18. Distribution-Free Tests

by David M. Lane

Benefits

by David M. Lane

Prerequisites

- Chapter 7: Normal Distributions
- Chapter 3: Shapes of Distributions
- Chapter 13: Introduction to Power
- Chapter 16: Transformations

Learning Objectives

1. State how distribution-free tests can avoid an inflated Type I error rate
2. State how how distribution-free tests can affect power

Most tests based on the normal distribution are said to be robust when the assumption of normality is violated. To the extent to which actual *probability values* differ from nominal probability values, the actual probability values tend to be higher than the nominal p values. For example, if the probability of a difference as extreme or more extreme were 0.04, the test might report that the probability value is 0.06. Although this sounds like a good thing because the *Type I error* rate is lower than the nominal rate, it has a serious downside: reduced *power*. When the *null hypothesis* is false, the probability of rejecting the null hypothesis can be substantially lower than it would have been if the distributions were distributed normally.

Tests assuming normality can have particularly low power when there are extreme values or outliers. A contributing factor is the sensitivity of the mean to extreme values. Although transformations can ameliorate this problem in some situations, they are not a universal solution.

Tests assuming normality often have low power for leptokurtic distributions. Transformations are generally less effective for reducing kurtosis than for reducing.

Because distribution-free tests do not assume normality, they can be less susceptible to non-normality and extreme values. Therefore, they can be more powerful than the standard tests of means that assume normality.

Randomization Tests: Two Conditions

by David M. Lane

Prerequisites

- Chapter 18: Permutations and Combinations
- Chapter 11: One- and Two-Tailed Tests

Learning Objectives

1. Explain the logic of randomization tests
2. Compute a randomization test of the difference between independent groups

The data in Table 1 are from a fictitious experiment comparing an experimental group with a control group. The scores in the Experimental Group are generally higher than those in the Control Group with the Experimental Group mean of 14 being considerably higher than the Control Group mean of 4. Would a difference this large or larger be likely if the two treatments had identical effects? The approach taken by randomization tests is to consider all possible ways the values obtained in the experiment could be assigned to the two groups. Then, the location of the actual data within the list is used to assess how likely a difference that large or larger would occur by chance.

Table 1. Fictitious data.

Experimental	Control
7	0
8	2
11	5
30	9

First, consider all possible ways the 8 values could be divided into two sets of 4. We can apply the formula from the section on *Permutations and Combinations* for the number of combinations of n items taken r at a time and find that there are 70 ways.

$$C_r^n = \frac{n!}{(n-r)!\,r!} = \frac{8!}{(8-4)!\,4!} = 70$$

Of these 70 ways of dividing the data, how many result in a difference between means of 10 or larger? From Table 1 you can see that there are two rearrangements that would lead to a bigger difference than 10: (a) the score of 7 could have been in the Control Group with the score of 9 in the Experimental Group and (b) the score of 8 could have been in the Control Group with the score of 9 in the Experimental Group. Therefore, including the actual data, there are 3 ways to produce a difference as large or larger than the one obtained. This means that if assignments to groups were made randomly, the probability of this large or a larger advantage of the Experimental Group is 3/70 = 0.0429. Since only one direction of difference is considered (Experimental larger than Control), this is a one-tailed probability. The *two-tailed* probability is 0.0857 since there are 6/70 ways to arrange the data so that the absolute value of the difference between groups is as large or larger than the one obtained.

Clearly, this type of analysis would be very time consuming for even moderate sample sizes. Therefore, it is most useful for very small sample sizes.

An alternate approach made practical by computer software is to randomly divide the data into groups thousands of times and count the proportion of times the difference is as big or bigger than that found with the actual data. If the number of times the data are divided randomly is very large, then this proportion will be very close to the proportion you would get if you listed all possible ways the data could be divided.

Randomization Tests: Two or More Conditions

by David M. Lane

Prerequisites
• Chapter 18: Randomization Tests (two conditions)

Learning Objectives
1. Compute a randomization test for differences among more than two conditions.

