Chapter 6
Two-Sample Tests

Abstract This chapter introduces permutation methods for two-sample tests. Included in this chapter are six example analyses illustrating computation of exact permutation probability values for two-sample tests, calculation of measures of effect size for two-sample tests, the effect of extreme values on conventional and permutation two-sample tests, exact and Monte Carlo permutation procedures for two-sample tests, application of permutation methods to two-sample rank-score data, and analysis of two-sample multivariate data. Included in this chapter are permutation versions of Student's two-sample t test, the Wilcoxon–Mann–Whitney two-sample rank-sum test, Hotelling's multivariate T^2 test for two independent samples, and a permutation-based alternative for the four conventional measures of effect size for two-sample tests: Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$.

This chapter presents exact and Monte Carlo permutation statistical methods for two-sample tests. Two-sample tests for experimental differences are of primary importance in basic research, whether that be in the behavioral, medical, biological, agricultural, or physical sciences. Statistical tests for differences between two samples are of two varieties. The first of the two varieties examines two sets of data obtained from two completely separate (independent) samples of subjects. For example, a study might seek to compare grades in an elementary statistics course for majors and non-majors, for female and male students, for transfer and non-transfer students, or for juniors and seniors. In a true experimental design with two independent samples a large pool of subjects is randomly assigned (randomized) to the treatments using a fair coin or a pseudo-random number generator.[1] More often than not, however, it is not possible to randomize subjects to treatments, especially in survey research. For example, it is not possible to randomly assign subjects to such

[1]For two treatments a fair coin works quite well with heads and tails. For three treatments, a fair die is often used with faces with one or two pips assigned to the first treatment, faces with 3 or 4 pips assigned to the second treatment, and faces with 5 or 6 pips assigned to the third treatment. For four treatments, a shuffled deck of cards works well with clubs (♣), diamonds (♢), hearts (♡), and spades (♠) assigned to Treatments 1, 2, 3, and 4, respectively.

K. J. Berry et al., *A Primer of Permutation Statistical Methods*,
https://doi.org/10.1007/978-3-030-20933-9_6

categories as gender, age, IQ, or educational level. The lack of random assignment to treatments can greatly compromise the results of two-sample tests.

The second variety of two-sample tests examines two sets of data obtained on the same or matched subjects. For example, a study might compare the same subjects at two different time periods, such as before and after an intervention, or matched subjects on two different diets: low- and high-carbohydrate. When compared with tests for two independent samples, matched-pairs tests generally have less variability between the two samples, provide more power with the same number of subjects, and because the sample sizes are the same for both treatments, matched-pairs tests produce larger test statistic values and smaller probability values than comparable tests for two independent samples, other factors being equal. Two-sample tests for independent samples are presented in this chapter. Matched-pairs tests for two related samples are presented in Chap. 7.[2]

6.1 Introduction

In this chapter permutation statistical methods for two-sample tests are illustrated with six example analyses. The first example utilizes a small set of data to illustrate the computation of exact permutation methods for two independent samples, wherein the permutation test statistic, δ, is developed and compared with Student's conventional t test for two independent samples. The second example develops a permutation-based measure of effect size as a chance-corrected alternative to the four conventional measures of effect size for two-sample tests: Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$. The third example compares permutation methods based on ordinary and squared Euclidean scaling functions, emphasizing methods of analysis for data sets containing extreme values. The fourth example compares and contrasts exact and Monte Carlo permutation methods, demonstrating the accuracy and efficiency of Monte Carlo statistical methods. The fifth example illustrates the application of permutation statistical methods to univariate rank-score data, comparing permutation statistical methods with the conventional Wilcoxon–Mann–Whitney two-sample rank-sum test. The sixth example illustrates the application of permutation statistical methods to multivariate data, comparing permutation statistical methods with the conventional Hotelling's multivariate T^2 test for two independent samples.

One of the most familiar and popular two-sample tests looks at the mean difference between two independent treatment groups. This is the classic test for a difference between a control group and an experimental group. For example, a researcher might want to compare the number of trials on a specified task for two groups of rats—one raised under normal conditions and the other raised in semi-

[2]In some disciplines tests on two independent samples are known as between-subjects tests and tests for two dependent or related samples are known as within-subjects tests.

darkness. Or it might be of interest to examine the differences in performance between two groups of students—one in a section of a course taught in a face-to-face lecture format and the other in a section of the same course taught in an on-line distance-learning format by the same instructor.

The most popular univariate test for two independent samples under the Neyman–Pearson population model of inference is Student's two-sample t test wherein the null hypothesis (H_0) posits no mean difference between the two populations from which the samples are presumed to have been drawn; for example, H_0: $\mu_1 = \mu_2$. Alternatively, H_0: $\mu_1 - \mu_2 = 0$. The test does not determine whether the null hypothesis is true, but only provides the probability that, if the null hypothesis is true, the samples have been drawn from the specified population(s). Student's t test is the standard test for a mean difference between two independent samples and is taught in every introductory course.

6.1.1 The Student Two-Sample t Test

Consider two independent samples of sizes n_1 and n_2. Under the Neyman–Pearson null hypothesis, H_0: $\mu_1 = \mu_2$, Student's t test for two independent samples is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{1/2}} ,$$

where the unbiased pooled estimate of the population variance is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2} ,$$

the sample estimate of the population variance for the ith treatment group is given by

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2 , \qquad i = 1, 2 ,$$

n_i denotes the number of objects in the ith of the two treatment groups,

$$N = \sum_{i=1}^{2} n_i$$

denotes the total number of objects in the two treatment groups, $\bar{x}_i$ denotes the arithmetic mean of the measurement scores for the ith of the two treatment groups, given by

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \ , \qquad i = 1, 2 \ ,$$

and x_{ij} is a measurement score for the jth object in the ith treatment group. Assuming independence, normality, and homogeneity of variance, test statistic t is asymptotically distributed as Student's t under the Neyman–Pearson null hypothesis with $N - 2$ degrees of freedom. The permissible probability of a type I error is denoted by α and if the observed value of t is more extreme than the critical values of $\pm t$ that define α, the null hypothesis is rejected with a probability of type I error equal to or less than α, under the assumptions of normality and homogeneity.

The assumptions underlying Student's t test for two independent samples are (1) the observations are independent, (2) the data are random samples from a well-defined population, (3) homogeneity of variance, that is $\sigma_1^2 = \sigma_2^2$, and (4) the target variable is normally distributed in the population. It should be noted that a number of textbooks have argued that what is important is that the sampling distribution of sample mean differences be normally distributed and not the target variable in the population. However, Student drew his random samples from populations of two sets of measurements on criminal anthropometry that had been published by William Robert Macdonell in *Biometrika* in 1902 [8]. Student's data consisted of two measurements obtained by Macdonell that were approximately normally distributed: (1) the height and (2) the length of the left middle finger of 3000 criminals over 20 years of age and serving sentences in the chief prisons of England and Wales. Moreover, Student proved in Sect. 2 of his 1908 paper that the mean and variance are independent and the normal distribution is the only distribution where this is always true, as noted by George Barnard [1, p. 169].[3]

6.2 A Permutation Approach

Now consider a test for two independent samples under the Fisher–Pitman permutation model of statistical inference. For the permutation model there is no null hypothesis specifying population parameters. Instead the null hypothesis is simply that all possible arrangements of the observed differences occur with equal chance [4]. Also, there is no alternative hypothesis under the permutation model and no specified α level. Moreover, there is no requirement of random sampling, no assumption of normality, and no assumption of homogeneity of variance. This is

[3] Also see a discussion by S.M. Stigler in *The Seven Pillars of Statistical Wisdom* [14, pp. 91–92].

not to say that the permutation model is unaffected by homogeneity of variance, but it is not a requirement as it is for Student's t test. Under the Neyman–Pearson null hypothesis, if the assumption of homogeneity is not met, t is no longer distributed as Student's t with $N - 2$ degrees of freedom.

A permutation alternative to the conventional test for two independent samples is easily defined. The permutation test statistic for two independent samples is given by

$$\delta = \sum_{i=1}^{2} C_i \xi_i \ , \tag{6.1}$$

where $C_i > 0$ is a positive treatment-group weight for $i = 1, 2$,

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} \Delta(j,k) \Psi_i(\omega_j) \Psi_i(\omega_k) \tag{6.2}$$

is the average distance-function value for all distinct pairs of objects in treatment group S_i for $i = 1, 2$,

$$\Delta(j,k) = \left| x_j - x_k \right|^v \ ,$$

is a symmetric distance-function value for paired objects j and k,

$$N = \sum_{i=1}^{2} n_i \ ,$$

and $\Psi(\cdot)$ is an indicator function given by

$$\Psi_i(\omega_j) = \begin{cases} 1 & \text{if } \omega_j \in S_i \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Under the Fisher–Pitman permutation model, the null hypothesis simply states that equal probabilities are assigned to each of the

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} \tag{6.3}$$

possible, equally-likely allocations of the N objects to the two treatment groups, S_1 and S_2. As noted in Chap. 5, it is imperative that the M possible arrangements of the observed data be generated systematically as expressed in Eq. (6.3), while preserving n_1 and n_2 for each arrangement. Only a systematic procedure guarantees M equally-likely arrangements. Simply shuffling values among the two treatment

groups does not ensure the M possible, equally-likely arrangements mandated by the Fisher–Pitman permutation null hypothesis: all possible arrangements of the observed data occur with equal chance [4].

Under the Fisher–Pitman permutation model, the probability value associated with an observed value of δ, say δ_o, is the probability under the null hypothesis of observing a value of δ as extreme or more extreme than δ_o. Thus an exact probability value for δ_o may be expressed as

$$P\left(\delta \leq \delta_o | H_0\right) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} \,. \tag{6.4}$$

When M is large, an approximate probability value for δ may be obtained from a Monte Carlo procedure, where

$$P\left(\delta \leq \delta_o | H_0\right) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{L}$$

and L denotes the number of randomly-sampled test statistic values. Typically, L is set to a large number to ensure accuracy; for example, $L = 1{,}000{,}000$ [6]. While $L = 1{,}000{,}000$ random arrangements does not guarantee that no two arrangements will be identical, the cycle lengths of modern pseudo-random number generators (PRNG) are sufficiently long that identical arrangements are either avoided or occur so rarely as to be inconsequential. For example, some pseudo-random generators utilize the expanded value of π where the cycle length is so long that it has yet to be determined. Older pseudo-random number generators had a cycle length of only

$$2^{32} - 1 = 4{,}294{,}967{,}295 \,.$$

The Mersenne twister is the current choice for a pseudo-random number generator and is by far the most widely-used general-purpose pseudo-random number generator, having been incorporated into a large number of computer statistical packages, including Microsoft Excel, GAUSS, GLib, Maple, MATLAB, Python, Stata, and the popular R statistical computing language. The cycle length for the Mersenne Twister is $2^{19937} - 1$, which is a very large number.

6.2.1 The Relationship Between Statistics t and δ

When the null hypothesis states H_0: $\mu_1 = \mu_2$, $v = 2$, and the treatment-group weights are given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2} \,,$$

the functional relationships between test statistic δ and Student's t test statistic are given by

$$\delta = \frac{2SS_{\text{Total}}}{t^2 + N - 2} \quad \text{and} \quad t = \left[\frac{2SS_{\text{Total}}}{\delta} - N + 2\right]^{1/2}, \tag{6.5}$$

where

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 \Big/ N$$

and x_i denotes a measurement score for the ith of N objects.

Because of the relationship between test statistic δ and Student's t test statistic, the exact probability values given by

$$P\big(\delta \leq \delta_{\text{o}}\big) = \frac{\text{number of } \delta \text{ values } \leq \delta_{\text{o}}}{M}$$

and

$$P\big(|t| \geq |t_{\text{o}}|\big) = \frac{\text{number of } |t| \text{ values } \geq |t_{\text{o}}|}{M}$$

are equivalent under the Fisher–Pitman null hypothesis, where δ_{o} and t_{o} denote the observed test statistic values of δ and t, respectively, and M is the number of possible, equally-likely arrangements of the observed data.

Also, given $v = 2$ and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2},$$

the two average distance-function values are related to the sample estimates of the population variance by

$$\xi_1 = 2s_1^2 \quad \text{and} \quad \xi_2 = 2s_2^2,$$

test statistic δ is related to the pooled estimate of the population variance by

$$\delta = 2s_p^2,$$

and the exact expected value of the M δ test statistic values is related to SS_{Total} by

$$\mu_\delta = \frac{2SS_{\text{Total}}}{N - 1}.$$

A chance-corrected measure of agreement among response measurement scores is given by

$$\Re = 1 - \frac{\delta}{\mu_\delta} , \tag{6.6}$$

where μ_δ is the arithmetic average of the M δ test statistic values calculated on all possible arrangements of the observed response measurement scores given by

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i . \tag{6.7}$$

6.3 Example 1: Test Statistics t and δ

A small example will serve to illustrate the relationship between test statistics t and δ. Consider a small set of data with $n_1 = 3$ female children in Group 1 and $n_2 = 4$ male children in Group 2, as given in Table 6.1, where the values indicate the ages of the children. Under the Neyman–Pearson population model with null hypothesis H_0: $\mu_1 = \mu_2$, $n_1 = 3$, $n_2 = 4$, $N = n_1 + n_2 = 3 + 4 = 7$, $\bar{x}_1 = 2.3333$, $\bar{x}_2 = 5.25$, $s_1^2 = 2.3333$, $s_2^2 = 2.9167$, the unbiased pooled estimate of the population variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2}$$
$$= \frac{(3 - 1)(2.3333) + (4 - 1)(2.9167)}{7 - 2} = 2.6833 ,$$

and the observed value of Student's t test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{1/2}} = \frac{2.3333 - 5.25}{\left[2.6833 \left(\frac{1}{3} + \frac{1}{4} \right) \right]^{1/2}} = -2.3313 .$$

Table 6.1 Example data for a test of two independent samples with $N = 7$ subjects

Group 1		Group 2	
Females	Age	Males	Age
1	1	4	3
2	2	5	5
3	4	6	6
		7	7

Under the Neyman–Pearson null hypothesis, H_0: $\mu_1 = \mu_2$, test statistic t is asymptotically distributed as Student's t with $N - 2$ degrees of freedom. With $N - 2 = 7 - 2 = 5$ degrees of freedom, the asymptotic two-tail probability value of $t = -2.3313$ is $P = 0.0671$, under the assumptions of normality and homogeneity.

