Introduction 1

Permutation statistical methods are a paradox of old and new. While permutation methods pre-date many traditional parametric statistical methods, only recently have permutation methods become part of the mainstream discussion regarding statistical testing. Permutation statistical methods follow a permutation model whereby a test statistic is computed on the observed data, then (1) the observed data are permuted over all possible arrangements of the observations—an exact permutation test, (2) the observed data are used for calculating the exact moments of the underlying discrete permutation distribution and the moments are fitted to an associated continuous distribution—a moment-approximation permutation test, or (3) the observed data are permuted over a random subset of all possible arrangements of the observations—a resampling-approximation permutation test [977, pp. 216–218].

1.1 Overview of This Chapter

This first chapter begins with a brief description of the advantages of permutation methods from statisticians who were, or are, advocates of permutation tests, followed by a description of the methods of permutation tests including exact, moment-approximation, and resampling-approximation permutation tests. The chapter continues with an example that contrasts the well-known Student t test and results from exact, moment-approximation, and resampling-approximation permutation tests using historical data. The chapter concludes with brief overviews of the remaining chapters.

Permutation tests are often described as the gold standard against which conventional parametric tests are tested and evaluated. Bakeman, Robinson, and Quera remarked that "like Read and Cressie (1988), we think permutation tests represent the standard against which asymptotic tests must be judged" [50, p. 6]. Edgington and Onghena opined that "randomization tests . . . have come to be recognized by many in the field of medicine as the 'gold standard' of statistical tests for randomized experiments" [396, p. 9]; Friedman, in comparing tests of significance

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_1,

for m rankings, referred to an exact permutation test as "the correct one" [486, p. 88]; Feinstein remarked that conventional statistical tests "yield reasonably reliable approximations of the more exact results provided by permutation procedures" [421, p. 912]; and Good noted that Fisher himself regarded randomization as a technique for validating tests of significance, i.e., making sure that conventional probability values were accurate [521, p. 263].

Early statisticians understood well the value of permutation statistical tests even during the period in which the computationally-intensive nature of the tests made them impractical. Notably, in 1955 Kempthorne wrote that "[t]ests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas their validity stems from randomization theory" [719, p. 947] and

> [w]hen one considers the whole problem of experimental inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using method of inference other randomization analysis [719, p. 966].

In 1966 Kempthorne re-emphasized that "the proper way to make tests of significance in the simple randomized experiments is by way of the randomization (or permutation) test" [720, p. 20] and "in the randomized experiment one should, logically, make tests of significance by way of the randomization test" [720, p. 21].[1] Similarly, in 1959 Scheffé stated that the conventional analysis of variance F test "can often be regarded as a good approximation to a permutation [randomization] test, which is an exact test under a less restrictive model" [1232, p. 313]. In 1968 Bradley indicated that "eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision" [201, p. 85].

With the advent of high-speed computing, permutation tests became more practical and researchers increasingly appreciated the benefits of the randomization model. In 1998, Ludbrook and Dudley stated that "it is our thesis that the randomization rather than the population model applies, and that the statistical procedures best adapted to this model are those based on permutation" [856, p. 127], concluding that "statistical inferences from the experiments are valid only under the randomization model of inference" [856, p. 131].

In 2000, Bergmann, Ludbrook, and Dudley, in a cogent analysis of the Wilcoxon–Mann–Whitney two-sample rank-sum test, observed that "the only accurate form of the Wilcoxon–Mann–Whitney procedure is one in which the exact permutation null distribution is compiled for the actual data" [100, p. 72] and concluded:

> [o]n theoretical grounds, it is clear that the only infallible way of executing the [Wilcoxon–Mann–Whitney] test is to compile the null distribution of the rank-sum statistic by exact permutation. This was, in effect, Wilcoxon's (1945) thesis and it provided the theoretical basis for his [two-sample rank-sum] test [100, p. 76].

[1] The terms "permutation test" and "randomization test" are often used interchangeably.

1.2 Two Models of Statistical Inference

Essentially, two models of statistical inference coexist: the population model and the permutation model; see for further discussion, articles by Curran-Everett [307], Hubbard [663], Kempthorne [721], Kennedy [748], Lachin [787], Ludbrook [849, 850], and Ludbrook and Dudley [854]. The population model, formally proposed by Jerzy Neyman and Egon Pearson in 1928 [1035, 1036], assumes random sampling from one or more specified populations. Under the population model, the level of statistical significance that results from applying a statistical test to the results of an experiment or a survey corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings from the same specified population(s). Because repeated sampling of the true population(s) is usually impractical, it is assumed that the sampling distribution of the test statistics generated under repeated random sampling conforms to an assumed, conjectured, hypothetical distribution, such as the normal distribution.

