Explorations in statistics: permutation methods

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 16 May 2012; accepted in final form 27 June 2012

Curran-Everett D. Explorations in statistics: permutation methods. Adv Physiol Educ 36: 181–187, 2012; doi:10.1152/advan.00072.2012.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This eighth installment of Explorations in Statistics explores permutation methods, empirical procedures we can use to test an experimental result—test a null hypothesis—when we are reluctant to trust statistical theory alone. Permutation methods operate on the observations—the data—we get from an experiment. A permutation procedure answers this question: out of all the possible ways we can rearrange the observations we got, in what proportion of those arrangements is the sample statistic we care about at least as extreme as the one we got? The answer to that question is the $P$ value.

Permutation methods give us an empiric approach to estimate the theoretic distribution of some statistic and to get the $P$ value associated with that statistic. When we explored correlation (8) we calculated a sample correlation coefficient of $r = 0.26$ for which the associated $P$ value was 0.05. Fisher calculated $t = 2.148$ (see Ref. 6). If the null hypothesis is true, we expect to observe a value of $t$ at least as big as 2.148 about 1 time in 20 ($P = 0.0497 \div 0.05$). Fisher concluded that cross-fertilized corn is barely taller than self-fertilized corn.

When we explored the bootstrap (7) we discovered that the result of a $t$ test is meaningful only if the theoretic distribution of the sample mean is roughly normal. One way this can happen is if our observations are drawn from a population that is distributed normally (4). In The Design of Experiments (20), Fisher mused about the extent to which his conclusion would be impacted if the observed differences came from something other than a normal distribution:

It has been mentioned that Student’s $t$ test is meaningful only if the theoretic distribution of the sample mean is roughly normal. One way this can happen is if our observations are drawn from a population that is distributed normally (4). In The Design of Experiments (20), Fisher mused about the extent to which his conclusion would be impacted if the observed differences came from something other than a normal distribution:

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application in order to show the possibility of an independent check on the more expeditious methods in common use.

Fisher then proceeded to detail the results of his tedious check: in 1,726 of the 32,768 possible ways of arranging the heights was the sum of the differences at least as extreme as the one Darwin actually observed. Therefore, $P = 1,726/32,768 = 0.05267 \div 0.05$,

which translates to, the observed differences are consistent with having come from a normal distribution that has a mean $\mu$ of 0.4. Or, from a practical standpoint, the heights of cross- and self-fertilized corn plants are the same. Using the 15 differences, Fisher calculated $t = 2.148$ (see Ref. 6). If the null hypothesis is true, we expect to observe a value of $t$ at least as big as 2.148 about 1 time in 20 ($P = 0.0497 \div 0.05$). Fisher concluded that cross-fertilized corn is barely taller than self-fertilized corn.

When we explored the bootstrap (7) we discovered that the result of a $t$ test is meaningful only if the theoretic distribution of the sample mean is roughly normal. One way this can happen is if our observations are drawn from a population that is distributed normally (4). In The Design of Experiments (20), Fisher mused about the extent to which his conclusion would be impacted if the observed differences came from something other than a normal distribution:

It has been mentioned that Student’s $t$ test is meaningful only if the theoretic distribution of the sample mean is roughly normal. One way this can happen is if our observations are drawn from a population that is distributed normally (4). In The Design of Experiments (20), Fisher mused about the extent to which his conclusion would be impacted if the observed differences came from something other than a normal distribution:

In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application in order to show the possibility of an independent check on the more expeditious methods in common use.

Fisher then proceeded to detail the results of his tedious check: in 1,726 of the 32,768 possible ways of arranging the heights was the sum of the differences at least as extreme as the one Darwin actually observed. Therefore, $P = 1,726/32,768 = 0.05267 \div 0.05$,
Permutation Methods: an Overview

When we explored hypothesis tests and \( P \) values (6), we developed the notion of a theoretic distribution of the test statistic \( t \) by drawing 1000 random samples from a standard normal distribution with mean \( \mu = 0 \). To assess whether the observations in our first sample were consistent with having come from this population, we calculated \( t = 3.407 \). We based our interpretation of this value of 3.407 on a theoretic distribution of \( t \) (see Ref. 6, Fig. 3). What did we conclude? Because the test statistic \( t = 3.407 \) (\( P = 0.005 \)) was more unusual than our critical significance level of \( \alpha = 0.10 \), we concluded that the sample observations were consistent with having come from a population that had a mean other than 0.

