of the 2×2 table approach to binary logistic regression is well described by Cox and Snell (1989).

Under conditional logistic regression (as in the Fisher exact test), structure is placed on the marginal totals. When leaving the row and column totals fixed, the results are the same as permuting results of 2×2 tables (or sets of 2×2 tables when there are more than one predictor variables). Further, this software uses a unique algorithm that speeds up processing when estimating the exact p-values. The LogExact package and SAS implementation of exact conditional logistic regression model these conditional tests exclusively. To use the software, predictor variables must either be categorical, or the sample size must be small enough to impose a categorical structure (e.g., the program will assume that the continuous variables are categorical). As an example we construct a $2 \times 2 \times 2$ contingency table based on the number of subjects that responded to an intervention by sex (β_1) and age (β_2) (see Table 5).

Table 6 shows the results using a logistic regression model under binomial distribution assumptions in

TABLE 4
Anova Results

Hypothesis Tested	*Permutational p-value*	*Likelihood Ratio p-value*
$\mu_1 = \mu_2 = \mu_3 = 0$	0.0275	0.0222
$\mu_1 - \mu_2 = 0$	0.1466	0.1436
$\mu_1 - \mu_3 = 0$	0.0090	0.0069
$\mu_2 - \mu_3 = 0$	0.0952	0.0945

TABLE 5
a $2 \times 2 \times 2$ Contingency Table

	Male		*Female*	
	<5 yrs	*5–10 yrs*	*<5 yrs*	*5–10 yrs*
Responder	0	1	0	2
Non-Responder	1	1	2	0

TABLE 6
Results from the Logistic Regression Model

Hypothesis Tested	*Permutational p-value*	*Likelihood Ratio p-value*[¥]
$\beta_1 = \beta_2 = 0$	0.2857	0.0336
$\beta_1 = 0$	1.000	0.0245
$\beta_2 = 0$	0.1617	0.6576

[¥]From SAS Proc Logistic.

contrast to the permutational p-value based on the conditional model using LogExact. The example shows that the permutational p-values (exact in this case) are different than the p-values that assume an underlying binomial sampling distribution. The discrepancy is due to sparse data (i.e., cells with zero).

This often happens with small sample sizes, or with many categories of data with respect to sample size. Data with many zero cells lead to what is sometimes called ''separation'' or ''quasi-separation'' of data points. In these circumstances the p-values based on the likelihood ratio test are invalid. When there is complete separation of data points, logistic regression programs will generally not converge or give any estimates. As an example, if the data depicted in Table 5 resulted in all responders being male between 5–10 years old (without the 1 nonresponder that was <5 years old) then this would be called complete separation of data points and programs written in commercial statistics packages would not converge. Thus, no results would be reported or there would be a warning that convergence criterion was not met. However, when the data result in quasi-separation, as shown in Table 5, some programs will converge on an incorrect estimate. A number of commercial statistical packages will give a warning message that the estimates and test statistics are not valid. Unfortunately, others do not give warning messages and can lead to incorrect statistical inference. In summary, there are two clear circumstances to apply exact methods for logistic regression: (1) when you have small sample sizes and (2) when there is complete or partial separation of data points. In other situations, approximate methods will result in the same inference as their exact counterparts.

ADVANCED TOPICS

Complications arise when applying permutation methods to multiple regression of any type because the regression framework has fewer model constraints. For instance, if there is a response variable and two covariates, what gets permuted? In a conditional setting, or when there is a structure imposed due the experimental design (ANOVA), permutation is done based on the design. Details about how to use permutation tests in multiple regression models can be found in papers by Anderson and Legendre (1999), Kennedy and Cade (1996), ter Braak (1992), and others. The relationship between test statistics and hypotheses in these unstructured (nonrandomized) models are discussed in LaFleur (1999).

Modern genomic analyses have incorporated permutation tests into the analysis of oligonucleotide microarray data (i.e., gene chip technology), specifically for probe-based tests of significance (e.g., one probe at a time, where a number of probes are associated with genes of interest). One rationale for using permutation tests in these settings is the small numbers of samples used. This is in contrast to the large numbers of genes being tested. The use of these exact tests do not protect from the problem of multiplicity (i.e., the probability of finding false positives). Methods associated with these tests that adjust p-values for this situation are discussed in the context of multiple testing in Westfall and Young (1993). When testing group differences between large numbers of probes individually, exact tests will be protective against normality and other large sample approximation assumptions on a gene-by-gene basis, a primary motivation for using permutation tests. A distinction needs to be made between permutation p-values and permutation adjusted p-values. The latter is a method for adjusting for multiple testing using the minimum p-value for all tests (as discussed in Westfall and Young (1993, 1998). The former is a p-value obtained by the act of physically permuting the data.