The size of a statistical test, e.g., 0.05, is the probability under a specified null hypothesis that repeated outcomes based on random samples of the same size are equal to or more extreme than the observed outcome. In the population model, assignment of treatments to subjects is viewed as fixed with the stochastic element taking the form of an error that would vary if the experiment was repeated [748]. Probability values are then calculated based on the potential outcomes of conceptual repeated draws of these errors. The model is sometimes referred to as the "conditional-on-assignment" model, as the distribution used for structuring the test is conditional on the treatment assignment of the observed sample; see for example, a comprehensive and informative 1995 article by Peter Kennedy in *Journal of Business & Economic Statistics* [748].

The permutation model was introduced by R.A. Fisher in 1925 [448] and further developed by R.C. Geary in 1927 [500], T. Eden and F. Yates in 1933 [379], and E.J.G. Pitman in 1937 and 1938 [1129–1131]. Permutation tests do not refer to any particular statistical tests, but to a general method of determining probability values. In a permutation statistical test the only assumption made is that experimental variability has caused the observed result. That assumption, or null hypothesis, is then tested. The smaller the probability, the stronger is the evidence against the assumption [648]. Under the permutation model, a permutation test statistic is computed for the observed data, then the observations are permuted over all possible arrangements of the observations and the test statistic is computed for each equally-likely arrangement of the observed data [307]. For clarification, an ordered sequence of n exchangeable objects $(\omega_1, \ldots, \omega_n)$ yields $n!$ equally-likely arrangements of the n objects, *vide infra*. The proportion of cases with test statistic values equal to or more extreme than the observed case yields the probability of the observed test statistic. In contrast to the population model, the assignment of errors to subjects is viewed as fixed, with the stochastic element taking the form of the assignment of treatments to subjects for each arrangement [748]. Probability values are then calculated according to all outcomes associated with assignments

of treatments to subjects for each case. This model is sometimes referred to as the "conditional-on-errors" model, as the distribution used for structuring the test is conditional on the individual errors drawn for the observed sample; see for example, a 1995 article by Peter Kennedy [748].

Exchangeability

A sufficient condition for a permutation test is the exchangeability of the random variables. Sequences that are independent and identically distributed (i.i.d.) are always exchangeable, but so is sampling without replacement from a finite population. However, while i.i.d. implies exchangeability, exchangeability does not imply i.i.d. [528, 601, 758]. Diaconis and Freedman present a readable discussion of exchangeability using urns and colored balls [346].

More formally, variables $X_1, X_2, \ldots, X_n$ are exchangeable if

$$P\left[\bigcap_{i=1}^{n}(X_i \leq x_i)\right] = P\left[\bigcap_{i=1}^{n}(X_i \leq x_{c_i})\right],$$

where $x_1, x_2, \ldots, x_n$ are n observed values and $\{c_1, c_2, \ldots, c_n\}$ is any one of the $n!$ equally-likely permutations of $\{1, 2, \ldots, n\}$ [1215].

1.3 Permutation Tests

Three types of permutation tests are common: exact, moment-approximation, and resampling-approximation permutation tests. While the three types are methodologically quite different, all three approaches are based on the same specified null hypothesis.

1.3.1 Exact Permutation Tests

Exact permutation tests enumerate all equally-likely arrangements of the observed data. For each arrangement, the desired test statistic is calculated. The obtained data yield the observed value of the test statistic. The probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the enumerated test statistics with values equal to or more extreme than the value of the observed test statistic. As sample sizes increase, the number of possible arrangements can become very large and exact methods become impractical. For example, permuting two small samples of sizes $n_1 = n_2 = 20$ yields

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(20 + 20)!}{(20!)^2} = 137{,}846{,}528{,}820$$

different arrangements of the observed data.

1.3.2 Moment-Approximation Permutation Tests

The moment-approximation of a test statistic requires computation of the exact moments of the test statistic, assuming equally-likely arrangements of the observed data. The moments are then used to fit a specified distribution. For example, the first three exact moments may be used to fit a Pearson type III distribution. Then, the Pearson type III distribution approximates the underlying discrete permutation distribution and provides an approximate probability value. For many years moment-approximation permutation tests provided an important intermediary approximation when computers lacked both the speed and the storage for calculating exact permutation tests. More recently, resampling-approximation permutation tests have largely replaced moment-approximation permutation tests, except when either the size of the data set is very large or the probability of the observed test statistic is very small.