This simulation illustrates the population model of statistical inference, a model advanced by Neyman and Pearson (41) and labeled by Lehmann (31). The academic difficulty with this model is that it is almost never realized in practice.

Instead, we have at hand subjects or experimental units—people, sheep, rabbits, cell cultures—whence we assign at random to different experimental groups.\(^9\) This randomization model of statistical inference was defined explicitly by Kempthorne (28). In the strict application of this model, the theoretic distribution of some test statistic is not relevant.

So the question becomes, in this situation, how do we obtain that ubiquitous \( P \) value?

Suppose we have six rabbits in which we want to examine the impact of some intervention on changes in brain blood flow. We randomly assign three rabbits to group 1, the control group, and the three remaining rabbits to group 2, the experimental group. We define our null hypothesis to be that changes in brain blood flow are similar in the two groups. As it turns out, the rabbits in group 2 had the biggest changes in brain blood flow: on average, brain blood flow increased \( 3 \) ml/min more than in group 1 (Table 1, arrangement 1).

How do we use the observations from the six rabbits to assess our null hypothesis? If the null hypothesis is true—if changes in brain blood flow are similar—then the observations are exchangeable (24). This means that out of all the ways we can rearrange our observations, the arrangement we got is a typical one (24, 36). How do we implement this abstract idea to get a \( P \) value? For each of the 19 other ways we can rearrange our 6 observations (see Table 1), we calculate the difference between sample means and then ask this question: out of all the possible ways we can rearrange the observations we got, in what proportion of those arrangements is the difference between sample means at least as extreme as the one we got? The answer to that question is our \( P \) value.

In 2 of the 20 possible arrangements of these observations was the magnitude of the difference between sample means at least as big as 3, the one we got (Table 1, arrangements 1 and 20). Therefore, \( P = 2/20 = 0.10 \). The Appendix extends this example.

---

\(^5\) Although Fisher was known for fierce exchanges with other statisticians (see Ref. 6), it is unlikely he felt affronted by Eden and Yates: they worked with him at Rothamsted Experimental Station.

\(^6\) This file is available through the Supplemental Material link for this article at the Advances in Physiology Education website.

\(^7\) Functions of the coin package perform conditional inference, inference based on actual observations (26).

\(^8\) The notation \( A \mid B \) means click A, then click B.
Table 1. Concept of a permutation method for a two-sample problem

<table>
<thead>
<tr>
<th>Arrangement</th>
<th>Group 1</th>
<th>Group 2</th>
<th>( \Delta \bar{y}_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 2 3</td>
<td>4 5 6</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1 2 4</td>
<td>3 5 6</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>1 2 5</td>
<td>4 3 6</td>
<td>1.7</td>
</tr>
<tr>
<td>4</td>
<td>1 2 6</td>
<td>4 5 3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1 4 3</td>
<td>2 5 6</td>
<td>1.7</td>
</tr>
<tr>
<td>6</td>
<td>1 5 3</td>
<td>4 2 6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1 6 3</td>
<td>4 5 2</td>
<td>0.3</td>
</tr>
<tr>
<td>8</td>
<td>4 2 3</td>
<td>1 5 6</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>5 2 3</td>
<td>4 1 6</td>
<td>0.3</td>
</tr>
<tr>
<td>10</td>
<td>6 2 3</td>
<td>4 5 1</td>
<td>-0.3</td>
</tr>
<tr>
<td>11</td>
<td>1 4 5</td>
<td>2 3 6</td>
<td>0.3</td>
</tr>
<tr>
<td>12</td>
<td>1 4 6</td>
<td>2 5 3</td>
<td>-0.3</td>
</tr>
<tr>
<td>13</td>
<td>1 5 6</td>
<td>4 2 3</td>
<td>-1</td>
</tr>
<tr>
<td>14</td>
<td>4 2 5</td>
<td>1 3 6</td>
<td>-0.3</td>
</tr>
<tr>
<td>15</td>
<td>4 2 6</td>
<td>1 5 3</td>
<td>-1</td>
</tr>
<tr>
<td>16</td>
<td>5 2 6</td>
<td>4 1 3</td>
<td>-1.7</td>
</tr>
<tr>
<td>17</td>
<td>4 5 3</td>
<td>1 2 6</td>
<td>-1</td>
</tr>
<tr>
<td>18</td>
<td>4 6 3</td>
<td>1 5 2</td>
<td>-1.7</td>
</tr>
<tr>
<td>19</td>
<td>5 6 3</td>
<td>4 2 2</td>
<td>-2.3</td>
</tr>
<tr>
<td>20</td>
<td>4 5 6</td>
<td>1 2 3</td>
<td>-3</td>
</tr>
</tbody>
</table>