An in-depth discussion about the use of permutation methods in logistic regression, and the differences between unconditional and conditional methods, can be found in a recent paper by Potter (2005). He describes the theoretical distinctions between these alternative methods. We refer readers interested in a more technical description to his paper and to the paper by Mehta and Patel (1995). In this tutorial, we focused on conditional methods because they are incorporated into commercial software packages and more accessible. For small sample sizes, Potter (2005) shows that the mid-p-values obtained using conditional methods are equivalent to unconditional p-values. Additional applications of unconditional methods can be found in LaFleur (1999). In this dissertation, unconditional methods are also extended to generalized linear models that include logistic, Poisson, gamma, and other distributions as described by McCullagh and Nelder (1989).

SOFTWARE

Most statistical packages have some permutation or other resampling test procedures. Many, however, are

specific to statistical tests used for categorical or logistic regression models. SAS has a procedure that allows for permutation-based testing for two sample t-tests and ANOVA models called Proc Multtest. Because both permutation p-values and adjusted permutation p-values can be output by the program, users must be careful not to confuse the two. StatExact is a stand alone package that includes a very comprehensive set of exact, permutation and other nonparametric test statistics for categorical data. SAS, SPlus, R, STATA, and SPSS have programs for exact categorical tests. The programs in SAS and SPSS are subsets of the capabilities from StatExact package. SAS and the stand-alone program LogExact can perform permutation tests for logistic regression.

STATA has incorporated some user-defined functions to perform these tests. All of the larger packages—MINITAB, SAS, SPlus, SPSS, STATA, and R—allow for user defined programming, enabling these methods to be implemented easily and efficiently. Another useful package, Resampling Stats, is affiliated with a Website that contains helpful information and tutorials covering multiple types of resampling methods (www.resample.com). Many of the textbooks referenced in this paper have sample programs that can be adapted into these packages (e.g., Lunneborg, 2000).

SUMMARY

We have presented an introduction to applying permutation hypothesis testing, using historical references and examples to clarify the process. Permutation hypothesis testing is optimal when distribution-based hypothesis test assumptions are questioned. Although there are arguments against their use, permutation-based hypothesis tests are exact and as powerful as their parametric counterparts when distributional assumptions are not met. In some cases, the bootstrap and jackknife resampling tests are more powerful. However, they are not exact as are the permutation tests.

A simple, and non-exhaustive, literature review offers four examples of when these methods might be used in practice. First is a study that presents permutational p-values for two-sample t-tests (Gambea, 2005). This study examined whether or not 13- and 14-month-old infants can comprehend references about missing objects. The investigators questioned whether the normality assumption of the scoring outcome was valid. They therefore presented permutational p-values based on the t-test in the article. The other three articles (i.e., Bolton, Park, Higgins, Griffiths, & Pickles, 2002; Busch et al., 2002; Chadwick, Taylor, Heptinstall, & Danckaerts, 1999) used conditional logistic regression because data points were partially or completely separated. Exact logistic regression is one of the most common permutation approaches (outside of Fisher's exact test) seen in the literature. It is readily available in commercial statistical software programs and applied primarily when standard logistic regression will not converge or produces suspicious output (e.g., very large standard errors). There are many other cases where a permutation approach is warranted and the authors encourage the use of these tests wherever practical, particularly when model assumptions may not be valid.

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *7*, 131–177.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1–10.

Anderson, M. J., & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, *62*, 271–303.

Bain, L. J., & Engelhardt, M. (1987). *Introduction to probability and mathematical statistics* (2nd ed.). Boston: PWS-Kent.

Basu, D. (1980). Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, *75*, 575–595.

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

Bolton, P. F., Park, R. J., Higgins, J. N., Griffiths, P. D., & Pickles, A. (2002). Neuro-epileptic determinants of autism spectrum disorders in tuberous sclerosis complex. *Brain*, *6*, 1247–1255.