1.3.3 Resampling-Approximation Permutation Tests

Resampling-approximation permutation tests generate and examine a Monte Carlo random subset of all possible equally-likely arrangements of the observed data. In the case of a resampling-approximation permutation test, the probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the resampled test statistics with values equal to or more extreme than the value of the observed test statistic [368, 649]. Thus, resampling permutation probability values are computationally quite similar to exact permutation tests, but the number of resamplings to be considered is decided upon by the researcher rather than by considering all possible arrangements of the observed data. With sufficient resamplings, a researcher can compute a probability value to any accuracy desired. Read and Cressie [1157], Bakeman, Robinson, and Quera [50], and Edgington and Onghena [396, p. 9] described permutation methods as the "gold standard" against which asymptotic methods must be judged. Tukey took it one step further, labeling resampling permutation methods the "platinum standard" of permutation methods [216, 1381, 1382].[2]

1.3.4 Compared with Parametric Tests

Permutation tests differ from traditional parametric tests based on an assumed population model in several ways.

[2] In a reversal Tukey could not have predicted, at the time of this writing gold was trading at $1,775 per troy ounce, while platinum was only $1,712 per troy ounce [275].

1. Permutation tests are data dependent, in that all the information required for analysis is contained within the observed data set; see a 2007 discussion by Mielke and Berry [965, p. 3].[3]
2. Permutation tests do not assume an underlying theoretical distribution; see a 1983 article by Gabriel and Hall [489].
3. Permutation tests do not depend on the assumptions associated with traditional parametric tests, such as normality and homogeneity; see articles by Kennedy in 1995 [748] and Berry, Mielke, and Mielke in 2002 [162].[4]
4. Permutation tests provide probability values based on the discrete permutation distribution of equally-likely test statistic values, rather than an approximate probability value based on a conjectured theoretical distribution, such as a normal, chi-squared, or F distribution; see a 2001 article by Berry, Johnston, and Mielke [117].
5. Whereas permutation tests are suitable when a random sample is obtained from a designated population, permutation tests are also appropriate for nonrandom samples, such as are common in biomedical research; see discussions by Kempthorne in 1977 [721], Gabriel and Hall in 1983 [489], Bear in 1995 [88], Frick in 1998 [482], Ludbrook and Dudley in 1998 [856], and Edgington and Onghena in 2007 [396, pp. 6–8].
6. Permutation tests are appropriate when analyzing entire populations, as permutation tests are not predicated on repeated random sampling from a specified population; see discussions by Ludbrook and Dudley in 1998 [856], Holford in 2003 [638], and Edgington and Onghena in 2007 [396, pp. 1–8].
7. Permutation tests can be defined for any selected test statistic; thus, researchers have the option of using a wide variety of test statistics, including the majority of statistics commonly utilized in traditional statistical approaches; see discussions by Mielke and Berry in 2007 [965].
8. Permutation tests are ideal for very small data sets, when conjectured, hypothetical distribution functions may provide very poor fits; see a 1998 article by Ludbrook and Dudley [856].
9. Appropriate permutation tests are resistant to extreme values, such as are common in demographic data, e.g., income, age at first marriage, number of children, and so on; see a discussion by Mielke and Berry in 2007 [965, pp. 52–53] and an article by Mielke, Berry, and Johnston in 2011 [978]. Consequently, the need for any data transformation is mitigated in the permutation context and in general is not recommended, e.g., square root, logarithmic, the use of

[3]Echoing Fisher's argument that inference must be based solely on the data at hand [460], Haber refers to data dependency as "the data at hand principle" [565, p. 148].

[4]Barton and David noted that it is desirable to make the minimum of assumptions, since, witness the oft-cited Bertrand paradox [163], that the assumptions made will often prejudice the conclusions reached [83, p. 455].

rank-order statistics,[5] and the choice of a distance function, in particular, may be very misleading [978].

10. Permutation tests provide data-dependent statistical inferences only to the actual experiment or survey that has been performed, and are not dependent on a contrived super population; see for example, discussions by Feinstein in 1973 [421] and Edgington and Onghena in 2007 [396, pp. 7–8].

1.3.5 The Bootstrap and the Jackknife

This chronicle is confined to permutation methods, although many researchers consider that permutation methods, bootstrapping, and the jackknife are closely related. Traditionally, jackknife (leave-one-out) methods have been used to reduce bias in small samples, calculate confidence intervals around parameter estimates, and test hypotheses [789, 876, 1376], while bootstrap methods have been used to estimate standard errors in cases where the distribution of the data is unknown [789]. In general, permutation methods are considered to be more powerful than either the bootstrap or (possibly) the jackknife approaches [789].