For arrangement \( j \), the difference \( \Delta \bar{y}_j \) between sample means is \( \Delta \bar{y}_j = \bar{y}_2 - \bar{y}_1 \). Arrangement 1 holds the observed sample data.

A More Realistic Two-Sample Example

Imagine that we want to study the impact of diet on plasma cholesterol in women.\(^{10}\) We randomly assign seven women to group 1, a diet of fish, and five women to group 2, a diet of meat. We define our null hypothesis to be that, after 1 year on a diet of fish or meat, plasma cholesterol levels will be similar in the two groups, and we establish a critical significance level—a benchmark for uncommonness—of \( \alpha = 0.05 \). Table 2 lists the observations from this fictitious experiment (see also Fig. 1).

In our second exploration we mentioned that a one-sample \( t \) statistic

\[
t = \frac{\bar{y} - \bar{y}_1}{SE(\bar{y} - \bar{y}_1)} ,
\]

(See Ref. 12, APPENDIX and Ref. 51, p. 100.) In this two-sample expression for \( t \), we opt for the more versatile form of \( SE(\bar{y} - \bar{y}_1) \),

\[
SE(\bar{y} - \bar{y}_1) = \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}} ,
\]

in which the subscripts 1 and 2 identify group number.

For the moment, we will set aside concerns that the theoretic distribution of \( t \) as defined in Eq. 1 may not apply under this randomized model.\(^{11}\) In this case, \( s_2^2 = 3.658, s_1^2 = 0.392, \) and \( SE(\bar{y} - \bar{y}_1) = 0.8875. \) Therefore,

\[
t = \frac{\bar{y} - \bar{y}_1}{SE(\bar{y} - \bar{y}_1)} = \frac{8.20 - 6.41}{0.8875} = 2.017 .
\]

The commands in lines 61–79 of Advances_Statistics_Code_Perm.R return these values. If the null hypothesis is true—if plasma cholesterol levels are similar—how usual is this value of \( t \)? If the null hypothesis is true, we expect to observe a value of \( |t| \) at least as big as 2.017 about 1 time in 10 \( (P = 0.104 = 0.10) \). This is more usual than our benchmark of 0.05. As a result, we fail to reject the null hypothesis and conclude that—although there is a suggestion that diet affects plasma cholesterol levels (see Ref. 11, Table 1)—the sample observations are consistent with having come from the same population: that is, plasma cholesterol levels are similar in the two groups.

<table>
<thead>
<tr>
<th>Fish 1</th>
<th>Meats 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.42</td>
<td>6.51</td>
</tr>
<tr>
<td>5.86</td>
<td>7.56</td>
</tr>
<tr>
<td>6.16</td>
<td>7.61</td>
</tr>
<tr>
<td>6.55</td>
<td>7.84</td>
</tr>
<tr>
<td>6.80</td>
<td>11.50</td>
</tr>
<tr>
<td>7.00</td>
<td></td>
</tr>
<tr>
<td>7.11</td>
<td></td>
</tr>
</tbody>
</table>

Values represent plasma cholesterol (in mmol/l) after 1 year on a diet of fish or meat. Simulated data are from Ludbrook and Dudley (35).