6.3.1 A Permutation Approach

Under the Fisher–Pitman permutation model, employing squared Euclidean scaling with $v = 2$ and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

for correspondence with Student's two-sample t test, the three symmetric distance-function values for Group 1 are

$$\Delta_{1,2} = |1 - 2|^2 = 1 , \quad \Delta_{1,3} = |1 - 4|^2 = 9 , \quad \Delta_{2,3} = |2 - 4|^2 = 4 ,$$

and the average distance-function value for Group 1 is

$$\xi_1 = \binom{n_1}{2}^{-1} (\Delta_{1,2} + \Delta_{1,3} + \Delta_{2,3}) = \binom{3}{2}^{-1} (1 + 9 + 4) = 4.6667 .$$

For Group 2 the six symmetric distance-function values are

$$\Delta_{4,5} = |3 - 5|^2 = 4 , \quad \Delta_{4,6} = |3 - 6|^2 = 9 , \quad \Delta_{4,7} = |3 - 7|^2 = 16 ,$$
$$\Delta_{5,6} = |5 - 6|^2 = 1 , \quad \Delta_{5,7} = |5 - 7|^2 = 4 , \quad \Delta_{6,7} = |6 - 7|^2 = 1 ,$$

and the average distance-function value for Group 2 is

$$\xi_2 = \binom{n_2}{2}^{-1} (\Delta_{4,5} + \Delta_{4,6} + \Delta_{4,7} + \Delta_{5,6} + \Delta_{5,7} + \Delta_{6,7})$$
$$= \binom{4}{2}^{-1} (4 + 9 + 16 + 1 + 4 + 1) = 5.8333 .$$

Then the observed permutation test statistic for the age data listed in Table 6.1 is

$$\delta = C_1 \xi_1 + C_2 \xi_2 = \left(\frac{3 - 1}{7 - 2} \right) (4.6667) + \left(\frac{4 - 1}{7 - 2} \right) (5.8333) = 5.3667 .$$

For the example data given in Table 6.1, the sum of the $N = 7$ observations is

$$\sum_{i=1}^{N} x_i = 1 + 2 + 4 + 3 + 5 + 6 + 7 = 28 ,$$

the sum of the $N = 7$ squared observations is

$$\sum_{i=1}^{N} x_i^2 = 1^2 + 2^2 + 4^2 + 3^2 + 5^2 + 6^2 + 7^2 = 140 ,$$

and the total sum-of-squares is

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N = 140 - (28)^2/7 = 28.00 .$$

Then based on the expressions given in Eq. (6.5) on p. 159, the observed value for test statistic δ with respect to the observed value of Student's t statistic is

$$\delta = \frac{2SS_{\text{Total}}}{t^2 + N - 2} = \frac{2(28.00)}{(-2.3313)^2 + 7 - 2} = 5.3667$$

and the observed value for Student's t test statistic with respect to the observed value of test statistic δ is

$$t = \left(\frac{2SS_{\text{Total}}}{\delta} - N + 2 \right)^{1/2} = \left[\frac{2(28.00)}{5.3667} - 7 + 2 \right]^{1/2} = \pm 2.3313 .$$

Under the Fisher–Pitman permutation model there are exactly

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(3 + 4)!}{3!\, 4!} = 35$$

possible, equally-likely arrangements in the reference set of all permutations of the age data listed in Table 6.1 on p. 160. Since $M = 35$ is a relatively small number, it is possible to list the $M = 35$ arrangements in Table 6.2, along with the corresponding values for ξ_1, ξ_2, δ, and $|t|$, ordered by δ values from low ($\delta_1 = 2.8000$) to high ($\delta_{35} = 11.2000$) and by $|t|$ values from high ($t_1 = 3.8730$) to low ($t_{35} = 0.0000$).

Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 5.3667$. The observed permutation test statistic, $\delta = 5.3667$, obtained for the realized arrangement is unusual since 31 of the 35 δ test statistic values exceed the observed value and only four of the δ test statistic values are

Table 6.2 Arrangements of the observed data listed in Table 6.1 with corresponding ξ_1, ξ_2, δ, and $|t|$ values

Number	Arrangement	ξ_1	ξ_2	δ	$\lvert t\rvert$
1*	1, 2, 3 4, 5, 6, 7	2.0000	3.3333	2.8000	3.8730
2*	5, 6, 7 1, 2, 3, 4	2.0000	3.3333	2.8000	3.8730
3*	1, 2, 4 3, 5, 6, 7	4.6667	5.8333	5.3667	2.3313
4*	4, 6, 7 1, 2, 3, 5	4.6667	5.8333	5.3667	2.3313
5	1, 2, 5 3, 4, 6, 7	8.6667	6.6667	7.4667	1.5811
6	1, 3, 4 2, 5, 6, 7	4.6667	9.3333	7.4667	1.5811
7	4, 5, 7 1, 2, 3, 6	4.6667	9.3333	7.4667	1.5811
8	3, 6, 7 1, 2, 4, 5	8.6667	6.6666	7.4667	1.5811
9	1, 2, 6 3, 4, 5, 7	14.0000	5.8333	9.1000	1.0742
10	1, 3, 5 2, 4, 6, 7	8.0000	9.8333	9.1000	1.0742
11	2, 3, 4 1, 5, 6, 7	2.0000	13.8333	9.1000	1.0742
12	2, 6, 7 1, 3, 4, 5	14.0000	5.8333	9.1000	1.0742
13	3, 5, 7 1, 2, 4, 6	8.0000	9.8333	9.1000	1.0742
14	4, 5, 6 1, 2, 3, 7	2.0000	13.8333	9.1000	1.0742
15	1, 2, 7 3, 4, 5, 6	20.6667	3.3333	10.2667	0.6742
16	1, 3, 6 2, 4, 5, 7	12.6667	8.6667	10.2667	0.6742
17	1, 4, 5 2, 3, 6, 7	8.6667	11.3333	10.2667	0.6742
18	1, 6, 7 2, 3, 4, 5	20.6667	3.3333	10.2667	0.6742
19	2, 3, 5 1, 4, 6, 7	4.6667	14.0000	10.2667	0.6742
20	2, 5, 7 1, 3, 4, 6	12.6667	8.6667	10.2667	0.6742
21	3, 4, 7 1, 2, 5, 6	8.6667	11.3333	10.2667	0.6742
22	3, 5, 6 1, 2, 4, 7	4.6667	14.0000	10.2667	0.6742
23	3, 4, 6 1, 2, 5, 7	4.6667	15.1667	10.9667	0.3262
24	1, 3, 7 2, 4, 5, 6	18.6667	5.8333	10.9667	0.3262
25	1, 4, 6 2, 3, 5, 7	12.6667	9.8333	10.9667	0.3262
26	1, 5, 7 2, 3, 4, 6	18.6667	5.8333	10.9667	0.3262
27	2, 3, 6 1, 4, 5, 7	8.6667	12.5000	10.9667	0.3262
28	2, 4, 5 1, 3, 6, 7	4.6667	15.1667	10.9667	0.3262
29	2, 4, 7 1, 3, 5, 6	12.6667	9.8333	10.9667	0.3262
30	2, 5, 6 1, 3, 4, 7	8.6667	12.5000	10.9667	0.3262
31	1, 4, 7 2, 3, 5, 6	18.0000	6.6667	11.2000	0.0000
32	1, 5, 6 2, 3, 4, 7	14.0000	9.3333	11.2000	0.0000
33	2, 3, 7 1, 4, 5, 6	14.0000	9.3333	11.2000	0.0000
34	2, 4, 6 1, 3, 5, 7	8.0000	13.3333	11.2000	0.0000
35	3, 4, 5 1, 2, 6, 7	2.0000	17.3333	11.2000	0.0000
Sum				326.6667	

equal to or less than the observed value. The rows containing the lowest four δ test statistic values are indicated with asterisks in Table 6.2. If all M arrangements of the observed data occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 5.3667$ computed on the $M = 35$ possible

arrangements of the observed data with $n_1 = 3$ and $n_2 = 4$ preserved for each arrangement is

$$P(\delta \le \delta_o) = \frac{\text{number of } \delta \text{ values } \le \delta_o}{M} = \frac{4}{35} = 0.1143 ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 7$ observations listed in Table 6.1.

Alternatively, there are only four $|t|$ test statistic values that are larger than the observed value of $|t| = 2.3313$. The rows containing the highest four $|t|$ values are indicated with asterisks in Table 6.2. Thus if all M arrangements of the observed data occur with equal chance, the exact probability value of $|t| = 2.3313$ under the Fisher–Pitman null hypothesis is

$$P(|t| \ge |t_o|) = \frac{\text{number of } |t| \text{ values } \ge |t_o|}{M} = \frac{4}{35} = 0.1143 ,$$

where t_o denotes the observed value of test statistic t. There is a considerable difference between the asymptotic probability value of $P = 0.0671$ and the exact probability value of $P = 0.1143$; that is,

$$\Delta_P = 0.1143 - 0.0671 = 0.0472 .$$

A continuous mathematical function such as Student's t cannot be expected to provide a precise fit to only $n_1 = 3$ and $n_2 = 4$ observed values.

For the example data listed in Table 6.1 on p. 160, the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{326.6667}{35} = 9.3333 .$$

Alternatively, under an analysis of variance model the exact expected value of test statistic δ is

$$\mu_\delta = \frac{2SS_{\text{Total}}}{N-1} = \frac{2(28.00)}{7-1} = 9.3333 ,$$

where the sum of the $N = 7$ observations listed in Table 6.1 is

$$\sum_{i=1}^{N} x_i = 1 + 2 + 4 + 3 + 5 + 6 + 7 = 28 ,$$

the sum of the $N = 7$ squared observations is

$$\sum_{i=1}^{N} x_i^2 = 1^2 + 2^2 + 4^2 + 3^2 + 5^2 + 6^2 + 7^2 = 140 ,$$

and the total sum-of-squares is

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N = 140 - (28)^2/7 = 28.00 .$$

Then the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{5.3667}{9.3333} = +0.4250 ,$$

indicating approximately 42% within-group agreement above what is expected by chance.

6.4 Example 2: Measures of Effect Size

Measures of effect size express the practical or clinical significance of a difference between independent sample means, as contrasted with the statistical significance of a difference. Five measures of effect size are commonly used for determining the magnitude of treatment effects in conventional tests for two independent samples: Cohen's $\hat{d}$ measure of effect size given by

$$\hat{d} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2}} ,$$

Pearson's r^2 measure of effect size given by

$$r^2 = \frac{t^2}{t^2 + N - 2} ,$$

Kelley's ϵ^2 measure of effect size given by

$$\epsilon^2 = \frac{t^2 - 1}{t^2 + N - 2} ,$$

Hays' $\hat{\omega}^2$ measure of effect size given by

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N - 1} ,$$

and Mielke and Berry's $\Re$ measure of effect size given by

$$\Re = 1 - \frac{\delta}{\mu_\delta} ,$$

where the permutation test statistic δ is defined in Eq. (6.1) on p. 157 and μ_δ is the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis given by

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i ,$$

where for a test of two independent samples, the number of possible arrangements of the observed data is given by

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} .$$

For the age data given in Table 6.1 on p. 160 for $N = 7$ subjects, Student's test statistic is $t = -2.3313$, Cohen's $\hat{d}$ measure of effect size is

$$\hat{d} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2}} = \frac{|2.3333 - 5.25|}{\sqrt{2.6833}} = 1.7805 ,$$

indicating a strong effect size ($\hat{d} \geq 0.80$); Pearson's r^2 measure of effect size is

$$r^2 = \frac{t^2}{t^2 + N - 2} = \frac{(-2.3313)^2}{(-2.3313)^2 + 7 - 2} = 0.5208 ,$$

also indicating a strong effect size ($r^2 \geq 0.25$); Kelley's ϵ^2 measure of effect size is

$$\epsilon^2 = \frac{t^2}{t^2 + N - 2} = \frac{(-2.3313)^2 - 1}{(-2.3313)^2 + 7 - 2} = 0.4250 ;$$

Hays' $\hat{\omega}^2$ measure of effect size is

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N - 1} = \frac{(-2.3313)^2 - 1}{(-2.3313)^2 + 7 - 1} = 0.3878 ;$$

and Mielke and Berry's $\Re$ measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{5.3667}{9.3333} = +0.4250\,,$$

where δ is defined in Eq. (6.1) on p. 157, μ_δ is the exact expected value of δ under the Fisher–Pitman null hypothesis given by

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{326.6667}{35} = 9.3333\,,$$

and the sum of the $M = 35$ δ test statistic values,

$$\sum_{i=1}^{M}\delta_i = 326.6667\,,$$

is calculated in Table 6.2 on p. 163.