While permutation methods and bootstrapping both involve computing simulations, and the rejection of the null hypothesis occurs when a common test statistic is extreme under both bootstrapping and permutation, they are conceptually and mechanically quite different. On the other hand, they do have some similarities, including equivalence in an asymptotic sense [358, 1189]. The two approaches differ in their distinct sampling methods. In resampling, a "new" sample is obtained by drawing the data without replacement, whereas in bootstrapping a "new" sample is obtained by drawing from the data with replacement [748, 1189]. Thus, bootstrapping and resampling are associated with sampling with and without replacement, respectively. Philip Good has been reported as saying that the difference between permutation tests and bootstrap tests is that "[p]ermutations test hypotheses concerning distributions; bootstraps test hypotheses concerning parameters."

Specifically, resampling is a data-dependent procedure, dealing with all finite arrangements of the observed data, and based on sampling without replacement. In contrast, bootstrapping involves repeated sampling from a finite population that conceptually yields an induced infinite population based on sampling with replacement. In addition, when bootstrapping is used with small samples it is necessary to make complex adjustments to control the risk of error; see for example, discussions by Hall and Wilson in 1991 [577], Efron and Tibshirani in 1993 [402], and Westfall and Young, also in 1993 [1437]. Finally, the bootstrap distribution may be viewed as an unconditional approximation to the null distribution of the

[5] Rank-order statistics were among the earliest permutation tests, transforming the observed data into ranks, e.g., from smallest to largest. While they were an important step in the history of permutation tests, modern computing has superseded the need for rank-order tests in the majority of cases.

test statistic, while the resampling distribution may be viewed as a conditional distribution of the test statistic [1189].

In 1991 Donegani argued that it is preferable to compute a permutation test based on sampling without replacement (i.e., resampling) than with replacement (i.e., bootstrap), although, as he noted, the two techniques are asymptotically equivalent [358]. In a thorough comparison and analysis of the two methods, he demonstrated that (1) the bootstrap procedure is "bad" for small sample sizes or whenever the alternative is close to the null hypothesis and (2) resampling tests should be used in order to take advantage of their flexibility in the choice of a distance criteria [358, p. 183].

In 1988 Tukey stated that the relationship between permutation procedures, on the one hand, and bootstrap and jackknife procedures, on the other hand, is "far from close" [1382]. Specifically, Tukey listed four major differences between bootstrap and jackknife procedures, which he called "resampling," and permutation methods, which he called "rerandomization" [1382].

1. Bootstrap and jackknife procedures need not begin until the data is collected. Rerandomization requires planning before the data collection is specified.
2. Bootstrap and jackknife procedures play games of omission of units with data already collected. Rerandomization plays games of exchange of treatments, while using all numerical results each time.
3. Bootstrap and jackknife procedures apply to experiences as well as experiments. Rerandomization only applies to randomized experiments.
4. Bootstrap and jackknife procedures give one only a better approximation to a desired confidence interval. Rerandomization gives one a "platinum standard" significance test, which can be extended in simple cases—by the usual devices—to a "platinum standard" confidence interval.

Thus, bootstrapping remains firmly in the conditional-on-assignment tradition, assuming that the true error distribution can be approximated by a discrete distribution with equal probability attached to each of the cases [850]. On the other hand, permutation tests view the errors as fixed in repeated samples [748]. Finally, some researchers have tacitly conceived of permutation methods in a Bayesian context. Specifically, this interpretation amounts to a primitive Bayesian analysis where the prior distribution is the assumption of equally-likely arrangements associated with the observed data, and the posterior distribution is the resulting data-dependent distribution of the test statistic induced by the prior distribution.

1.4 Student's *t* Test

Student's pooled t test [1331] for two independent samples is a convenient vehicle to illustrate permutation tests and to compare a permutation test with its parametric counterpart. As a historical note, Student's 1908 publication used z for the test statistic, and not t. The first mention of t appeared in a letter from William Sealy Gosset ("Student") to R.A. Fisher in November of 1922. It appears that the decision to change from z to t originated with Fisher, but the choice of the letter t was due

to Student. Eisenhart [408] and Box [196] provide historical commentaries on the transition from Student's z test to Student's t test.

Student's pooled t test for two independent samples is well-known, familiar to most researchers, widely used in quantitative analyses, and elegantly simple. The pooled t test evaluates the mean difference between two independent random samples. Under the null hypothesis, H_0: $\mu_1 = \mu_2$, Student's pooled t test statistic is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} ,$$

where the standard error of the sampling distribution of differences between two independent sample means is given by

$$s_{\bar{x}_1 - \bar{x}_2} = \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{n_1 + n_2}{n_1 n_2} \right) \right]^{1/2} ,$$

μ_1 and μ_2 denote the hypothesized population means, $\bar{x}_1$ and $\bar{x}_2$ denote the sample means, s_1^2 and s_2^2 denote the sample variances, and t follows Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, assuming the data samples are from independent normal distributions with equal variances.