\(^{10}\) Ludbrook and Dudley (35) used this example. Assume that the women have the same plasma cholesterol value beforehand.

\(^{11}\) Some statisticians say that \( t \) tests apply meaningfully to the randomization model of inference (see Ref. 39).
This is more unusual—less likely to occur—than our problem.

**Table 3.** Exact permutation method for a two-sample problem

<table>
<thead>
<tr>
<th>Arrangement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>7</th>
<th>22</th>
<th>57</th>
<th>792</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Fish</td>
<td>5.42</td>
<td>5.42</td>
<td>5.42</td>
<td>5.42</td>
<td>5.42</td>
<td>5.42</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>5.86</td>
<td>5.86</td>
<td>5.86</td>
<td>5.86</td>
<td>5.86</td>
<td>5.86</td>
<td>7.61</td>
</tr>
<tr>
<td></td>
<td>6.16</td>
<td>6.16</td>
<td>6.16</td>
<td>6.16</td>
<td>6.16</td>
<td>6.16</td>
<td>7.84</td>
</tr>
<tr>
<td></td>
<td>6.80</td>
<td>6.80</td>
<td>6.80</td>
<td>6.51</td>
<td>6.80</td>
<td>6.80</td>
<td>6.80</td>
</tr>
<tr>
<td></td>
<td>7.00</td>
<td>6.51</td>
<td>6.51</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
<td>7.00</td>
</tr>
<tr>
<td></td>
<td>6.51</td>
<td>7.11</td>
<td>7.56</td>
<td>7.11</td>
<td>7.11</td>
<td>7.11</td>
<td>7.11</td>
</tr>
<tr>
<td>2 Meat</td>
<td>7.11</td>
<td>7.00</td>
<td>7.00</td>
<td>6.80</td>
<td>6.55</td>
<td>6.51</td>
<td>6.51</td>
</tr>
<tr>
<td></td>
<td>7.56</td>
<td>7.56</td>
<td>7.11</td>
<td>7.56</td>
<td>7.56</td>
<td>7.56</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>7.61</td>
<td>7.61</td>
<td>7.61</td>
<td>7.61</td>
<td>7.61</td>
<td>7.61</td>
<td>5.86</td>
</tr>
<tr>
<td></td>
<td>7.84</td>
<td>7.84</td>
<td>7.84</td>
<td>7.84</td>
<td>7.84</td>
<td>7.84</td>
<td>6.16</td>
</tr>
<tr>
<td></td>
<td>11.50</td>
<td>11.50</td>
<td>11.50</td>
<td>11.50</td>
<td>11.50</td>
<td>11.50</td>
<td>6.55</td>
</tr>
<tr>
<td>Δy&lt;sup&gt;∗&lt;/sup&gt;</td>
<td>2.00</td>
<td>1.96</td>
<td>1.80</td>
<td>1.89</td>
<td>1.80</td>
<td>1.79</td>
<td>-1.82</td>
</tr>
</tbody>
</table>

Arrangements of the data in Table 2 for which the magnitude of the difference Δy<sup>∗</sup> between sample means is at least as extreme as the observed difference of 1.79. *Arrangement 57* holds the observed sample data.

But we know from our exploration of the bootstrap (7) that this conclusion is meaningful only if the theoretic distribution of the difference between sample means is roughly normal. The bootstrap distribution provides clear evidence it is not (Fig. 2). Great. Now what? You guessed it: a randomization test.

The first question is, in how many ways can we rearrange our 12 observations? The answer is, in

\[
\binom{12}{5} = \frac{12!}{5!7!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7!}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 7!} = 792 \text{ ways}.^{12}
\]

The next question is, in how many of these possible arrangements is the difference between sample means at least as extreme as 1.79, the actual difference we observed? The answer is, in seven of them (Table 3 and Fig. 3). Therefore, if the null hypothesis is true, the probability that we would have observed a difference at least as big as 11.79 is 7/792 = 0.009. This is more unusual—less likely to occur—than our benchmark of 0.05. In contrast to our conclusion above, we reject the null hypothesis and conclude that plasma cholesterol levels differ between these two groups of women. We revisit this discrepancy in Practical Considerations.