Bruce, P., Simon, S., & Oswald, T. (1995). *Resampling stats user's guide*. Arlington: Resampling Stats, Inc.

Busch, B., Biederman, J., Cohen, L. G., Sayer, J. M., Monuteaux, M. C., Mick, E., Zallen, B., & Faraone, S. V. (2002). Correlates of ADHD among children in pediatric and psychiatric clinics. *Psychiatric Services*, *53*, 1103–1111.

Chadwick, O., Taylor, E., Taylor, A., Heptinstall, E., & Danckaerts, M. (1999). Hyperactivity and reading disability: A longitudinal study of the nature of the association. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*, 1039–1050.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). Boca Raton: Chapman Hall.

Draper, N. R., & Stoneman, D. M. (1966a). Errata: Testing for the inclusion of variables in linear regression by a randomization technique. *Technometrics*, *11*, 627.

Draper, N. R., & Stoneman, D. M. (1966b). Testing for the inclusion of variables in linear regression by a randomization technique. *Technometrics*, *8*, 695–698.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187.

Edgington, E. S. (1995). *Randomization tests*. New York: Marcel Dekker.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman Hall.

Fisher, R. A. (1935). *Design of Experiments* (Ist ed.). New York: Oliver Boyd.

Ganea, P. A. (2005). Contextual factors affect absent reference comprehension in 14-month-olds. *Child Development*, *76*, 989–998.

Good, P. (2004). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer-Verlag.

Harter, S. (1985). *The self-perception profile for children: Revision of the perceived Competence scale for children*. Denver, CO: University of Denver.

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*, 851–853.

Hinkelmann, K., & Kempthorne, O. (1994). *Design and analysis of experiments*. New York: Wiley.

Hirji, K. F., Mehta, C. R., & Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, *82*, 1110–1117.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, *40*, 633–643.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, *50*, 946–967.

Kempthrone, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association*, *61*, 11–34.

Kempthrone, O. (1972). Theories of inference and data analysis. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor* (pp. 167–191). Ames: Iowa State University Press.

Kempthrone, O. (1975). Inference from experiments and randomization. In J. N. Srivastava (Ed.), *A survey of statistical design and linear models* (pp. 303–331).

Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika*, *56*, 231–248.

Kennedy, P. E. (1995). Randomization tests in econometrics. *Journal of Business and Economic Statistics*, *13*, 85–94.

Kennedy, P. E., & Cade, B. S. (1996). Randomization tests for multiple regression. *Communications in Statistics*, *25*, 937–952.

Koch, G. G., & Edwards, S. (1988). Clinical efficacy trials with categorical data. In K. E. Peace (Ed.), *Biopharmaceutical statistics for drug development* (pp. 403–57). New York: Marcel-Dekker.

LaFleur, B. J. (1999) Application of permutation methods to the generalized linear model. Ph.D. dissertation, University of Colorado Health Sciences Center, United States, Colorado. (Publication No. AAT 9951494).

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. California: Duxbury Press.

Manly, B. F. J. (1997). *Randomization and monte carlo methods in biology* (2nd ed.). London: Chapman Hall.

MathSoft (1997). *S-Plus User's Guide*. MathSoft, Inc.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, *11*, 59–67.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman Hall.

Mehta, C. R., & Patel, N. (1993a). *LogXact for Windows*. Cambridge: Cytel Software Corporation.

Mehta, C. R., & Patel, N. (1993b). *StatXact for Windows*. Cambridge: Cytel Software Corporation.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, *14*, 2143–2160.

Pesarin, F. (2001). *Multivariate permutation tests*. Chichester: Wiley.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society*, *Supplement 4*, 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any population. II the correlation test. *Journal of the Royal Statistical Society*, *Supplement 4*, 225–232.

Pitman, E. J. G. (1937c). Significance tests which may be applied to samples from any population. III the analysis of variance test. *Biometrika*, *29*, 322–335.

Potter, D. M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine*, *24*, 693–708.

Raz, J. (1989). Analysis of repeated measurements using non-parametric smoothers and randomization tests. *Biometrics*, *45*, 851–871.

Zerbe, G. O. (1979). Randomization analysis of the completely randomized design extended to growth curves. *Journal of the American Statistical Association*, *74*, 215–221.