1.4.1 An Exact Permutation *t* Test

Exact permutation tests are based on all possible arrangements of the observed data. For the two-sample t test, the number of permutations of the observed data is given by

$$M = \frac{N!}{n_1!\, n_2!} ,$$

where $N = n_1 + n_2$.

Let x_{ij} denote the ith observed score in the jth independent sample, $j = 1, 2$ and $i = 1, \ldots, n_j$, let t_o denote the Student t statistic computed on the observed data, and let t_k denote the Student t statistic computed on each permutation of the observed data for $k = 1, \ldots, M$. For the first permutation of the observed data set, interchange x_{13} and x_{12}, compute t_1, and compare t_1 with t_o. For the second permutation, interchange x_{12} and x_{22}, compute t_2, and compare t_2 with t_o. Continue the process for $k = 1, \ldots, M$.

To illustrate the exact permutation procedure, consider two independent samples of $n_1 = n_2 = 3$ observations and let $\{x_{11}, x_{21}, x_{31}\}$ denote the $n_1 = 3$ observations in Sample 1 and $\{x_{12}, x_{22}, x_{32}\}$ denote the $n_2 = 3$ observations in Sample 2. Table 1.1 depicts the

Table 1.1 Illustrative $M = 20$ permutations of $N = 6$ observations in two independent samples with $n_1 = n_2 = 3$

	Sample 1			Sample 2			
Permutation	1	2	3	1	2	3	t
1	x_{11}	x_{21}	x_{31}	x_{12}	x_{22}	x_{32}	t_1
2	x_{11}	x_{21}	x_{12}	x_{31}	x_{22}	x_{32}	t_2
3	x_{11}	x_{21}	x_{22}	x_{31}	x_{12}	x_{32}	t_3
4	x_{11}	x_{21}	x_{32}	x_{31}	x_{12}	x_{22}	t_4
5	x_{11}	x_{31}	x_{12}	x_{21}	x_{22}	x_{32}	t_5
6	x_{11}	x_{31}	x_{22}	x_{21}	x_{12}	x_{32}	t_6
7	x_{11}	x_{31}	x_{32}	x_{21}	x_{12}	x_{22}	t_7
8	x_{11}	x_{12}	x_{22}	x_{21}	x_{31}	x_{32}	t_8
9	x_{11}	x_{12}	x_{32}	x_{21}	x_{31}	x_{22}	t_9
10	x_{11}	x_{22}	x_{32}	x_{21}	x_{31}	x_{12}	t_{10}
11	x_{21}	x_{31}	x_{12}	x_{11}	x_{22}	x_{32}	t_{11}
12	x_{21}	x_{31}	x_{22}	x_{11}	x_{12}	x_{32}	t_{12}
13	x_{21}	x_{31}	x_{32}	x_{11}	x_{12}	x_{22}	t_{13}
14	x_{21}	x_{12}	x_{22}	x_{11}	x_{31}	x_{32}	t_{14}
15	x_{21}	x_{12}	x_{32}	x_{11}	x_{31}	x_{22}	t_{15}
16	x_{21}	x_{22}	x_{32}	x_{11}	x_{31}	x_{12}	t_{16}
17	x_{31}	x_{12}	x_{22}	x_{11}	x_{21}	x_{32}	t_{17}
18	x_{31}	x_{12}	x_{32}	x_{11}	x_{21}	x_{22}	t_{18}
19	x_{31}	x_{22}	x_{32}	x_{11}	x_{21}	x_{12}	t_{19}
20	x_{12}	x_{22}	x_{32}	x_{11}	x_{21}	x_{31}	t_{20}

$$M = \frac{6!}{3!\,3!} = 20$$

arrangements of $n_1 = n_2 = 3$ observations in each of the two independent samples where $t_o = t_1$, the subscripts denote the original position of each observation in either Sample 1 or Sample 2, and the position of the observation in Table 1.1 on either the left side of the table in Sample 1 or the right side of the table in Sample 2 indicates the placement of the observation after permutation. The exact two-sided probability (P) value is then given by

$$P = \frac{\text{number of } |t_k| \text{ values} \geq |t_o|}{M} \qquad \text{for } k = 1, \ldots, M \ .$$

1.4.2 A Moment-Approximation *t* Test

Moment-approximation permutation tests filled an important gap in the development of permutation statistical methods. Prior to the advent of modern computers, exact tests were impossible to compute except for extremely small samples, and even resampling-approximation permutation tests were limited in the number of

random permutations of the data possible, thus yielding too few places of accuracy for research purposes.

A moment-approximation permutation test is based, for example, on the first three exact moments of the underlying discrete permutation distribution, yielding the exact mean, variance, and skewness, i.e., μ_x, σ_x^2, and γ_x. Computational details for the exact moments are given in Sect. 4.15 of Chap. 4. An approximate probability value is obtained by fitting the exact moments to the associated Pearson type III distribution, which is completely characterized by the first three moments, and integrating the obtained Pearson type III distribution.