In our overview of permutation methods, we said that if the null hypothesis is true, then the observations are exchangeable. This means we assume the observations came from populations that share a common standard deviation. In some situations, a randomization test can detect a difference in standard deviation rather than a difference in mean, but this is not inevitable (Fig. 4), and methods exist to deal with this phenomena (40).

**An Example in Correlation**

When we explored correlation, we used data from Snedecor and Cochran (51) to show that one observation can distort the magnitude of a correlation coefficient (see Ref. 8, Fig. 8). We can use the same data (Fig. 5) to demonstrate the application of permutation methods to correlation.

In the setting of correlation, if there is no relationship between two variables, then every possible arrangement of those variables is equally likely (16, 24). Now the question is, how do we rearrange our observations? We fix the x values and permute the y values. For these data, there are 11! = 39,916,800 possible arrangements of the y values (Table 4 and APPENDIX). Rather than calculate a correlation coefficient for each of those 11! arrangements, we calculate a correlation coefficient for 10,000 of those arrangements chosen at random.^{14}

In how many of the 10,000 arrangements is the correlation coefficient r at least as extreme as 0.56l, the actual magnitude we observed? The answer is, in 770 of them (Fig. 6). Therefore, if the null hypothesis is true, the probability that we would have observed a correlation at least as big as 0.56l is 770 times

\[
\binom{12}{5} = \frac{12!}{5!7!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7!}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 7!} = 792 \text{ ways}.^{12}
\]
in 10,000 ($P = \frac{770}{10,000} \approx 0.08$). The commands in lines 186–202 of Advances_Statistics_Code_Perm.R return these values. Your values will differ slightly.

**Practical Considerations**

In our examples, we used the same strategy (38):

1. Define the problem—the null hypothesis—we care about.
2. Identify and then calculate a sample statistic that is relevant to the null hypothesis.
3. Rearrange the observations in a way that is consistent with the null hypothesis. For each arrangement, calculate the sample statistic.
4. Calculate the proportion of sample statistics in the permutation distribution that are more extreme than the value we observed in the actual arrangement of the observations.

It is the final step that produces the $P$ value. This strategy applies to a broad spectrum of experimental designs (16, 24, 36).

In the 1930s, when Eden, Yates, Fisher, Pitman, and Welch pioneered permutation methods (15, 20, 46–48, 55), the computational effort was prohibitive. Today, that effort is trivial.

Table 4. Approximate permutation method for a problem in correlation

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>Arrangement $j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>69</td>
<td>66 64 .... 64</td>
</tr>
<tr>
<td>68</td>
<td>64</td>
<td>62 62 69</td>
</tr>
<tr>
<td>66</td>
<td>65</td>
<td>65 62 65</td>
</tr>
<tr>
<td>67</td>
<td>63</td>
<td>62 65 65</td>
</tr>
<tr>
<td>70</td>
<td>65</td>
<td>63 59 64</td>
</tr>
<tr>
<td>71</td>
<td>61</td>
<td>59 63 63</td>
</tr>
<tr>
<td>70</td>
<td>65</td>
<td>64 59 65</td>
</tr>
<tr>
<td>73</td>
<td>64</td>
<td>65 65 66</td>
</tr>
<tr>
<td>72</td>
<td>66</td>
<td>64 65 62</td>
</tr>
<tr>
<td>65</td>
<td>59</td>
<td>69 66 59</td>
</tr>
<tr>
<td>66</td>
<td>62</td>
<td>65 64 62</td>
</tr>
</tbody>
</table>

In columns $x$ and $y$ are the actual observations (see Fig. 5). To effect a permutation procedure, we fix the $x$ values and permute the $y$ values for which there are $11! = 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 39,916,800$ possible arrangements. The correlation coefficient between the actual $x$ values and arrangement $j$ of the $y$ values is $r_j$. The commands in lines 173–181 of Advances_Statistics_Code_Perm.R print three arrangements of the actual $y$ observations.