The method of randomization for testing differences among more than two means is essentially very similar to the method when there are exactly two means. Table 1 shows the data from a fictitious experiment with three groups.

Table 1. Fictitious data.

T1	T2	Control
7	14	0
8	19	2
11	21	5
12	122	9

The first step in a randomization test is to decide on a test statistic. Then we compute the proportion of the possible arrangements of the data for which that test statistic is as large as or larger than the arrangement of the actual data. When comparing several means, it is convenient to use the F ratio. The F ratio is computed not to test for significance directly, but as a measure of how different the groups are. For these data, the F ratio for a one-way ANOVA is 2.06.

The next step is to determine how many arrangements of the data result in as large or larger F ratios. There are 6 arrangements that lead to the same F of 2.06: the six arrangements of the three columns. One such arrangement is shown in Table 2. The six are:

```
(1) T1, T2, Control
(2) T1, Control, T2
(3) T2, T1, Control
(4) T2, Control, T1
(5) Control, T1, T2
(6) Control, T2, T1
```

For each of the 6 arrangements there are two changes that lead to a higher F ratio: swapping the 7 for the 9 (which gives an F of 2.08) and swapping the 8 for the 9 (which gives an F of 2.07). The former of these two is shown in Table 3.

Table 2. Fictitious data with data for T1 and T2 swapped

T1	Control	T2
7	14	0
8	19	2
11	21	5
12	122	9

Table 3. Data from Table 1 with the 7 and the 9 swapped.

T1	T2	Control
9	14	0
8	19	2
11	21	5
12	122	7

Thus, there are six arrangements, each with two swaps that lead to a larger F ratio. Therefore, the number of arrangements with an F as large or larger than the actual arrangement is 6 (for the arrangements with the same F) + 12 (for the arrangements with a larger F), which makes 18 in all.

The next step is to determine the total number of possible arrangements. This can be computed from the following formula:

$$Arrangements = (n!)^k = (4!)^3 = 13{,}824$$

where n is the number of observations in each group (assumed to be the same for all groups), and k is the number of groups. Therefore, the proportion of arrangements with an F as large or larger than the F of 2.06 obtained with the data is

```
18/13,824 = 0.0013.
```

Thus, if there were no treatment effect, it is very unlikely that an F as large or larger than the one obtained would be found.

Randomization Tests: Association (Pearson's r)

by David M. Lane

Prerequisites

- Chapter 14: Inferential Statistics for b and r

Learning Objectives

1. Compute a randomization test for Pearson's r.

A significance test for Pearson's r is described in the section *inferential statistics for b and r*. The significance test described in that section assumes normality. This section describes a method for testing the significance of r that makes no distributional assumptions.

Table 1. Example data.

X	Y
1.0	1.0
2.4	2.0
3.8	2.3
4.0	3.7
11.0	2.5

The approach is to consider the X variable fixed and compare the correlation obtained in the actual data to the correlations that could be obtained by rearranging the Y variable. For the data shown in Table 1, the correlation between X and Y is 0.385. There is only one arrangement of Y that would produce a higher correlation. This arrangement is shown in Table 2 and the r is 0.945. Therefore, there are two arrangements of Y that lead to correlations as high or higher than the actual data.

Table 2. The example data arranged to give the highest r.

X	Y
1.0	1.0
2.4	2.0
3.8	2.3

4.0	2.5
11.0	3.7

The next step is to calculate the number of possible arrangements of Y. The number is simply N! where N is the number of pairs of scores. Here, the number of arrangements is 5! = 120. Therefore, the probability value is 2/120 = 0.017. Note that this is a one-tailed probability since it is the proportion of arrangements that give an r as large or larger. For the two-tailed probability, you would also count arrangements for which the value of r were less than or equal to -0.385. In randomization tests, the two-tailed probability is not necessarily double the one-tailed probability.