It is readily apparent that for a test of two independent samples, the five measures of effect size, $\hat{d}$, r^2, ϵ^2, $\hat{\omega}^2$, and $\Re$ provide similar results when $v = 2$,

$$C_1 = \frac{n_1 - 1}{N - 2}\,, \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}\,,$$

and are directly related to each other and to Student's t test statistic for two independent samples. It can easily be shown that Kelley's ϵ^2 and Mielke and Berry's $\Re$ are identical measures of effect size for two independent samples under the Neyman–Pearson population model; that is,

$$\epsilon^2 = \Re = \frac{t^2 - 1}{t^2 + N - 2} = \frac{(-2.3313)^2 - 1}{(-2.3313)^2 + 7 - 2} = +0.4250\,.$$

6.4.1 *Efficient Calculation of μ_δ*

Although the exact expected value of test statistic δ is defined as

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i\,, \tag{6.8}$$

there is a more efficient way to calculate the expected value of δ than utilizing the expression given in Eq. (6.8) [11]. Because permutation methods are by their very nature computationally intensive methods, efficient calculation of the permutation

test statistic δ and the exact expected value of δ is imperative. Define

$$\mu_\delta = \frac{(N-2)!}{N!} \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta(i,j) \ , \tag{6.9}$$

where the symmetric distance function between paired objects i and j is given by $\Delta(i,j) = |x_i - x_j|^v$.

Table 6.3 illustrates the calculation of μ_δ for the example data listed in Table 6.1 with $v = 2$. Given the double sum,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \Delta(i,j) = 392$$

Table 6.3 Example data for a test of two independent samples with $N = 7$ subjects and $N^2 = 7^2 = 49$ possible values

Index	$\Delta(i,j)$	Index	$\Delta(i,j)$
1	$\Delta(1,1) = \|1-1\|^2 = 0$	26	$\Delta(4,5) = \|4-5\|^2 = 1$
2	$\Delta(1,2) = \|1-2\|^2 = 1$	27	$\Delta(4,6) = \|4-6\|^2 = 4$
3	$\Delta(1,3) = \|1-3\|^2 = 9$	28	$\Delta(4,7) = \|4-7\|^2 = 9$
4	$\Delta(1,4) = \|1-3\|^2 = 4$	29	$\Delta(5,1) = \|5-1\|^2 = 16$
5	$\Delta(1,5) = \|1-5\|^2 = 16$	30	$\Delta(5,2) = \|5-2\|^2 = 9$
6	$\Delta(1,6) = \|1-6\|^2 = 25$	31	$\Delta(5,3) = \|5-4\|^2 = 1$
7	$\Delta(1,7) = \|1-7\|^2 = 36$	32	$\Delta(5,4) = \|5-3\|^2 = 4$
8	$\Delta(2,1) = \|2-1\|^2 = 1$	33	$\Delta(5,5) = \|5-5\|^2 = 0$
9	$\Delta(2,2) = \|2-2\|^2 = 0$	34	$\Delta(5,6) = \|5-6\|^2 = 1$
10	$\Delta(2,3) = \|2-4\|^2 = 4$	35	$\Delta(5,7) = \|5-7\|^2 = 4$
11	$\Delta(2,4) = \|2-3\|^2 = 1$	36	$\Delta(6,1) = \|6-1\|^2 = 25$
12	$\Delta(2,5) = \|2-5\|^2 = 9$	37	$\Delta(6,2) = \|6-2\|^2 = 16$
13	$\Delta(2,6) = \|2-6\|^2 = 16$	38	$\Delta(6,3) = \|6-4\|^2 = 4$
14	$\Delta(2,7) = \|1-2\|^2 = 25$	39	$\Delta(6,4) = \|6-3\|^2 = 9$
15	$\Delta(3,1) = \|3-1\|^2 = 4$	40	$\Delta(6,5) = \|6-5\|^2 = 1$
16	$\Delta(3,2) = \|3-2\|^2 = 1$	41	$\Delta(6,6) = \|6-6\|^2 = 0$
17	$\Delta(3,3) = \|3-4\|^2 = 1$	42	$\Delta(6,7) = \|6-7\|^2 = 1$
18	$\Delta(3,4) = \|3-3\|^2 = 0$	43	$\Delta(7,1) = \|7-1\|^2 = 36$
19	$\Delta(3,5) = \|3-5\|^2 = 4$	44	$\Delta(7,2) = \|7-2\|^2 = 25$
20	$\Delta(3,6) = \|3-6\|^2 = 9$	45	$\Delta(7,3) = \|7-4\|^2 = 9$
21	$\Delta(3,7) = \|3-7\|^2 = 16$	46	$\Delta(7,4) = \|7-3\|^2 = 16$
22	$\Delta(4,1) = \|4-1\|^2 = 9$	47	$\Delta(7,5) = \|7-5\|^2 = 4$
23	$\Delta(4,2) = \|4-2\|^2 = 4$	48	$\Delta(7,6) = \|7-6\|^2 = 1$
24	$\Delta(4,3) = \|4-4\|^2 = 0$	49	$\Delta(7,7) = \|7-7\|^2 = 0$
25	$\Delta(4,4) = \|4-3\|^2 = 1$		
Sum			392

calculated in Table 6.3,

$$\mu_\delta = \frac{(N-2)!}{N!}\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta(i,\,j) = \frac{(7-2)!}{7!}(392) = \frac{47{,}040}{5{,}040} = 9.3333\,.$$

Thus the actual computation of μ_δ involves only N^2 operations to obtain the exact expected value of test statistic δ. For example, if $n_1 = n_2 = 15$ there are

$$M = \frac{(n_1+n_2)!}{n_1!\,n_2!} = \frac{(15+15)!}{15!\;15!} = 155{,}117{,}520$$

δ test statistic values to be computed using the expression for μ_δ given in Eq. (6.8), but only $(15{+}15)^2 = 30^2 = 900\;\Delta(i,\,j)$ values to be computed using the expression for μ_δ given in Eq. (6.9) for $i, j = 1, \ldots, N$—a much more efficient solution resulting in a substantial savings in computation time.

6.4.2 Comparisons of Effect Size Measures

The four measures of effect size and Student's t test statistic are all interrelated. Any one of the measures can be derived from any of the other measures. The functional relationships between Student's t test statistic and Mielke and Berry's $\Re$ measure of effect size for tests of two independent samples are given by

$$t = \left[\frac{\Re(N-2)+1}{1-\Re}\right]^{1/2} \quad \text{and} \quad \Re = \frac{t^2-1}{t^2+N-2}\,, \tag{6.10}$$

the relationships between Pearson's r^2 measure of effect size and Mielke and Berry's $\Re$ measure of effect size are given by

$$r^2 = \Re + \left(t^2+N-2\right)^{-1} \quad \text{and} \quad \Re = r^2 - \left(t^2+N-2\right)^{-1}\,, \tag{6.11}$$

the relationships between Hays' $\hat{\omega}^2$ measure of effect size and Mielke and Berry's $\Re$ measure of effect size are given by

$$\hat{\omega}^2 = \Re\left(\frac{t^2+N-2}{t^2+N-1}\right) \quad \text{and} \quad \Re = \hat{\omega}^2\left(\frac{t^2+N-1}{t^2+N-2}\right)\,, \tag{6.12}$$

the relationships between Cohen's $\hat{d}$ measure of effect size and Mielke and Berry's $\Re$ measure of effect size are given by

$$\hat{d} = \left[\frac{\Re N(N-2)+N}{n_1 n_2(1-\Re)}\right]^{1/2} \quad \text{and} \quad \Re = \frac{n_1 n_2 \hat{d}^{\,2} - N}{n_1 n_2 \hat{d}^{\,2} + N(N-2)}\,, \tag{6.13}$$

the relationships between Cohen's $\hat{d}$ measure of effect size and Student's t test statistic are given by

$$\hat{d} = \left(\frac{Nt^2}{n_1 n_2}\right)^{1/2} \quad \text{and} \quad t = \left(\frac{n_1 n_2 \hat{d}^{\,2}}{N}\right)^{1/2}, \tag{6.14}$$

the relationships between Pearson's r^2 measure of effect size and Student's t test statistic are given by

$$r^2 = \frac{t^2}{t^2 + N - 2} \quad \text{and} \quad t = \left[\frac{r^2(N-2)}{1-r^2}\right]^{1/2}, \tag{6.15}$$

the relationships between Pearson's r^2 measure of effect size and Cohen's $\hat{d}$ measure of effect size are given by

$$r^2 = \frac{n_1 n_2 \hat{d}^{\,2}}{n_1 n_2 \hat{d}^{\,2} + N(N-2)} \quad \text{and} \quad \hat{d} = \left[\frac{r^2 N(N-2)}{n_1 n_2 (1-r^2)}\right]^{1/2}, \tag{6.16}$$

the relationships between Pearson's r^2 measure of effect size and Hays' $\hat{\omega}^2$ measure of effect size are given by

$$r^2 = \frac{\hat{\omega}^2(N-1)+1}{\hat{\omega}^2 + N - 1} \quad \text{and} \quad \hat{\omega}^2 = \frac{r^2(N-1)-1}{N-(1+r^2)}, \tag{6.17}$$

the relationships between Student's t test statistic and Hays' $\hat{\omega}^2$ measure of effect size are given by

$$t = \left[\frac{\hat{\omega}^2(N-1)+1}{1-\hat{\omega}^2}\right]^{1/2} \quad \text{and} \quad \hat{\omega}^2 = \frac{t^2-1}{t^2+N-1}, \tag{6.18}$$

and the relationships between Cohen's $\hat{d}$ measure of effect size and Hays' $\hat{\omega}^2$ measure of effect size are given by

$$\hat{d} = \left\{\frac{N\left[\hat{\omega}^2(N-1)+1\right]}{n_1 n_2 (1-\hat{\omega}^2)}\right\}^{1/2} \quad \text{and} \quad \hat{\omega}^2 = \frac{n_1 n_2 \hat{d}^{\,2} - N}{n_1 n_2 \hat{d}^{\,2} + N(N-1)}. \tag{6.19}$$

It is important to note that the relationships between Student's t and Mielke and Berry's $\Re$, Pearson's r^2 and Mielke and Berry's $\Re$, Hays' $\hat{\omega}^2$ and Mielke and Berry's $\Re$, Cohen's $\hat{d}$ and Mielke and Berry's $\Re$, Cohen's $\hat{d}$ and Student's t, Pearson's r^2 and Student's t, Pearson's r^2 and Cohen's $\hat{d}$, Pearson's r^2 and Hays' $\hat{\omega}^2$, Student's t and Hays' $\hat{\omega}^2$, and Cohen's $\hat{d}$ and Hays' $\hat{\omega}^2$ hold only for Student's *pooled* two-sample t test. The measures of effect size, $\hat{d}$, r^2, and $\hat{\omega}^2$, all require homogeneity

of variance and the relationships listed above do not hold for Student's non-pooled two-sample t test. On the other hand, $\Re$ does not require homogeneity of variance and is appropriate for both pooled and non-pooled two-sample tests [5].

6.4.3 Example Effect Size Comparisons

In this section, comparisons of Student's t, Cohen's $\hat{d}$, Mielke and Berry's $\Re$, Hays' $\hat{\omega}^2$, and Pearson's r^2 are illustrated with the example data listed in Table 6.1 on p. 160 with $n_1 = 3$, $n_2 = 4$, and $N = 7$ observations.