1.4.3 A Resampling-Approximation *t* Test

When M is very large, exact permutation tests are impractical, even with high-speed computers, and resampling-approximation permutation tests become an important alternative. Resampling-approximation tests provide more precise probability values than moment-approximation tests and are similar in structure to exact tests, except that only a random sample of size L selected from all possible permutations, M, is generated, where L is usually a large number to guarantee accuracy to a specified number of places. For instance, $L = 1{,}000{,}000$ will likely ensure three places of accuracy [696]. The resampling two-sided approximate probability value is then given by

$$\hat{P} = \frac{\text{number of } |t_k| \text{ values} \geq |t_o|}{L} \qquad \text{for } k = 1, \ldots, L \ .$$

1.5 An Example Data Analysis

The English poor laws, the relief expenditure act, and a comparison of two English counties provide vehicles to illustrate exact, moment-approximation, and resampling-approximation permutation tests.

The English Poor Laws

Up until the Reformation, it was considered a Christian duty in England to undertake the seven corporal works of mercy. In accordance with Matthew 25:32–46, Christians were to feed the hungry, give drink to the thirsty, welcome a stranger, clothe the naked, visit the sick, visit the prisoner, and bury the dead. After the Reformation and the establishment of the Church of England, many of these precepts were neglected, the poor were left without adequate assistance, and it became necessary to regulate relief of the poor

(continued)

by statute. The Poor Laws passed during the reign of Elizabeth I played a determining role in England's system of welfare, signaling a progression from private charity to a welfare state, where care of the poor was embodied in law. Boyer [198] provides an exhaustive description of the historical development of the English Poor Laws.

In 1552, Parish registers of the poor were introduced to ensure a well-documented official record, and in 1563, Justices of the Peace were empowered to raise funds to support the poor. In 1572, it was made compulsory that all people pay a poor tax, with those funds used to help the deserving poor. In 1597, Parliament passed a law that each parish appoint an Overseer of the Poor who calculated how much money was needed for the parish, set the poor tax accordingly, collected the poor rate from property owners, dispensed either food or money to the poor, and supervised the parish poor house. In 1601, the Poor Law Act was passed by Parliament, which brought together all prior measures into one legal document. The act of 1601 endured until the Poor Law Amendment Act was passed in 1834.

Consider an example data analysis utilizing Student's pooled two-sample t test based on historical parish-relief expenditure data from the 1800s [697]. To investigate factors that contributed to the level of relief expenditures, Boyer [198] assembled a data set comprised of a sample of 311 parishes in 20 counties in the south of England in 1831. The relief expenditure data were obtained from Blaug [172].[6] Table 1.2 contains the 1831 per capita relief expenditures, in shillings, for 36 parishes in two counties: Oxford and Hertford. For this example, the data were rounded to four places.

The relief expenditure data from Oxford and Hertford counties are listed in Table 1.2. Oxford County consisted of 24 parishes with a sample mean relief of $\bar{x}_1 = 20.28$ shillings and a sample variance of $s_1^2 = 58.37$ shillings. Hertford County consisted of 12 parishes with a sample mean relief of $\bar{x}_2 = 13.47$ shillings and a sample variance of $s_2^2 = 37.58$ shillings. A conventional two-sample t test yields $t_o = +2.68$ and, with $24 + 12 - 2 = 34$ degrees of freedom, a two-sided approximate probability value of $\hat{P} = .0113$. Although there are

$$M = \frac{36!}{24!\,12!} = 1{,}251{,}677{,}700$$

possible arrangements of the observed data and an exact permutation test is therefore not practical, it is not impossible. For the Oxford and Hertford relief expenditure

[6] The complete data set is available in several formats at the Cambridge University Press site: http://uk.cambridge.org/resources/0521806631.

Table 1.2 Average per capita relief expenditures for Oxford and Hertford counties in shillings: 1831

Oxford County				Hertford County	
Parish	Expenditure	Parish	Expenditure	Parish	Expenditure
1	20.3619	13	25.4683	1	27.9748
2	29.0861	14	12.5632	2	6.4173
3	14.9318	15	13.2780	3	10.4841
4	24.1232	16	27.3030	4	10.0057
5	18.2075	17	29.6055	5	9.7699
6	20.7287	18	13.6132	6	15.8665
7	8.1195	19	11.3714	7	19.3424
8	14.0201	20	21.5248	8	17.1452
9	18.4248	21	20.9408	9	13.1342
10	34.5466	22	11.5952	10	10.0420
11	16.0927	23	18.2355	11	15.0838
12	24.6166	24	37.8809	12	6.3985

data in Table 1.2, an exact permutation analysis yields a two-sided probability value of $P = 10{,}635{,}310/1{,}251{,}677{,}700 = 0.0085$.