To generate three arrangements as in this table, highlight and submit the lines of code from Table 4: first line to Table 4: last line.

Commercial software packages such as StatXact (13) simplify the process.\(^{15}\)

But still, are permutation methods really worth the effort? When John Tukey writes

No other class of approach provides significance information of comparative quality.

(Ref. 53, p. 18)

\(^{15}\) See Refs. 16, 24, and 36 for other software packages.

Fig. 4. The impact of variability on a two-sample randomization test using the difference between sample means. In each of 3 situations, we draw at random 30 observations from each of 2 populations, calculate the difference between sample means for 10,000 arrangements chosen at random, and obtain a $P$ value. Then, we repeat the simulation 999 times. Each time, we draw the observations for sample 1 from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$; that is, from a $N(0,1)$. In situation A, we draw the observations for sample 2 from a $N(1,1)$; 95% of the $P$ values from the randomization test are less than 0.03 (left). In situation B, we draw the observations for sample 2 from a $N(1.5); 50\%$ of the $P$ values are less than 0.25 (middle); the fivefold difference in variability blunts out our ability to detect that the observations were drawn from populations of different means. In situation C, we draw the observations for sample 2 from a $N(0.5)$ and note that the fivefold difference in variability does not masquerade as a difference in mean: 50% of the $P$ values are less than 0.53 (right), a greater value than in situation B.

Fig. 5. Data used by Snedecor and Cochran (51) to illustrate correlation. See also Ref. 8. For these 11 observations, the sample correlation coefficient $r = 0.56$ ($P = 0.07$).

Fig. 6. Permutation distribution of 10,000 correlation coefficients for the data in Fig. 5. (See also Table 4.) This empiric distribution estimates the theoretic distribution of the correlation coefficient when the null hypothesis is true. The magnitude of 770 correlation coefficients (gray) is at least extreme as the observed value of 0.56. If the null hypothesis is true—if the true correlation coefficient is 0—then we expect to observe a correlation as big as 0.56 about 770 times in 10,000 ($P = \frac{770}{10,000} \approx 0.08$). The commands in lines 159–219 of Advances_Statistics_Code_Perm.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 6: first line to Figure 6: last line.
and when Brad Efron writes

When there is something to permute . . . it is a good idea to do so, even if other methods like the bootstrap are also brought to bear.

(Ref. 17, p. 218)

and when other statisticians (18, 24, 28, 35) also endorse the procedures, it is worth our while to pay attention.

You are starting to warm to permutation methods. But when do you use them? Although some statisticians advocate that we use a permutation method for the actual statistical analysis whenever we can (17, 18, 24, 28, 35, 53), we can also use a permutation method as we did the bootstrap (7): to assess whether an inference we make from a more traditional hypothesis test is justified. How do we do this? If our conclusion from permutation matches our conclusion from the traditional test, then the assumptions for the traditional procedure are reasonably well met (16, 20, 24, 36).

We have one last question to address: what happens if our conclusion from permutation conflicts with our conclusion from the traditional test? We should suspect that the assumptions for the traditional procedure have not been met, as in our two-sample example. It almost goes without saying that we want to opt not for the statistical procedure that produces the result of our dreams but for the statistical procedure that has its assumptions best satisfied.

Summary

As this exploration has demonstrated, permutation methods give us an approach we can use to assess an experimental result when we are reluctant to trust statistical theory alone (16, 24, 36). Like the bootstrap, permutation methods are empiric: they operate on the observations—the data—we get from an experiment. Unlike the bootstrap, the primary utility of a permutation procedure is in the test of a scientific null hypothesis. Although we have explored permutation methods using a basic two-sample problem and a problem in correlation, we can use them to reach conclusions about scientific results in other experimental designs (16, 24, 36).

In the next installment of this series, we will explore the analysis of ratios. As researchers, we use ratios to normalize a numerator to some denominator: to control for differences in the denominator when the thing we really care about is the numerator. In physiology, ratios are quite common. The prob-

lem is that the analysis of ratios is quite complex. In the next installment, we will see why.