Given the age data listed in Table 6.1 and following the expressions given in Eq. (6.10) for Student's t test statistic and Mielke and Berry's $\Re$ measure of effect size, the observed value for Student's t test statistic with respect to the observed value of Mielke and Berry's $\Re$ measure of effect size is

$$t = \left[\frac{\Re(N-2)+1}{1-\Re}\right]^{1/2} = \left[\frac{0.4250(7-2)+1}{1-0.4250}\right]^{1/2} = \pm 2.3313$$

and the observed value for Mielke and Berry's $\Re$ measure of effect size with respect to the observed value of Student's t test statistic is

$$\Re = \frac{t^2-1}{t^2+N-2} = \frac{(-2.3313)^2-1}{(-2.3313)^2+7-2} = +0.4250\,.$$

Following the expressions given in Eq. (6.11) for Pearson's r^2 measure of effect size and Mielke and Berry's $\Re$ measure of effect size, the observed value for Pearson's r^2 measure of effect size with respect to the observed value of Mielke and Berry's $\Re$ measure of effect size is

$$r^2 = \Re + \left(t^2+N-2\right)^{-1} = 0.4250 + \left[(-2.3313)^2+7-2\right]^{-1} = 0.5208$$

and the observed value for Mielke and Berry's $\Re$ measure of effect size with respect to the observed value of Pearson's r^2 measure of effect size is

$$\Re = r^2 - \left(t^2+N-2\right)^{-1} = 0.5208 - \left[(-2.3313)^2+7-2\right]^{-1} = +0.4250\,.$$

Following the expressions given in Eq. (6.12) for Hays' $\hat{\omega}^2$ measure of effect size and Mielke and Berry's $\Re$ measure of effect size, the observed value for Hays' $\hat{\omega}^2$ measure of effect size with respect to the observed value of Mielke and Berry's $\Re$ measure of effect size is

$$\hat{\omega}^2 = \Re\left(\frac{t^2+N-2}{t^2+N-1}\right) = \left[\frac{(-2.3313)^2+7-2}{(-2.3313)^2+7-1}\right] = 0.3878$$

and the observed value for Mielke and Berry's $\Re$ measure of effect size with respect to the observed value of Hays' $\hat{\omega}^2$ measure of effect size is

$$\Re = \hat{\omega}^2 \left(\frac{t^2 + N - 1}{t^2 + N - 2} \right) = \left[\frac{(-2.3313)^2 + 7 - 1}{(-2.3313)^2 + 7 - 2} \right] = +0.4250 \, .$$

Following the expressions given in Eq. (6.13) for Cohen's $\hat{d}$ measure of effect size and Mielke and Berry's $\Re$ measure of effect size, the observed value for Cohen's $\hat{d}$ measure of effect size with respect to the observed value of Mielke and Berry's $\Re$ measure of effect size is

$$\hat{d} = \left[\frac{\Re N(N-2) + N}{n_1 n_2 (1 - \Re)} \right]^{1/2} = \left[\frac{(0.4250)(7)(7-2) + 7}{(3)(4)(1 - 0.4250)} \right]^{1/2} = \pm 1.7805$$

and the observed value for Mielke and Berry's $\Re$ measure of effect size with respect to the observed value of Cohen's $\hat{d}$ measure of effect size is

$$\Re = \frac{n_1 n_2 \hat{d}^{\,2} - N}{n_1 n_2 \hat{d}^{\,2} + N(N-2)} = \frac{(3)(4)(1.7805)^2}{(3)(4)(1.7805)^2 + (7)(7-2)} = +0.4250 \, .$$

Following the expressions given in Eq. (6.14) for Cohen's $\hat{d}$ measure of effect size and Student's t test statistic, the observed value for Cohen's $\hat{d}$ measure of effect size with respect to the observed value of Student's t statistic is

$$\hat{d} = \left(\frac{N t^2}{n_1 n_2} \right)^{1/2} = \left[\frac{7(-2.3313)^2}{(3)(4)} \right]^{1/2} = \pm 1.7805$$

and the observed value of Student's t test statistic with respect to the observed value of Cohen's $\hat{d}$ measure of effect size is

$$t = \left(\frac{n_1 n_2 \hat{d}^{\,2}}{N} \right)^{1/2} = \left[\frac{(3)(4)(1.7805)^2}{7} \right]^{1/2} = \pm 2.3313 \, .$$

Following the expressions given in Eq. (6.15) for Pearson's r^2 measure of effect size and Student's t test statistic, the observed value for Pearson's r^2 measure of effect size with respect to the observed value of Student's t statistic is

$$r^2 = \frac{t^2}{t^2 + N - 2} = \frac{(-2.3313)^2}{(-2.3313)^2 + 7 - 2} = 0.5208$$

and the observed value for Student's t test statistic with respect to the observed value of Pearson's r^2 measure of effect size is

$$t = \left[\frac{r^2(N-2)}{1-r^2}\right]^{1/2} = \left[\frac{0.5208(7-2)}{1-0.5208}\right]^{1/2} = \pm 2.3313 \, .$$

Following the expressions given in Eq. (6.16) for Pearson's r^2 measure of effect size and Cohen's $\hat{d}$ measure of effect size, the observed value for Pearson's r^2 measure of effect size with respect to the observed value of Cohen's $\hat{d}$ measure of effect size is

$$r^2 = \frac{n_1 n_2 \hat{d}^2}{n_1 n_2 \hat{d}^2 + N(N-2)} = \frac{(3)(4)(-1.7805)^2}{(3)(4)(-1.7805)^2 + 7(7-2)} = 0.5208$$

and the observed value for Cohen's $\hat{d}$ measure of effect size with respect to the observed value of Pearson's r^2 measure of effect size is

$$\hat{d} = \left[\frac{r^2 N(N-2)}{n_1 n_2 (1-r^2)}\right]^{1/2} = \left[\frac{(0.5208)(7)(7-2)}{(3)(4)(1-0.5208)}\right]^{1/2} = \pm 1.7805 \, .$$

Following the expressions given in Eq. (6.17) for Pearson's r^2 measure of effect size and Hays' $\hat{\omega}^2$ measure of effect size, the observed value for Pearson's r^2 measure of effect size with respect to the observed value of Hays' $\hat{\omega}^2$ measure of effect size is

$$r^2 = \frac{\hat{\omega}^2(N-1)+1}{\hat{\omega}^2 + N - 1} = \frac{(0.3878)(7-1)+1}{0.3878 + 7 - 1} = 0.5208$$

and the observed value for Hays' $\hat{\omega}^2$ measure of effect size with respect to the observed value of Pearson's r^2 measure of effect size is

$$\hat{\omega}^2 = \frac{r^2(N-1)-1}{N-(1+r^2)} = \frac{(0.5208)(7-1)-1}{7-(1+0.5208)} = 0.3878 \, .$$

Following the expressions given in Eq. (6.18) for Student's t test statistic and Hays' $\hat{\omega}^2$ measure of effect size, the observed value for Student's t statistic with respect to the observed value of Hays' $\hat{\omega}^2$ measure of effect size is

$$t = \left[\frac{\hat{\omega}^2(N-1)+1}{1-\hat{\omega}^2}\right]^{1/2} = \left[\frac{(0.3878)(7-1)+1}{1-0.3878}\right]^{1/2} = \pm 2.3313$$

and the observed value for Hays' $\hat{\omega}^2$ measure of effect size with respect to the observed value of Student's t test statistic is

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N - 1} = \frac{(-2.3313)^2 - 1}{(-2.3313)^2 + 7 - 1} = 0.3878 \, .$$

And following the expressions given in Eq. (6.19) for Cohen's $\hat{d}$ measure of effect size and Hays' $\hat{\omega}^2$ measure of effect size, the observed value for Cohen's $\hat{d}$ measure of effect size with respect to the observed value of Hays' $\hat{\omega}^2$ measure of effect size is

$$\hat{d} = \left\{ \frac{N\big[\hat{\omega}^2(N-1)+1\big]}{n_1 n_2 (1 - \hat{\omega}^2)} \right\}^{1/2} = \left\{ \frac{7\big[0.3878(7-1)+1\big]}{(3)(4)(1-0.3878)} \right\}^{1/2} = \pm 1.7805$$

and the observed value for Hays' $\hat{\omega}^2$ measure of effect size with respect to the observed value of Cohen's $\hat{d}$ measure of effect size is

$$\hat{\omega}^2 = \frac{n_1 n_2 \hat{d}^{\,2} - N}{n_1 n_2 \hat{d}^{\,2} + N(N-1)} = \frac{(3)(4)(-1.7805)^2 - 7}{(3)(4)(-1.7805)^2 + 7(7-1)} = 0.3878 \, .$$

6.5 Example 3: Analyses with $v = 2$ and $v = 1$

For a third example of tests of differences for two independent samples, consider the error scores obtained for two groups of experimental animals running a maze under two different treatment conditions: treatment Group 1 without a reward and treatment Group 2 with a reward. The example data are given in Table 6.4.

Under the Neyman–Pearson population model with H_0: $\mu_1 = \mu_2$, $n_1 = 8$, $n_2 = 6$, $N = 14$, $\bar{x}_1 = 11.00$, $\bar{x}_2 = 8.00$, $s_1^2 = 57.7143$, $s_2^2 = 63.60$, the unbiased

Table 6.4 Example data for a test of two independent samples with $N = 14$ subjects

Group 1		Group 2	
Subject	Error	Subject	Error
1	16	9	20
2	9	10	5
3	4	11	1
4	23	12	16
5	19	13	2
6	10	14	4
7	5		
8	2		

pooled estimate of the population variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2}$$
$$= \frac{(8 - 1)(57.7143) + (6 - 1)(63.60)}{14 - 2} = 60.1667 ,$$

and the observed value of Student's t test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{1/2}} = \frac{11.00 - 8.00}{\left[60.1667 \left(\frac{1}{8} + \frac{1}{6}\right)\right]^{1/2}} = +0.7161 .$$

Under the Neyman–Pearson null hypothesis, H_0: $\mu_1 = \mu_2$, test statistic t is asymptotically distributed as Student's t with $N - 2$ degrees of freedom. With $N - 2 = 14 - 2 = 12$ degrees of freedom, the asymptotic two-tail probability value of $t = +0.7161$ is $P = 0.4876$, under the assumptions of normality and homogeneity.

6.5.1 An Exact Analysis with $v = 2$

Under the Fisher–Pitman permutation model, employing squared Euclidean scaling with $v = 2$ and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

for correspondence with Student's two-sample t test, the average distance-function values for treatment Groups 1 and 2 are

$$\xi_1 = 115.4286 \quad \text{and} \quad \xi_2 = 127.20 ,$$

respectively, and the observed permutation test statistic value is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{8 - 1}{14 - 2}\right)(115.4286) + \left(\frac{6 - 1}{14 - 2}\right)(127.20) = 120.3333 .$$

Alternatively, in terms of Student's t test statistic the average distance-function values are

$$\xi_1 = 2s_1^2 = 2(57.7143) = 115.4286 , \quad \xi_2 = 2s_2^2 = 2(63.60) = 127.20 ,$$

and the observed permutation test statistic value is

$$\delta = 2s_p^2 = 2(60.1667) = 120.3333 \, .$$

For the example data listed in Table 6.4, the sum of the $N = 14$ observations is

$$\sum_{i=1}^{N} x_i = 16 + 9 + 4 + 23 + 19 + 10 + 5 + 2 + 20 + 5 + 1 + 16 + 2 + 4 = 136 \, ,$$

the sum of the $N = 14$ squared observations is

$$\sum_{i=1}^{N} x_i^2 = 16^2 + 9^2 + 4^2 + 23^2 + 19^2 + 10^2 + 5^2 + 2^2 + 20^2 + 5^2 + 1^2 + 16^2 + 2^2 + 4^2 = 2074 \, ,$$

and the total sum-of-squares is

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N = 2074 - (136)^2/14 = 752.8571 \, .$$

Then based on the expressions given in Eq. (6.5), the observed value for test statistic δ with respect to the observed value of Student's t test statistic is

$$\delta = \frac{2SS_{\text{Total}}}{t^2 + N - 2} = \frac{2(752.8571)}{(+0.7161)^2 + 14 - 2} = 120.3333$$

and the observed value for Student's t test statistic with respect to the observed value of test statistic δ is

$$t = \left(\frac{2SS_{\text{Total}}}{\delta} - N + 2 \right)^{1/2} = \left[\frac{2(752.8571)}{120.3333} - 14 + 2 \right]^{1/2} = \pm 0.7161 \, .$$

Under the Fisher–Pitman permutation model there are only

$$M = \frac{(n_1 + n_2)!}{n_1! \, n_2!} = \frac{(8 + 6)!}{8! \, 6!} = 3003$$

possible, equally-likely arrangements in the reference set of all permutations of the error data listed in Table 6.4, making an exact permutation analysis possible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely

arrangements of the observed data that are equal to or less than the observed value of $\delta = 120.3333$. There are exactly 1487 δ test statistic values that are equal to or less than the observed value of $\delta = 120.3333$. If all M arrangements of the $N = 14$ observations listed in Table 6.4 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 120.3333$ computed on the $M = 3003$ possible arrangements of the observed data with $n_1 = 8$ and $n_2 = 6$ preserved for each arrangement is

$$P\big(\delta \le \delta_{\mathrm{o}}\big) = \frac{\text{number of } \delta \text{ values } \le \delta_{\mathrm{o}}}{M} = \frac{1487}{3003} = 0.5275\,,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 14$ observations listed in Table 6.4. Alternatively, the exact two-tail probability value of $|t| = 0.7161$ is

$$P\big(|t| \ge |t_{\mathrm{o}}|\big) = \frac{\text{number of } |t| \text{ values } \ge |t_{\mathrm{o}}|}{M} = \frac{1487}{3003} = 0.5275\,,$$

where t_{o} denotes the observed value of test statistic t.