A moment-approximation permutation analysis of the Oxford and Hertford relief expenditure data in Table 1.2 based on the Pearson type III distribution, yields a two-sided approximate probability value of $\hat{P} = 0.0100$.

Finally, a resampling analysis of the Oxford and Hertford relief expenditure data based on $L = 1{,}000{,}000$ random arrangements of the observed data in Table 1.2, yields 8,478 calculated t values equal to or more extreme than the observed value of $t_o = +2.68$, and a two-sided approximate probability value of $\hat{P} = 8{,}478/1{,}000{,}000 = 0.0085$.

1.6 Overviews of Chaps. 2–6

Chapters 2–6 describe the birth and development of statistical permutation methods. Chapter 2 covers the period from 1920 to 1939; Chap. 3, the period from 1940 to 1959; Chap. 4, the period from 1960 to 1979; and Chap. 5, the period from 1980 to 2000. Chapter 6 looks beyond the year 2000, summarizing the development of permutation methods from 2001 to 2010. Following Chap. 6 is a brief epilogue summarizing the attributes that distinguish permutation statistical methods from conventional statistical methods.

Chapter 2: 1920–1939

Chapter 2 chronicles the period from 1920 to 1939 when the earliest discussions of permutation methods appeared in the literature. In this period J. Spława-Neyman, R.A. Fisher, R.C. Geary, T. Eden, F. Yates, and E.J.G. Pitman laid the foundations of permutation methods as we know them today. As is evident in this period,

permutation methods had their roots in agriculture and, from the beginning, were widely recognized as the gold standard against which conventional methods could be verified and confirmed.

In 1923 Spława-Neyman introduced a permutation model for the analysis of field experiments [1312], and in 1925 Fisher calculated an exact probability using the binomial distribution [448]. Two years later in 1927, Geary used an exact analysis to support the use of asymptotic methods for correlation and regression [500], and in 1933 Eden and Yates used a resampling-approximation permutation approach to validate the assumption of normality in an agricultural experiment [379].

In 1935, Fisher's well-known hypothesized experiment involving "the lady tasting tea" was published in the first edition of *The Design of Experiments* [451]. In 1936, Fisher used a shuffling technique to demonstrate how a permutation test works [453], and in the same year Hotelling and Pabst utilized permutation methods to calculate exact probability values for the analysis of rank data [653].

In 1937 and 1938, Pitman published three seminal articles on permutation methods. The first article dealt with permutation methods in general, with an emphasis on the two-sample test; the second article with permutation methods as applied to bivariate correlation; and the third article with permutation methods as applied to a randomized blocks analysis of variance [1129–1131].

In addition to laying the foundations for permutation tests, the 1920s and 1930s were also periods in which tools to ease the computation of permutation tests were developed. Probability tables provided exact values for small samples, rank tests simplified the calculations, and desktop calculators became more available. Importantly, statistical laboratories began to appear in the United States in the 1920s and 1930s, notably at the University of Michigan and Iowa State College of Agriculture (now, Iowa State University). These statistical centers not only resulted in setting the foundations for the development of the computing power that would eventually make permutation tests feasible, they also initiated the formal study of statistics as a stand-alone discipline.

Chapter 3: 1940–1959

Chapter 3 explores the period between 1940 and 1959 with attention to the continuing development of permutation methods. This period may be considered as a bridge between the early years where permutation methods were first conceptualized and the next period, 1960–1979, in which gains in computer technology provided the necessary tools to successfully employ specific permutation tests.

Between 1940 and 1959, the work on establishing permutation statistical methods that began in the 1920s continued. In the 1940s, researchers applied known permutation techniques to create tables of exact probability values for small samples, among them tables for 2×2 contingency tables; the Spearman and Kendall rank-order correlation coefficients; the Wilcoxon, Mann–Whitney, and Festinger two-sample rank-sum tests; and the Mann test for trend.

Theoretical work, driven primarily by the computational challenges of calculating exact permutation probability values, was also completed during this period. Instead of the focus being on new permutation tests, however, attention turned to developing more simple alternatives to do calculations by converting data to rank-order statistics. Examples of rank tests that were developed between 1940 and 1959 include non-parametric randomization tests, exact tests for randomness based on serial correlation, and tests of significance when the underlying probability distribution is unknown.

While this theoretical undertaking continued, other researchers worked on developing practical non-parametric rank tests. Key among these tests were the Kendall rank-order correlation coefficient, the Kruskal–Wallis one-way analysis of variance rank test, the Wilcoxon and Mann–Whitney two-sample rank-sum tests, and the Mood median test.