APPENDIX

This appendix reviews nomenclature related to permutation methods and illustrates a practical extension of our overview of permutation methods.

Nomenclature

To be honest, I struggled with the nomenclature associated with permutation methods. In part I struggled because the nomenclature is divorced from the mechanics of the procedures. A permutation test refers to the possible rearrangements of observations among groups of subjects when those subjects have been drawn at random from their populations (16, 29). A randomization test refers to the possible rearrangements of observations among groups of subjects when those subjects have been drawn not at random but simply assigned at random to different experimental groups (16, 29).

To make matters worse, I had vague memories from high school of the definitions (54) of permutations and combinations. In permutations, order matters. Suppose we have n things and we want to ask, in how many ways can we arrange these n things? The first time we choose, we have n things from which to choose. After we have chosen one of those n things, we have n − 1 things from which to choose. This continues until we have chosen n − 1 things, at which point we have just one thing from which to choose:

\[
\frac{n}{\left(\frac{n-1}{r} \cdot \frac{n-2}{r-1} \ldots \frac{1}{r!(n-r)!}\right)}
\]

In how many ways can we arrange these n things? In \(n(n-1)(n-2)\ldots1 = n!\) ways, or in \(n!\) permutations. These are the mechanics we used with correlation (see Table 4).

In combinations, order does not matter. Suppose we have n things and we want to ask, in how many ways can we choose r of these n things? This means we will have two groups: one with r things and one with n − r things. We can calculate the number of ways—the number of combinations—as n choose r :

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!}
\]

These are the mechanics we used with our two-sample example (see Table 3).

Regardless of whether we use permutations or combinations to do a permutation test or a randomization test, the concept is the same: out of all the possible ways we can rearrange the observations we got, in what proportion of those arrangements is our sample statistic at least as extreme as the one we got?

A Practical Extension

In our overview, we used two groups of three observations to illustrate the framework of permutation (see Table 1). We discovered that in 2 of the 20 possible arrangements of those nonoverlapping observations, the magnitude of our sample statistic was at least as extreme as the actual difference we got (\(P = 2/20 = 0.10\)).

I have reviewed papers in which the authors analyzed two groups of three observations using a t test. This is reckless. The only way a t test can be applied meaningfully to two groups with so few observations is if the physiological thing we care about is distributed normally. With three observations in each group, there is simply no way to know. Using a permutation procedure, the smallest \(P\) value we can achieve with two groups of three observations is \(P = 0.10\).

Suppose we have two groups of four observations. We still have no good way of knowing if the physiological thing we care about is distributed normally, and so we still dismiss the t test. If we use a permutation approach, there are

\[
\binom{8}{4} = \frac{8!}{4!4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4!}{4 \cdot 3 \cdot 2 \cdot 4!} = 70 \text{ ways}
\]

we can rearrange the observations. In only 2 of those 70 arrangements,

<table>
<thead>
<tr>
<th>Control</th>
<th>1, 2, 3, 4</th>
<th>5, 6, 7, 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>5, 6, 7, 8</td>
<td>1, 2, 3, 4</td>
</tr>
</tbody>
</table>

can the sample statistic be at least as extreme as the one we got. This means that two groups of four nonoverlapping observations will result in \(P = 2/70 = 0.03\).

ACKNOWLEDGMENTS

I thank John Ludbrook (Department of Surgery, The University of Melbourne, Melbourne, Victoria, Australia), Gerald DiBona (University of Iowa College of Medicine, Iowa City, Iowa), and Bryan Manly (Western Ecosystems Technology, Cheyenne, Wyoming) for helpful comments and suggestions, and I thank Cyrus Mehta and Cytel Software for graciously granting me the opportunity to explore StatXact.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).
REFERENCES


47. Pitman EJ. Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. J Roy Stat Soc Suppl 4: 225–232, 1937.

48. Pitman EJ. Significance tests which may be applied to samples from any populations. III. The analysis of variance. Biometrika 29: 322–335, 1938.