6.5.2 Measures of Effect Size

For the example data listed in Table 6.4 on p. 174, Cohen's $\hat{d}$ measure of effect size is

$$\hat{d} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2}} = \frac{|11.00 - 8.00|}{\sqrt{60.1667}} = 0.3868\,,$$

Pearson's r^2 measure of effect size is

$$r^2 = \frac{t^2}{t^2 + N - 2} = \frac{(+0.7161)^2}{(+0.7161)^2 + 14 - 2} = 0.0410\,,$$

Kelley's ϵ^2 measure of effect size is

$$\epsilon^2 = \frac{t^2 - 1}{t^2 + N - 2} = \frac{(+0.7161)^2 - 1}{(+0.7161)^2 + 14 - 2} = -0.0389\,, \tag{6.20}$$

Hays' $\hat{\omega}^2$ measure of effect size is

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N - 1} = \frac{(+0.7161)^2 - 1}{(+0.7161)^2 + 14 - 1} = -0.0361\,, \tag{6.21}$$

and Mielke and Berry's $\Re$ measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{120.3333}{115.8242} = -0.0389 ,$$

where, for the example data listed in Table 6.4, the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{347{,}820}{3003} = 115.8242 .$$

Alternatively, under an analysis of variance model,

$$\mu_\delta = \frac{2SS_{\text{Total}}}{N-1} = \frac{2(752.8571)}{14-1} = 115.8242 ,$$

where

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N = 2074 - (136)^2/14 = 752.8571 .$$

6.5.3 Chance-Corrected Measures of Effect Size

As is evident in Eqs. (6.20) and (6.21), some squared measures of effect size can be negative; in this case, Kelley's $\epsilon^2 = -0.0389$ and Hays' $\hat{\omega}^2 = -0.0361$. It is somewhat disconcerting, to say the least, to try to interpret squared coefficients with negative values. It is also important to recognize that negative values cannot simply be dismissed on theoretical grounds [10, p. 1000]. A number of authors have suggested that negative values be treated as zero [9]. It is not widely recognized that, like Mielke and Berry's $\Re$ measure of effect size, Kelley's ϵ^2 and Hays' $\hat{\omega}^2$ are chance-corrected measures of effect size. In fact $\Re$ and ϵ^2 are equivalent measures of effect size for tests of two independent samples. This places Kelley's ϵ^2 and Hays' $\hat{\omega}^2$ into the family of chance-corrected measures that includes such well-known members as Scott's π coefficient of inter-coder agreement [12], Cohen's κ coefficient of weighted agreement [2], Kendall and Babington Smith's u measure of agreement [7], and Spearman's footrule measure [13]. Negative values simply indicate that the magnitude of the differences between the two samples is less than expected by chance. It can easily be shown that, for the two-sample t test, the minimum value of $\Re$ and ϵ^2 is given by $-1/(N-2)$. Thus, for the example data listed in Table 6.4,

$$\min(\Re) = \min(\epsilon^2) = \frac{-1}{N-2} = \frac{-1}{14-2} = -0.0833 .$$

Incidentally, the minimum value for Hays' $\hat{\omega}^2$ measure of effect size is given by $-1/(N-1)$. Thus for the data listed in Table 6.4, the minimum value of Hays' $\hat{\omega}^2$ is $-1/(14-1) = -0.0769$.

6.5.4 An Exact Analysis with $v = 1$

Consider an analysis of the error data listed in Table 6.4 on p. 174 under the Fisher–Pitman permutation model with $v = 1$ and treatment-group weights given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2} .$$

For $v = 1$, the average distance-function values for the two treatment groups are

$$\xi_1 = 9.1429 \quad \text{and} \quad \xi_2 = 9.20 ,$$

respectively, and the observed permutation test statistic is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{8-1}{14-2}\right)(9.1429) + \left(\frac{6-1}{14-2}\right)(9.20) = 9.1667 .$$

There are only

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(8+6)!}{8!\, 6!} = 3003$$

possible, equally-likely arrangements in the reference set of all permutations of the error data listed in Table 6.4, making an exact permutation analysis possible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 9.1667$. There are exactly 2114 δ test statistic values that are equal to or less than the observed value of $\delta = 9.1667$. If all M arrangements of the $N = 14$ observations listed in Table 6.4 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 9.1667$ computed on the $M = 3003$ possible arrangements of the observed data with $n_1 = 8$ and $n_2 = 6$ preserved for each arrangement is

$$P(\delta \leq \delta_o) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{2114}{3003} = 0.7040 ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 14$ observations listed in Table 6.4.

No comparison is made with Student's t test statistic for two independent samples as Student's t is undefined for ordinary Euclidean scaling.

For the example data listed in Table 6.4, the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{26{,}400}{3003} = 8.7912$$

and the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{9.1667}{8.7912} = -0.0427\,,$$

indicating less than chance within-group agreement. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ conventional measures of effect size for two independent samples as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for ordinary Euclidean scaling.

6.5.5 The Effects of Extreme Values

For the example data listed in Table 6.4 on p. 174, the exact probability value employing squared Euclidean scaling with $v = 2$ is $P = 0.5275$ and the exact probability value employing ordinary Euclidean scaling with $v = 1$ is $P = 0.7040$. The difference between the two probability values of

$$\Delta_P = 0.7040 - 0.5275 = 0.1765$$

is entirely due to the squared and non-squared differences obtained with $v = 2$ and $v = 1$, respectively, under the Fisher–Pitman permutation model. Permutation test statistics employing squared Euclidean scaling with $v = 2$ are based on the sample mean ($\bar{x}$) and permutation test statistics employing ordinary Euclidean scaling with $v = 1$ are based on the sample median ($\tilde{x}$). Median-based statistics are highly resistant to extreme values and both treatment Group 1 and treatment Group 2 contain extreme values: $x_{14} = 23$ for Group 1 and $x_{21} = 20$ for Group 2. While these two values are not highly extreme, they are sufficiently removed from their respective mean values of $\bar{x}_1 = 11.00$ and $\bar{x}_2 = 8.00$ to strongly affect the probability value with $v = 2$. Incidentally, the median value for Group 1 is $\tilde{x} = 9.50$ and the median value for Group 2 is $\tilde{x} = 4.50$.

Extreme values are prevalent in applied research. Most variables are not even close to normally distributed and many are highly skewed, often positively. Some examples of positively-skewed variables are family income, net worth, prices of houses sold in a given month, age at first marriage, length of first marriage, and

Table 6.5 Raw-score observed values for two samples with $n_1 = n_2 = 13$ objects randomly assigned to each sample

Sample 1		Sample 2	
Object	Value	Object	Value
1	264.3	1	263.4
2	264.6	2	263.7
3	264.6	3	263.7
4	264.6	4	263.7
5	264.9	5	264.0
6	264.9	6	264.0
7	264.9	7	264.0
8	264.9	8	264.3
9	265.2	9	264.3
10	265.2	10	264.3
11	265.2	11	264.3
12	265.5	12	264.6
13	265.5	13	w

student debt. Consider the case of student debt: in 2017 upon graduation the average student debt was reported to be approximately $34,000, while the median student debt was only approximately $12,000. The mean is pulled higher than the median due to a small proportion of students with substantial debt. Graduate and professional students in veterinary medicine, dental school, law school, and medical school often graduate with hundreds of thousands of dollars in student debt.[4]

To demonstrate the difference between analyses based on squared Euclidean scaling with $v = 2$ and ordinary Euclidean scaling with $v = 1$, consider the two-sample data listed in Table 6.5. While the $n_1 = 13$ values in Sample 1 are fixed, one value in Sample 2, indicated by w, is allowed to vary in order to determine its effect on the exact probability values. Table 6.6 lists 21 values for w ranging from a low value of $w = 40$ up to a high value of $w = 988$, the exact permutation probability values with $v = 1$ and $v = 2$, and the two-tail probability values for Student's two-sample t test, under the usual assumptions of normality, homogeneity, and independence. Each of the exact probability values in Table 6.6 is based on

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(13 + 13)!}{13!\, 13!} = 10{,}400{,}600$$

possible, equally-likely arrangements of the $N = 26$ data values listed in Table 6.5, with the assigned value for w included. The two-tail probability values for the classical two-sample t test listed in Table 6.6 are based on Student's t distribution with $n_1 + n_2 - 2 = 13 + 13 - 2 = 24$ degrees of freedom.

[4] In 2017 the average student debt for law-school graduates was reported to be $141,000 and the average student debt for medical-school graduates was reported to be $192,000.

Table 6.6 Probability value comparisons for exact permutation tests with $v = 1$ and $v = 2$ and the classical Student two-sample t test for the data listed in Table 6.5

	Exact permutation test		
w	$v = 1$	$v = 2$	Student's t test
40	0.4038×10^{-5}	0.4038×10^{-5}	0.3026
80	0.4038×10^{-5}	0.4038×10^{-5}	0.2975
120	0.4038×10^{-5}	0.4038×10^{-5}	0.2895
160	0.4038×10^{-5}	0.4038×10^{-5}	0.2759
200	0.4038×10^{-5}	0.4038×10^{-5}	0.2470
240	0.4038×10^{-5}	0.4038×10^{-5}	0.1481
258	0.4038×10^{-5}	0.4038×10^{-5}	0.8538×10^{-2}
261	0.4038×10^{-5}	0.4038×10^{-5}	0.2646×10^{-3}
264	0.4038×10^{-5}	0.4038×10^{-5}	0.5837×10^{-6}
267	0.9115×10^{-4}	0.0157	0.0159
270	0.9115×10^{-4}	0.4728	0.3455
273	0.9115×10^{-4}	0.9772	0.7459
276	0.9115×10^{-4}	1.0000	1.0000
288	0.9115×10^{-4}	1.0000	0.6222
388	0.9115×10^{-4}	1.0000	0.3753
488	0.9115×10^{-4}	1.0000	0.3533
588	0.9115×10^{-4}	1.0000	0.3451
688	0.9115×10^{-4}	1.0000	0.3409
788	0.9115×10^{-4}	1.0000	0.3382
888	0.9115×10^{-4}	1.0000	0.3365
988	0.9115×10^{-4}	1.0000	0.3352

As illustrated in Table 6.6, the exact probability values for the two-sample permutation test with $v = 1$ are stable, consistent, and relatively unaffected by the extreme values of w in either direction. The small change in the exact probability values from $P = 0.4038\times10^{-5}$ to $P = 0.9115\times10^{-4}$; that is,

$$\Delta_P = 0.9115\times10^{-4} - 0.4038\times10^{-5} = 0.8711\times10^{-4} ,$$

with $v = 1$ occurs when w changes from $w = 264$ to $w = 267$ and passes the median value of $\tilde{x} = 264.9$. In contrast, the exact probability values for the two-sample permutation test with $v = 2$ range from $P = 0.4038\times10^{-5}$ for small values of w up to $P = 1.0000$ for large values of w, relative to the fixed values. Finally, the asymptotic two-tail probability values for the classical two-sample t test approach a common value as w becomes very small or very large, relative to the fixed values, and the classical t test is unable to detect the obvious differences in location between Samples 1 and 2.

6.5.6 *Treatment-Group Weights*

The treatment-group weighting functions with $N - 2$ degrees of freedom given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

are essential for Student's t test for two independent samples, but are not required for a permutation test, as degrees of freedom are irrelevant for nonparametric, distribution-free permutation methods.[5] For a reanalysis of the example data listed in Table 6.4 on p. 174, the treatment-group weighting functions are set to

$$C_1 = \frac{n_1}{N} \quad \text{and} \quad C_2 = \frac{n_2}{N} \, ,$$

simply weighting each treatment group proportional to the number of observations in the group and setting $v = 1$, employing ordinary Euclidean difference between the pairs of values. For the example data listed in Table 6.4 on p. 174 the permutation test statistic is $\delta = 9.1673$.

Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 9.1673$. There are exactly 2127 δ test statistic values that are equal to or less than the observed value of $\delta = 9.1673$. If all M arrangements of the $N = 14$ observed values listed in Table 6.4 on p. 174 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 9.1673$ computed on the $M = 3003$ possible arrangements of the observed data with $n_1 = 8$ and $n_2 = 6$ preserved for each arrangement is

$$P\big(\delta \leq \delta_{\mathrm{o}}\big) = \frac{\text{number of } \delta \text{ values } \leq \delta_{\mathrm{o}}}{M} = \frac{2127}{3003} = 0.7083 \, ,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 14$ observations listed in Table 6.4.

For comparison, the exact probability values based on squared Euclidean scaling with $v = 2$ and ordinary Euclidean scaling with $v = 1$ and

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2} \, ,$$

[5] Degrees of freedom are not relevant for any nonparametric, distribution-free statistic. However, it is noteworthy that degrees of freedom may be required for a test statistic that is nonparametric but is not distribution-free, such as Pearson's χ^2 test statistics for goodness of fit and independence.

are $P = 0.5275$ and $P = 0.7040$, respectively. No comparison is made with Student's two-sample t test for two independent samples as Student's t is undefined for $C_i = n_i/N$, $i = 1, 2$.

The exact expected value of the $M = 3003$ δ test statistic values with $v = 1$ is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{26{,}400}{3003} = 8.7912$$

and the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{9.1673}{8.7912} = -0.0428\,,$$

indicating less than chance within-group agreement. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ conventional measures of effect size for two independent samples as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for $C_i = n_i/N$, $i = 1, 2$.