Chapter 4: 1960–1979

Chapter 4 surveys the development of permutation methods in the period between 1960 and 1979 that was witness to dramatic improvements in computer technology, a process that was integral to the further development of permutation statistical methods. Prior to 1960, computers were based on vacuum tubes[7] and were large, slow, expensive, and availability was severely limited. Between 1960 and 1979 computers increasingly became based on transistors and were smaller, faster, more affordable, and more readily available to researchers. As computers became more accessible to researchers, work on permutation tests continued with much of the focus of that work driven by computer limitations in speed and storage.

During this period, work on permutation methods fell primarily into three categories: writing algorithms that efficiently generated permutation sequences; designing exact permutation analogs for existing parametric statistics; and, for the first time, developing statistics specifically designed for permutation methods. Numerous algorithms were published in the 1960s and 1970s with a focus on increasing the speed and efficiency of the routines for generating permutation sequences. Other researchers focused on existing statistics, creating permutation counterparts for well-known conventional statistics, notably the Fisher exact probability test for 2×2 contingency tables, the Pitman test for two independent samples, the F test for randomized block designs, and the chi-squared test for goodness of fit. The first procedures designed specifically for permutation methods, multi-response permutation procedures (MRPP), appeared during this period.

[7]The diode and triode vacuum tubes were invented in 1906 and 1908, respectively, by Lee de Forest.

Chapter 5: 1980–2000

Chapter 5 details the development of permutation methods during the period 1980 to 2000. It is in this period that permutation tests may be said to have arrived. One measure of this arrival was the expansion in the coverage of permutation tests, branching out from the traditional coverage areas in computer technology and statistical journals, and into such diverse subject areas as anthropology, atmospheric science, biomedical science, psychology, and environmental health. A second measure of the arrival of permutation statistical methods was the sheer number of algorithms that continued to be developed in this period, including the development of a pivotal network algorithm by Mehta and Patel in 1980 [919]. Finally, additional procedures designed specifically for permutation methods, multivariate randomized block permutation (MRBP) procedures, were published in 1982 by Mielke and Iyer [984].

This period was also home to the first books that dealt specifically with permutation tests, including volumes by Edgington in 1980, 1987 and 1995 [392–394], Hubert in 1987 [666], Noreen in 1989 [1041], Good in 1994 and 1999 [522–524], Manly in 1991 and 1997 [875, 876], and Simon in 1997 [1277], among others. Permutation versions of known statistics continued to be developed in the 1980s and 1990s, and work also continued on developing permutation statistical tests that did not possess existing parametric analogs.

Chapter 6: Beyond 2000

Chapter 6 describes permutation methods after the year 2000, an era in which permutation tests have become much more commonplace. Computer memory and speed issues that hampered early permutation tests are no longer factors and computers are readily available to virtually all researchers. Software packages for permutation tests now exist for well-known statistical programs such as StatXact, SPSS, Stata, and SAS. A number of books on permutation methods have been published in this period, including works by Chihara and Hesterberg in 2011, Edgington and Onghena in 2007 [396], Good in 2000 and 2001 [525–527], Lunneborg in 2000 [858], Manly in 2007 [877], Mielke and Berry in 2001 and 2007 [961, 965], and Pesarin and Salmaso in 2010 [1122].

Among the many permutation methods considered in this period are analysis of variance, linear regression and correlation, analysis of clinical trials, measures of agreement and concordance, rank tests, ridit analysis, power, and Bayesian hierarchical analysis. In addition, permutation methods expanded into new fields of inquiry, including animal research, bioinformatics, chemistry, clinical trials, operations research, and veterinary medicine.

The growth in the field of permutations is made palpable by a search of The Web of Science® using the key word "permutation." Between 1915 and 1959, the key word search reveals 43 journal articles. That number increases to 540 articles

for the period between 1960 and 1979 and jumps to 3,792 articles for the period between 1980 and 1999. From 2000 to 2010, the keyword search for permutation results in 9,259 journal articles.

Epilogue

A brief coda concludes the book. Chapter 2 contains a description of the celebrated "lady tasting tea" experiment introduced by Fisher in 1935 [451, pp. 11–29], which is the iconic permutation test. The Epilogue returns full circle to the lady tasting tea experiment, analyzing the original experiment to summarize the attributes that distinguish permutation tests from conventional tests in general.

Researchers early on understood the superiority of permutation tests for calculating exact probability values. These same researchers also well understood the limitations of trying to calculate exact probability values. While some researchers turned to developing asymptotic solutions for calculating probability values, other researchers remained focused on the continued development of permutation tests. This book chronicles the search for better methods for calculating permutation tests, the development of permutation counterparts for existing parametric statistical tests, and the development of separate, unique permutation tests.