6.6 Example 4: Exact and Monte Carlo Analyses

For a fourth, larger example of a test for two independent samples, consider the data on $N = 28$ subjects under the Neyman–Pearson population model, randomly divided into two groups of $n_1 = n_2 = 14$ subjects each and listed in Table 6.7. For the example data listed in Table 6.7, the null hypothesis is H_0: $\mu_1 = \mu_2$; that

Table 6.7 Example data for a test of two independent samples with $N = 28$ subjects

Group 1		Group 2	
Case	Value	Case	Value
1	72.87	15	72.92
2	72.78	16	72.86
3	72.61	17	72.85
4	72.55	18	72.80
5	72.53	19	72.74
6	72.50	20	72.73
7	72.47	21	72.69
8	72.47	22	72.66
9	72.44	23	72.66
10	72.42	24	72.62
11	72.38	25	72.57
12	72.31	26	72.51
13	72.17	27	72.36
14	72.14	28	72.25

is, no mean difference is expected between the two populations from which the samples are presumed to have been drawn. The two groups are of equal size with $n_1 = n_2 = 14$, the mean of treatment Group 1 is $\bar{x}_1 = 72.4743$, the mean of treatment Group 2 is $\bar{x}_2 = 72.6586$, the estimated population variance for Group 1 is $s_1^2 = 0.0402$, the estimated population variance for Group 2 is $s_2^2 = 0.0358$, the unbiased pooled estimate of the population variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{N - 2} = \frac{(14 - 1)(0.0402) + (14 - 1)(0.0358)}{28 - 2} = 0.0380 ,$$

and the observed value of Student's t test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{1/2}} = \frac{72.4743 - 72.6586}{\left[0.0380 \left(\frac{1}{14} + \frac{1}{14} \right) \right]^{1/2}} = -2.5011 .$$

Under the Neyman–Pearson null hypothesis, test statistic t is asymptotically distributed as Student's t with $N-2$ degrees of freedom. With $N-2 = 28-2 = 26$ degrees of freedom, the asymptotic two-tail probability value of $t = -2.5011$ is $P = 0.0190$, under the assumptions of normality and homogeneity.

6.6.1 A Monte Carlo Analysis with $v = 2$

Under the Fisher–Pitman permutation model, employing squared Euclidean scaling with $v = 2$ and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

for correspondence with Student's two-sample t test, the average distance-function values for Groups 1 and 2 are

$$\xi_1 = 0.0804 \quad \text{and} \quad \xi_2 = 0.0717 ,$$

respectively, and the observed permutation test statistic is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{14 - 1}{28 - 2} \right) (0.0804) + \left(\frac{14 - 1}{28 - 2} \right) (0.0717) = 0.0760 .$$

Alternatively, in terms of Student's t test statistic,

$$\xi_1 = 2s_1^2 = 2(0.0402) = 0.0804 \,, \quad \xi_2 = 2s_2^2 = 2(0.0358) = 0.0717 \,,$$

and

$$\delta = 2s_p^2 = 2(0.0380) = 0.0760 \,.$$

For the example data listed in Table 6.7, the sum of the $N = 28$ observations is

$$\sum_{i=1}^{N} x_i = 72.87 + 72.78 + \cdots + 72.36 + 72.25 = 2031.8600 \,,$$

the sum of the $N = 28$ squared observations is

$$\sum_{i=1}^{N} x_i^2 = 72.87^2 + 72.78^2 + \cdots + 72.36^2 + 72.25^2 = 147{,}446.0494 \,,$$

and the total sum-of-squares is

$$SS_{\text{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N$$
$$= 147{,}446.0494 - (2031.8600)^2/28 = 1.2258 \,.$$

Based on the expressions given in Eq. (6.5) on p. 159, the observed value for test statistic δ with respect to the observed value of Student's t statistic is

$$\delta = \frac{2SS_{\text{Total}}}{t^2 + N - 2} = \frac{2(1.2258)}{(-2.5011)^2 + 28 - 2} = 0.0760$$

and the observed value for Student's t statistic with respect to the observed value of test statistic δ is

$$t = \left(\frac{2SS_{\text{Total}}}{\delta} - N + 2 \right)^{1/2} = \left[\frac{2(1.2258)}{0.0760} - 28 + 2 \right]^{1/2} = \pm 2.5011 \,.$$

Under the Fisher–Pitman permutation model there are

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(14 + 14)!}{14!\, 14!} = 40{,}116{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the observed data listed in Table 6.7, making an exact permutation analysis impractical. Under the Fisher–Pitman permutation model, the Monte Carlo probability of an observed δ is the proportion of δ test statistic values computed on the randomly-selected, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 0.0760$. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, there are exactly 20,439 δ test statistic values that are equal to or less than the observed value of $\delta = 0.0760$. If all M arrangements of the $N = 28$ observations listed in Table 6.7 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $\delta = 0.0760$ computed on $L = 1{,}000{,}000$ random arrangements of the observed data with $n_1 = n_2 = 14$ preserved for each arrangement is

$$P\left(\delta \leq \delta_{\mathrm{o}}\right) = \frac{\text{number of } \delta \text{ values } \leq \delta_{\mathrm{o}}}{L} = \frac{20{,}439}{1{,}000{,}000} = 0.0204 \, ,$$

where δ_{o} denotes the observed value of test statistic δ and L is the number of randomly-selected, equally-likely arrangements of the $N = 28$ observations listed in Table 6.7. Alternatively, the Monte Carlo probability value of $|t| = 2.5011$ under the Fisher–Pitman null hypothesis is

$$P\left(|t| \geq |t_{\mathrm{o}}|\right) = \frac{\text{number of } |t| \text{ values } \geq |t_{\mathrm{o}}|}{L} = \frac{20{,}439}{1{,}000{,}000} = 0.0204 \, ,$$

where t_{o} denotes the observed value of test statistic t.

For the example data listed in Table 6.7 the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_{\delta} = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{3{,}642{,}715}{40{,}116{,}600} = 0.0908 \, .$$

Alternatively, in terms of an analysis of variance model the exact expected value of test statistic δ is

$$\mu_{\delta} = \frac{2SS_{\mathrm{Total}}}{N-1} = \frac{2(1.2258)}{28-1} = 0.0908 \, ,$$

where

$$SS_{\mathrm{Total}} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2 \Big/ N$$

$$= 147{,}446.0494 - (2031.8600)/28 = 1.2258 \, .$$

Finally, the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.0760}{0.0908} = +0.1629 ,$$

indicating approximately 16% within-group agreement above what is expected by chance.

6.6.2 An Exact Analysis with $v = 2$

While an exact analysis may be impractical with $M = 40{,}116{,}600$ possible arrangements of the observed data, it is not impossible. For an exact test under the Fisher–Pitman permutation model with $v = 2$, the observed value of δ is still $\delta = 0.0760$, the exact expected value of δ under the Fisher–Pitman null hypothesis is $\mu_\delta = 0.0908$, there are exactly 815,878 δ test statistic values that are equal to or less than the observed value of $\delta = 0.0760$, and the exact probability value based on all $M = 40{,}116{,}600$ arrangements of the observed data under the Fisher–Pitman null hypothesis is

$$P\left(\delta \leq \delta_o | H_0\right) = \frac{\text{number of } \delta \text{ values } \leq \delta_o}{M} = \frac{815{,}878}{40{,}116{,}600} = 0.0203 ,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 28$ observations listed in Table 6.7.

Alternatively, the exact two-tail probability value of $|t| = 2.5011$ under the null hypothesis is

$$P\left(|t| \geq |t_o|\right) = \frac{\text{number of } |t| \text{ values } \geq |t_o|}{M} = \frac{815{,}878}{40{,}116{,}600} = 0.0203 ,$$

where t_o denotes the observed value of test statistic t. The observed chance-corrected measure of effect size is unchanged at

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.0760}{0.0908} = +0.1629 ,$$

indicating approximately 16% within-group agreement above what is expected by chance.

6.6.3 Measures of Effect Size

For comparison, for the example data with $N = 28$ observations listed in Table 6.7 Cohen's $\hat{d}$ measure of effect size is

$$\hat{d} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p^2}} = \frac{|72.4743 - 72.6586}{\sqrt{0.0380}} = 0.9454 ,$$

Pearson's r^2 measure of effect size is

$$r^2 = \frac{t^2}{t^2 + N - 2} = \frac{(-2.5011)^2}{(-2.5011)^2 + 28 - 2} = 0.1939 ,$$

Kelley's ϵ^2 measure of effect size is

$$\epsilon^2 = \frac{t^2 - 1}{t^2 + N - 2} = \frac{(-2.5011)^2 - 1}{(-2.5011)^2 + 28 - 2} = 0.1629 ,$$

and Hays' $\hat{\omega}^2$ measure of effect size is

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N - 1} = \frac{(-2.5011)^2 - 1}{(-2.5011)^2 + 28 - 2} = 0.1580 .$$

There is a considerable difference between the value for Cohen's measure of effect size ($\hat{d} = 0.9454$) and the other three measures (Pearson's $r^2 = 0.1939$, Kelley's $\epsilon^2 = 0.1629$, and Hays' $\hat{\omega}^2 = 0.1580$). In general, members of the r family, such as Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$, produce measures of effect size that vary between the limits of 0 and 1, while members of the d family, such as Cohen's $\hat{d}$, produce measures of effect size in standard deviation units and, theoretically, can vary between 0 and ∞.

For comparison purposes, Cohen's $\hat{d}$ measure of effect size can be converted to the r family of measures of effect size. Cohen's $\hat{d}$ can then be compared with Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$. Thus,

$$r^2 = \frac{n_1 n_2 \hat{d}^2}{n_1 n_2 \hat{d}^2 + N(N-2)} = \frac{(14)(14)(0.9454)^2}{(14)(14)(0.9454)^2 + 28(28-2)} = 0.1939 ,$$

which is similar to Kelley's $\epsilon^2 = 0.1629$ and Hays' $\hat{\omega}^2 = 0.1580$.

6.6.4 A Monte Carlo Analysis with $v = 1$

Consider an analysis of the error data listed in Table 6.7 under the Fisher–Pitman permutation model employing ordinary Euclidean scaling with $v = 1$, $N = 28$, and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2} .$$

For $v = 1$, the average distance-function values for treatment Groups 1 and 2 are

$$\xi_1 = 0.2288 \quad \text{and} \quad \xi_2 = 0.2167 ,$$

respectively, and the observed permutation test statistic is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{14 - 1}{28 - 2}\right)(0.2288) + \left(\frac{14 - 1}{28 - 2}\right)(0.2167) = 0.2227 .$$

There are exactly

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(14 + 14)!}{14!\ 14!} = 40{,}116{,}600$$

possible, equally-likely arrangements in the reference set of all permutations of the example data listed in Table 6.7. Under the Fisher–Pitman permutation model the Monte Carlo probability of an observed δ is the proportion of δ test statistic values computed on the randomly-selected, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 0.2227$. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, there are exactly 14,493 δ test statistic values that are equal to or less than the observed value of $\delta = 0.2227$. If all M arrangements of the $N = 28$ observations listed in Table 6.7 occur with equal chance under the Fisher–Pitman null hypothesis, the Monte Carlo probability value of $\delta = 0.2227$ computed on $L = 1{,}000{,}000$ random arrangements of the observed data with $n_1 = n_2 = 14$ preserved for each arrangement is

$$P\left(\delta \le \delta_o | H_0\right) = \frac{\text{number of } \delta \text{values } \le \delta_o}{L} = \frac{14{,}493}{1{,}000{,}000} = 0.0145 ,$$

where δ_o denotes the observed value of test statistic δ and L is the number of randomly-selected, equally-likely arrangements of the $N = 28$ observations listed in Table 6.7. No comparison is made with Student's t test statistic for two independent samples as Student's t is undefined for ordinary Euclidean scaling.

For the data listed in Table 6.7, the exact expected value of test statistic δ under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{10{,}997{,}042}{40{,}116{,}600} = 0.2741\ ,$$

and the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.2227}{0.2741} = +0.1875\ ,$$

indicating approximately 19% within-group agreement above what is expected by chance. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ measures of effect size for two independent samples as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for ordinary Euclidean scaling.

6.6.5 *An Exact Analysis with $v = 1$*

For comparison, with $v = 1$ and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}\ ,$$

the exact probability value of $\delta = 0.2227$ computed on the $M = 40{,}116{,}600$ possible arrangements of the observed data with $n_1 = n_2 = 14$ preserved for each arrangement is

$$P\big(\delta \le \delta_{\mathrm{o}}\big) = \frac{\text{number of } \delta \text{ values } \le \delta_{\mathrm{o}}}{M} = \frac{583{,}424}{40{,}116{,}600} = 0.0145\ ,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 28$ observations listed in Table 6.7 on p. 184.

6.7 Example 5: Rank-Score Permutation Analyses

Oftentimes in conventional research it becomes necessary to analyze rank-score data, either because the observed data are collected as ranks or because the necessary parametric assumptions cannot be met and the raw data are subsequently converted to ranks. There is never any reason to convert raw scores to ranks with permutation statistical methods [3], so this example merely serves to demonstrate the relationship

between a quotidian two-sample test of rank-score data and a permutation test of rank-score data. The conventional approach to rank-score data for two independent samples under the Neyman–Pearson population model is the Wilcoxon–Mann–Whitney (WMW) two-sample rank-sum test.

6.7.1 The Wilcoxon–Mann–Whitney Test

Consider a two-sample rank test for N univariate rank scores with n_1 and n_2 rank scores in the first and second samples, respectively. Under the Neyman–Pearson population model, the WMW two-sample rank-sum test is given by

$$W = \sum_{i=1}^{n_1} R_i \ ,$$

where R_i denotes the rank function of the ith response measurement and n_1 is, typically, the smaller of the two-sample sizes.

For an example analysis of rank-score data, consider the rank scores listed in Table 6.8, where for two samples, $n_1 = 8$, $n_2 = 12$, $N = n_1 + n_2 = 8 + 12 = 20$ total scores, and there are no tied rank scores. For this application, let $n_1 = 8$ denote the rank scores in Sample 1 and $n_2 = 12$ denote the rank scores in Sample 2.

The conventional Wilcoxon–Mann–Whitney two-sample rank-sum test on the $N = 20$ rank scores listed in Table 6.8 yields an observed test statistic value of

$$W = \sum_{i=1}^{n_1} R_i = 1 + 2 + 3 + 4 + 5 + 6 + 8 + 11 = 40 \ ,$$

Table 6.8 Example rank-score data for a conventional Wilcoxon–Mann–Whitney two-sample rank-sum test with $n_1 = 8$ and $n_2 = 12$ subjects

Sample 1		Sample 2	
Subject	Score	Subject	Score
1	1	9	7
2	2	10	9
3	3	11	10
4	4	12	12
5	5	13	13
6	6	14	14
7	8	15	15
8	11	16	16
		17	17
		18	18
		19	19
		20	20

where statistic W is asymptotically distributed $N(0, 1)$ under the Neyman–Pearson null hypothesis as $N \to \infty$. For the rank scores listed in Table 6.8, the mean value of test statistic W is

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{8(20+1)}{2} = 84 \,,$$

the variance of test statistic W is

$$\sigma_W^2 = \frac{n_1 n_2(N+1)}{12} = \frac{(8)(12)(20+1)}{12} = 168 \,,$$

the standard score, corrected for continuity, is[6]

$$z = \frac{W + 0.5 - \mu_W}{\sqrt{\sigma_W^2}} = \frac{40 + 0.5 - 84}{\sqrt{168}} = -3.3561 \,,$$

and the asymptotic two-tail $N(0, 1)$ probability value is $P = 0.3952 \times 10^{-3}$.

6.7.2 *An Exact Analysis with $v = 2$*

For an analysis of the rank-score data listed in Table 6.8 under the Fisher–Pitman permutation model let $v = 2$, employing squared Euclidean differences between the pairs of rank scores, and let the treatment-group weights be given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

for correspondence with the Wilcoxon–Mann–Whitney two-sample rank-sum test. Following Eq. (6.2) on p. 157, the average distance-function values for Samples 1 and 2 are

$$\xi_1 = 21.7143 \quad \text{and} \quad \xi_2 = 33.7576 \,,$$

respectively, and the observed value of the permutation test statistic δ is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{8-1}{20-2}\right)(21.7143) + \left(\frac{12-1}{20-2}\right)(33.7576) = 29.0741 \,.$$

[6] When fitting a continuous mathematical function, such as the normal probability distribution, to a discrete permutation distribution, it is oftentimes necessary to correct the fit by adding or subtracting 0.5 to compensate for the discreteness of the distribution.

Although no self-respecting researcher would seriously consider calculating an estimated population variance on rank scores, since the WMW two-sample rank-sum test is directly derived from Student's two-sample t test, certain relationships still hold. Thus,

$$\xi_1 = 2s_1^2 = 2(10.8571) = 21.7143\ , \quad \xi_2 = 2s_2^2 = 2(16.8788) = 33.7576\ ,$$

and

$$\delta = 2s_p^2 = 2(14.5370) = 29.0741\ ,$$

where $s_1^2 = 10.8571$ and $s_2^2 = 16.8788$ are calculated on the rank-score data listed in Table 6.8.

Because there are only

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(8 + 12)!}{8!\ 12!} = 125{,}970$$

possible, equally-likely arrangements in the reference set of all permutations of the $N = 20$ rank scores listed in Table 6.8, an exact permutation analysis is feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 29.0741$. There are exactly 24 δ test statistic values that are equal to or less than the observed value of $\delta = 29.0741$. If all M arrangements of the $N = 20$ rank scores listed in Table 6.8 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 29.0741$ computed on the $M = 125{,}970$ possible arrangements of the observed data with $n_1 = 8$ and $n_2 = 12$ preserved for each arrangement is

$$P\big(\delta \le \delta_{\mathrm{o}} | H_0\big) = \frac{\text{number of } \delta \text{ values } \le \delta_{\mathrm{o}}}{M} = \frac{24}{125{,}970} = 0.1905{\times}10^{-3}\ ,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 20$ observations listed in Table 6.8.

The functional relationships between test statistics δ and W are given by

$$\delta = \frac{2}{N(N-2)}\left[NT - S^2 - \frac{(NW - n_1 S)^2}{n_1 n_2}\right] \tag{6.22}$$

and

$$W = \frac{n_1 S}{N} - \left\{\frac{n_1 n_2}{N^2}\left[NT - S^2 - \frac{N(N-2)\delta}{2}\right]\right\}^{1/2}\ , \tag{6.23}$$

where

$$S = \sum_{i=1}^{N} R_i \quad \text{and} \quad T = \sum_{i=1}^{N} R_i^2 \, .$$

In the absence of any tied rank scores, it is well known that S and T may simply be expressed as

$$S = \sum_{i=1}^{N} i = \frac{N(N+1)}{2} \quad \text{and} \quad T = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} \, .$$

The relationships between test statistics δ and W are confirmed as follows. For the $N = 20$ rank scores listed in Table 6.8 with no tied values, the observed value of S is

$$S = \sum_{i=1}^{N} i = \frac{N(N+1)}{2} = \frac{20(20+1)}{2} = 210$$

and the observed value of T is

$$T = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} = \frac{20(20+1)[2(20)+1]}{6} = 2870 \, .$$

Then following Eq. (6.22), the observed value of test statistic δ with respect to the observed value of test statistic W for the rank scores listed in Table 6.8 is

$$\begin{aligned} \delta &= \frac{2}{N(N-2)} \left[NT - S^2 - \frac{(NW - n_1 S)^2}{n_1 n_2} \right] \\ &= \frac{2}{20(20-2)} \left\{ 20(2,870) - (210)^2 - \frac{[20(40) - 8(210)]^2}{(8)(12)} \right\} \\ &= \frac{2}{360} \left(13,300 - \frac{774,400}{96} \right) = 29.0741 \end{aligned}$$

and following Eq. (6.23), the observed value of test statistic W with respect to the observed value of test statistic δ is

$$\begin{aligned} W &= \frac{n_1 S}{N} - \left\{ \frac{n_1 n_2}{N^2} \left[NT - S^2 - \frac{N(N-2)\delta}{2} \right] \right\}^{1/2} \\ &= \frac{(8)(210)}{20} - \left\{ \frac{(8)(12)}{20^2} \left[(20)(2870) - (210)^2 - \frac{20(20-2)(29.0741)}{2} \right] \right\} \\ &= 84 - \left[(0.24)(8,066.6667) \right]^{1/2} = 40 \, . \end{aligned}$$

Because of the relationship between test statistics δ and W, the exact probability of $W = 40$ is the same as the exact probability of $\delta = 29.0741$. Thus,

$$P\big(\delta \le \delta_o | H_0\big) = \frac{\text{number of } \delta \text{ values } \le \delta_o}{M} = \frac{24}{125{,}970} = 0.1905{\times}10^{-3}$$

and

$$P\big(W \ge W_o | H_0\big) = \frac{\text{number of } W \text{ values } \ge W_o}{M} = \frac{24}{125{,}970} = 0.1905{\times}10^{-3} \,,$$

where δ_o and W_o denote the observed values of test statistics δ and W, respectively, and M is the number of possible, equally-likely arrangements of the $N = 20$ rank scores listed in Table 6.8.

Following Eq. (6.7) on p. 160, the exact expected value of the $M = 125{,}970$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{8{,}817{,}900}{125{,}970} = 70.00$$

and following Eq. (6.6) on p. 160, the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{29.0741}{70.00} = +0.5847 \,,$$

indicating approximately 58% within-group agreement above what is expected by chance. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ measures of effect size for two independent samples as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for rank-score data.

6.7.3 An Exact Analysis with $v = 1$

For a reanalysis of the rank-score data listed in Table 6.8 on p. 192 under the Fisher–Pitman permutation model let $v = 1$ instead of $v = 2$, employing ordinary Euclidean differences between the pairs of rank scores, and let the treatment-group weights be given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2} \,.$$

Following Eq. (6.2) on p. 157, the average distance-function values for Samples 1 and 2 are

$$\xi_1 = 3.9286 \quad \text{and} \quad \xi_2 = 4.9091 \,,$$

respectively, and the observed value of test statistic δ is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{8-1}{20-2}\right)(3.9286) + \left(\frac{12-1}{20-2}\right)(4.9091) = 4.5278 \,.$$

Because there are only

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(8+12)!}{8!\ 12!} = 125{,}970$$

possible, equally-likely arrangements in the reference set of all permutations of the $N = 20$ rank scores listed in Table 6.8, an exact permutation analysis is feasible. Under the Fisher–Pitman permutation model, the exact probability of an observed δ is the proportion of δ test statistic values computed on all possible, equally-likely arrangements of the observed data that are equal to or less than the observed value of $\delta = 4.5278$. There are exactly 24 δ test statistic values that are equal to or less than the observed value of $\delta = 4.5278$. If all M arrangements of the $N = 20$ rank scores listed in Table 6.8 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 4.5278$ computed on the $M = 125{,}970$ possible arrangements of the observed data with $n_1 = 8$ and $n_2 = 12$ preserved for each arrangement is

$$P\big(\delta \le \delta_{\mathrm{o}} | H_0\big) = \frac{\text{number of } \delta \text{ values } \le \delta_{\mathrm{o}}}{M} = \frac{24}{125{,}970} = 0.1905{\times}10^{-3} \,,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 20$ observations listed in Table 6.8. For comparison, the exact probability value based on $v = 2$, $M = 125{,}970$, and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

in the previous analysis was also $P = 0.1905{\times}10^{-3}$. No comparison is made with the conventional Wilcoxon–Mann–Whitney two-sample rank-sum test as the WMW test is undefined for ordinary Euclidean scaling.

Following Eq. (6.7) on p. 160, the exact expected value of the $M = 125{,}970$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{881{,}790}{125{,}970} = 7.00$$

and, following Eq. (6.6) on p. 160, the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{4.5278}{7.00} = +0.3532\,,$$

indicating approximately 35% within-group agreement above what is expected by chance. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ measures of effect size for two independent samples as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for rank-score data.

Finally, it should be noted that for the example data listed in Table 6.8 the observed ξ_1 and ξ_2 values differ for $v = 2$ ($\xi_1 = 21.7143$ and $\xi_2 = 33.7576$) and $v = 1$ ($\xi_1 = 3.9286$ and $\xi_2 = 4.9091$), the observed δ values differ for $v = 2$ ($\delta = 29.0741$) and $v = 1$ ($\delta = 4.5278$), and the exact values for μ_δ also differ for $v = 2$ ($\mu_\delta = 70.00$) and $v = 1$ ($\mu_\delta = 7.00$). However, the probability values for $v = 2$ ($P = 0.1905\times10^{-3}$) and $v = 1$ ($P = 0.1905\times10^{-3}$) do not differ. This is always true for two-sample tests of rank scores under the Fisher–Pitman permutation model. Unlike two-sample tests of raw (interval-level) values, there is never any difference in probability values for $v = 2$ and $v = 1$ with rank-score (ordinal-level) data.

6.8 Example 6: Multivariate Permutation Analyses

Oftentimes a research design calls for a test of difference between two independent treatment groups when $r \geq 2$ response measurements have been obtained for each subject. The conventional approach to such a research design under the Neyman–Pearson population model is Hotelling's multivariate T^2 test for two independent samples given by

$$T^2 = \frac{n_1 n_2}{N}(\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_2)\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\,, \tag{6.24}$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ denote vectors of mean differences between treatment Groups 1 and 2, respectively, n_1 and n_2 are the number of multivariate measurement scores in treatment Groups 1 and 2, respectively, $N = n_1 + n_2$, and $\mathbf{S}$ is a variance–covariance

matrix given by

$$\begin{bmatrix} \frac{1}{N}\sum_{I=1}^{N}(y_{1I}-\bar{y}_1)^2 & \cdots & \frac{1}{N}\sum_{I=1}^{N}(y_{1I}-\bar{y}_1)(y_{rI}-\bar{y}_r) \\ \vdots & & \vdots \\ \frac{1}{N}\sum_{I=1}^{N}(y_{rI}-\bar{y}_r)(y_{1I}-\bar{y}_1) & \cdots & \frac{1}{N}\sum_{I=1}^{N}(y_{rI}-\bar{y}_r)^2 \end{bmatrix} .$$

The observed value of Hotelling's T^2 is conventionally transformed into an F test statistic by

$$F = \frac{N-r-1}{r(N-2)}\,T^2 ,$$

which is asymptotically distributed as Snedecor's F under the Neyman–Pearson null hypothesis with $\nu_1 = r$ and $\nu_2 = N - r - 1$ degrees of freedom.

6.8.1 The Hotelling Two-Sample T^2 Test

To illustrate a conventional multivariate analysis under the Neyman–Pearson population model, consider the multivariate measurement scores listed in Table 6.9, where $r = 2$, $n_1 = 4$, $n_2 = 6$, and $N = n_1 + n_2 = 4 + 6 = 10$.

A conventional two-sample Hotelling T^2 test of the $N = 10$ multivariate measurement scores listed in Table 6.9 yields

$$\bar{y}_{11} = 2.7750 ,$$
$$s_{11}^2 = 3.1092 ,$$
$$\bar{y}_{12} = 4.5250 ,$$
$$s_{12}^2 = 5.1892 ,$$

Table 6.9 Example multivariate response measurement scores with $r = 2$, $n_1 = 4$, $n_2 = 6$, and $N = n_1 + n_2 = 10$

Treatment	
1	2
(1.2, 3.1)	(3.7, 6.1)
(2.9, 6.8)	(6.1, 8.3)
(1.8, 2.1)	(6.2, 7.9)
(5.2, 6.1)	(4.8, 9.7)
	(5.1, 9.9)
	(4.2, 7.8)

$$\begin{aligned} \text{cov}(1,2)_1 &= +2.9042\,, \\ \bar{y}_{21} &= 5.0167 \\ s_{21}^2 &= 1.0057\,, \\ \bar{y}_{22} &= 8.2833\,, \\ s_{22}^2 &= 1.9537\,, \end{aligned}$$

and

$$\text{cov}(1,2)_2 = +0.5323\,.$$

Then the vector of mean differences for treatment Group 1 is

$$\bar{\mathbf{y}}_1 = \bar{y}_{11} - \bar{y}_{21} = 2.7550 - 5.0167 = -2.2417$$

and the vector of mean differences for treatment Group 2 is

$$\bar{\mathbf{y}}_2 = \bar{y}_{12} - \bar{y}_{22} = 4.5250 - 8.2833 = -3.7583\,.$$

The variance–covariance matrices for Treatments 1 and 2 are

$$\hat{\mathbf{\Sigma}}_1 = \begin{bmatrix} 3.1092 & +2.9042 \\ +2.9042 & 5.1892 \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{\Sigma}}_2 = \begin{bmatrix} 1.0057 & +0.5323 \\ +0.5323 & 1.9537 \end{bmatrix},$$

respectively, and the pooled variance–covariance matrix and its inverse are

$$\mathbf{S} = \begin{bmatrix} 1.7945 & +1.4218 \\ +1.4218 & 3.1670 \end{bmatrix} \quad \text{and} \quad \mathbf{S}^{-1} = \begin{bmatrix} +0.8649 & -0.3883 \\ -0.3883 & +0.4901 \end{bmatrix},$$

respectively.

Following Eq. (6.24) on p. 198, the observed value of Hotelling's T^2 is

$$\begin{aligned} T^2 &= \frac{n_1 n_2}{N} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \\ &= \frac{(4)(6)}{10} \begin{bmatrix} -2.2417 & -3.7583 \end{bmatrix} \begin{bmatrix} +0.8649 & -0.3883 \\ -0.3883 & +0.4901 \end{bmatrix} \begin{bmatrix} -2.2417 \\ -3.7583 \end{bmatrix} \\ &= (2.40)(4.7260) = 11.3423 \end{aligned}$$

and the F test statistic for Hotelling's T^2 is

$$F = \frac{N-r-1}{r(N-2)}\, T_{\mathrm{o}}^2 = \frac{10-2-1}{2(10-2)}\,(11.3423) = 4.9623\ ,$$

where T_{o}^2 denotes the observed value of Hotelling's T^2.

Assuming independence, normality, homogeneity of variance, and homogeneity of covariance, Hotelling's F test statistic is asymptotically distributed as Snedecor's F with $\nu_1 = r = 2$ and $\nu_2 = N-r-1 = 10-2-1 = 7$ degrees of freedom. Under the Neyman–Pearson null hypothesis, the observed value of $F = 4.9623$ yields an asymptotic probability value of $P = 0.0455$.

6.8.2 An Exact Analysis with $v = 2$

For an analysis under the Fisher–Pitman permutation model let $v = 2$, employing squared Euclidean differences between pairs of measurement scores and let the treatment-group weights be given by

$$C_1 = \frac{n_1 - 1}{N-2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N-2}$$

for correspondence with Hotelling's T^2 test for two independent samples. Since there are only

$$M = \frac{(n_1+n_2)!}{n_1!\, n_2!} = \frac{(4+6)!}{4!\, 6!} = 210$$

possible, equally-likely arrangements in the reference set of all permutations of the $N = 10$ multivariate measurement scores listed in Table 6.9, an exact permutation analysis is feasible. The multivariate measurement scores listed in Table 6.9 yield average distance-function values for Treatments 1 and 2 of

$$\xi_1 = 0.4862 \quad \text{and} \quad \xi_2 = 0.2737\ ,$$

respectively, and the observed permutation test statistic is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{4-1}{10-2}\right)(0.4862) + \left(\frac{6-1}{10-2}\right)(0.2737) = 0.3534\ .$$

If all M arrangements of the $N = 10$ observed multivariate measurement scores listed in Table 6.9 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 0.3534$ computed on the $M = 210$ possible

arrangements of the observed data with $n_1 = 4$ and $n_2 = 6$ scores preserved for each arrangement is

$$P\big(\delta \le \delta_o | H_0\big) = \frac{\text{number of } \delta \text{ values } \le \delta_o}{M} = \frac{12}{210} = 0.0571\,,$$

where δ_o denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 10$ multivariate observations listed in Table 6.9.

Following Eq. (6.7) on p. 160, the exact expected value of the $M = 210$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M}\sum_{i=1}^{M}\delta_i = \frac{93.3333}{210} = 0.4444$$

and following Eq. (6.6) on p. 160, the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.3534}{0.4444} = +0.2049\,,$$

indicating approximately 20% within-group agreement above what is expected by chance. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ measures of effect size for two-sample tests as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for multivariate data.

The identity relating Hotelling's two-sample T^2 test and the permutation test statistic is given by

$$\delta = \frac{2\big[r - V^{(s)}\big]}{N - 2}\,, \tag{6.25}$$

where

$$V^{(s)} = \frac{T^2}{T^2 + N - 2} \tag{6.26}$$

and $s = \min(g - 1, r)$; in this case with $g - 1 = 2 - 1 = 1$ and $r = 2$, $s = \min(2 - 1, 2) = 1$. Thus, following Eqs. (6.25) and (6.26), the observed value of $V^{(1)}$ is

$$V^{(1)} = \frac{T^2}{T^2 + N - 2} = \frac{11.3423}{11.3423 + 10 - 2} = \frac{11.3423}{19.3423} = 0.5864$$

and the observed value of δ is

$$\delta = \frac{2(r - V^{(s)})}{N - g} = \frac{2(2 - 0.5864)}{10 - 2} = \frac{2.8272}{8} = 0.3534\,.$$

6.8.3 An Exact Analysis with $v = 1$

Under the Fisher–Pitman permutation model, it is not necessary to set $v = 2$, thereby illuminating the squared differences between pairs of measurement scores. For a reanalysis of the measurement scores listed in Table 6.9 on p. 199, let the treatment-group weights be given by

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

as in the previous example, but set $v = 1$ instead of $v = 2$, employing ordinary Euclidean differences between pairs of measurement scores. Following Eq. (6.2) on p. 157, the average distance-function values for Treatments 1 and 2 are

$$\xi_1 = 3.7865 \quad \text{and} \quad \xi_2 = 2.2200 \,,$$

respectively, and following Eq. (6.1) on p. 157, the observed value of the permutation test statistic is

$$\delta = \sum_{i=1}^{2} C_i \xi_i = \left(\frac{4-1}{10-2}\right)(3.7865) + \left(\frac{6-1}{10-2}\right)(2.2200) = 2.8074 \,.$$

If all M arrangements of the $N = 10$ observed multivariate measurement scores listed in Table 6.9 occur with equal chance under the Fisher–Pitman null hypothesis, the exact probability value of $\delta = 2.8074$ computed on the $M = 210$ possible arrangements of the observed data with $n_1 = 4$ and $n_2 = 6$ measurement scores preserved for each arrangement is

$$P\big(\delta \le \delta_{\mathrm{o}} | H_0\big) = \frac{\text{number of } \delta \text{ values } \le \delta_{\mathrm{o}}}{M} = \frac{4}{210} = 0.0190 \,,$$

where δ_{o} denotes the observed value of test statistic δ and M is the number of possible, equally-likely arrangements of the $N = 10$ multivariate observations listed in Table 6.9. For comparison, the exact probability value based on $v = 2$, $M = 210$, and treatment-group weights

$$C_1 = \frac{n_1 - 1}{N - 2} \quad \text{and} \quad C_2 = \frac{n_2 - 1}{N - 2}$$

in the previous example is $P = 0.0571$. No comparison is made with Hotelling's multivariate two-sample T^2 test as T^2 is undefined for ordinary Euclidean scaling.

Following Eq. (6.7) on p. 160, the exact expected value of the $M = 210$ δ test statistic values under the Fisher–Pitman null hypothesis is

$$\mu_\delta = \frac{1}{M} \sum_{i=1}^{M} \delta_i = \frac{790.1880}{210} = 3.7628$$

and, following Eq. (6.6) on p. 160, the observed chance-corrected measure of effect size is

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{2.8074}{3.7628} = +0.2539\,,$$

indicating approximately 25% within-group agreement above what is expected by chance. No comparisons are made with Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, or Hays' $\hat{\omega}^2$ measures of effect size for two-sample tests as $\hat{d}$, r^2, ϵ^2, and $\hat{\omega}^2$ are undefined for multivariate data.

6.9 Summary

This chapter examined tests for two independent samples where the null hypothesis under the Neyman–Pearson population model typically posits no difference between the means of two populations; that is, H_0: $\mu_1 = \mu_2$. The conventional tests for two independent samples and four measures of effect size under the Neyman–Pearson population model were described and illustrated: Student's two-sample t test and Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$ measures of effect size, respectively.

Under the Fisher–Pitman permutation model, test statistic δ and associated measure of effect size $\Re$ were introduced and illustrated for tests of two independent samples. Test statistic δ was related to Student's t test statistic and shown to be flexible enough to incorporate either ordinary or squared Euclidean scaling with $v = 1$ and $v = 2$, respectively. Effect-size measure $\Re$ was shown to be applicable to either $v = 1$ or $v = 2$ without modification and to have a clear and meaningful chance-corrected interpretation.

Six examples illustrated permutation statistics δ and $\Re$. In the first example, a small sample of $N = 7$ values was utilized to describe and illustrate the calculations of δ and $\Re$ for two independent samples. The second example demonstrated the permutation-based, chance-corrected measure of effect size, $\Re$, and related $\Re$ to the four conventional measures of effect size for two independent samples: Cohen's $\hat{d}$, Pearson's r^2, Kelley's ϵ^2, and Hays' $\hat{\omega}^2$. The third example with $N = 14$ values was designed to illustrate the effects of extreme values on both conventional and permutation tests for two independent samples. The fourth example utilized a larger sample with $N = 28$ observations to compare and contrast exact and Monte Carlo permutation tests for two independent samples. The fifth example applied permutation methods to univariate rank-score data and compared the permutation results with conventional results from the Wilcoxon–Mann–Whitney two-sample rank-sum test. Finally, the sixth example illustrated the application of permutation methods to multivariate data and compared the permutation results with conventional results from Hotelling's T^2 test for two independent samples.

Chapter 7 continues the presentation of permutation statistical methods for two samples, but examines research designs in which the subjects in the two samples have been matched on specific characteristics; that is to say, not independent. Research designs that posit no mean difference between two matched treatment groups in which univariate measurements have been obtained are ubiquitous in the statistical literature. Matched-pairs tests are the simplest of the tests in an extensive class of randomized-blocks tests and are taught in every introductory course.

References

1. Barnard, G.A.: 2 × 2 tables. A note on E. S. Pearson's paper. Biometrika **34**, 168–169 (1947)
2. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. **70**, 213–220 (1968)
3. Feinstein, A.R.: Clinical biostatistics XXIII: the role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2). Clin. Pharmacol. Ther. **14**, 898–915 (1973)
4. Hotelling, H., Pabst, M.R.: Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Stat. **7**, 29–43 (1936)
5. Johnston, J.E., Berry, K.J., Mielke, P.W.: A measure of effect size for experimental designs with heterogeneous variances. Percept. Motor Skill. **98**, 3–18 (2004)
6. Johnston, J.E., Berry, K.J., Mielke, P.W.: Permutation tests: precision in estimating probability values. Percept. Motor Skill. **105**, 915–920 (2007)
7. Kendall, M.G., Babington Smith, B.: On the method of paired comparisons. Biometrika **31**, 324–345 (1940)
8. Macdonell, W.R.: On criminal anthropometry and the identification of criminals. Biometrika **1**, 177–227 (1902)
9. Maxwell, S.E., Camp, C.J., Arvey, R.D.: Measures of strength of association: a comparative examination. J. Appl. Psychol. **66**, 525–534 (1981)
10. McHugh, R.B., Mielke, P.W.: Negative variance estimates and statistical dependence in nested sampling. J. Am. Stat. Assoc. **63**, 1000–1003 (1968)
11. Mielke, P.W., Berry, K.J., Johnson, E.S.: Multi-response permutation procedures for a priori classifications. Commun. Stat. Theor. Methods **5**, 1409–1424 (1976)
12. Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Public Opin. Quart. **19**, 321–325 (1955)
13. Spearman, C.E.: 'Footrule' for measuring correlation. Brit. J. Psychol. **2**, 89–108 (1906)
14. Stigler, S.M.: The Seven Pillars of Statistical Wisdom. Harvard University Press, Cambridge (2016)