Randomization Tests: Contingency Tables: (Fisher's Exact Test)

by David M. Lane

Prerequisites

- Chapter 17: Contingency Tables

Learning Objectives

1. State the situation when Fisher's exact test can be used
2. Calculate Fisher's exact test
3. Describe how conservative the Fisher exact test is relative to a Chi Square test

The chapter on Chi Square showed one way to test the relationship between two nominal variables. A special case of this kind of relationship is the difference between proportions. This section shows how to compute a significance test for a difference in proportions using a randomization test. Suppose, in a fictitious experiment, 4 subjects in an Experimental Group and 4 subjects in a Control Group are asked to solve an anagram problem. Three of the 4 subjects in the Experimental Group and none of the subjects in the Control Group solved the problem. Table 1 shows the results in a contingency table.

Table 1. Anagram Problem Data.

	Experimental	Control	Total
Solved	3	0	3
Did not Solve	1	4	5
Total	4	4	8

The significance test we are going to perform is called the Fisher Exact Test. The basic idea is to take the row totals and column totals as "given" and add the probability of obtaining the pattern of frequencies obtained in the experiment and the probabilities of all other patterns that reflect a greater difference between conditions. The formula for obtaining any given pattern of frequencies is:

$$\frac{n!\,(N-n)!\,R!\,(N-R)!}{r!\,(n-r)!\,(R-r)!\,(N-n-R+r)!\,N!}$$

where N is the total sample size (8), n is the sample size for the first group (4), r is the number of successes in the first group (3), and R is the total number of successes (3). For this example, the probability is

$$\frac{4!\,(8-4)!\,3!\,(8-3)!}{3!\,(4-3)!\,(3-3)!\,(8-4-3+3)!\,8!} = 0.0714$$

Since more extreme outcomes do not exist given the row and column totals, the p value is 0.0714. This is a one-tailed probability since it only considers outcomes as extreme or more extreme favoring the Experimental Group. An equally extreme outcome favoring the Control Group is shown in Table 2, which also has a probability of 0.0714. Therefore, the two-tailed probability is 0.1428. Note that in the Fisher Exact Test, the two-tailed probability is not necessarily double the one-tailed probability.

Table 2. Anagram Problem Favoring Control Group.

	Experimental	Control	Total
Solved	0	3	3
Did not Solve	4	1	5
Total	4	4	8

The Fisher Exact Test is “exact” in the sense that it is not based on a statistic that is approximately distributed as, for example, Chi Square. However, because it assumes that both marginal totals are fixed, it can be considerably less powerful than the Chi Square test. Even though the Chi Square test is an approximate test, the approximation is quite good in most cases and tends to have too low a Type I error rate more often than too high a Type I error rate.

Rank Randomization: Two Conditions (Mann-Whitney U, Wilcoxon Rank Sum)

by David M. Lane

Prerequisites

- Chapter 5: Permutations and Combinations
- Chapter 17: Randomization Tests for Two Conditions

Learning Objectives

1. State the difference between a randomization test and a rank randomization test
2. Describe why rank randomization tests are more common
3. Be able to compute a Mann-Whitney U test

The major problem with randomization tests is that they are very difficult to compute. Rank randomization tests are performed by first converting the scores to ranks and then computing a randomization test. The primary advantage of rank randomization tests is that there are tables that can be used to determine significance. The disadvantage is that some information is lost when the numbers are converted to ranks. Therefore, rank randomization tests are generally less powerful than randomization tests based on the original numbers.

There are several names for rank randomization tests for differences in central tendency. The two most common are the Mann-Whitney U test and the Wilcoxon Rank Sum Test

Consider the data shown in Table that were used as an example in the section on *randomization tests*.

Table 1. Fictitious data.

Experimental	Control
7	0
8	2
11	5
30	9

A rank randomization test on these data begins by converting the numbers to ranks.

Table 2. Fictitious data converted to ranks. Rank sum = 24.

Experimental	Control
4	1
5	2
7	3
8	6

The probability value is determined by computing the proportion of the possible arrangements of these ranks that result in a difference between ranks of as large or larger than those in the actual data (Table 2). Since the sum of the ranks (the numbers 1-8) is a constant (36 in this case), we can use the computational shortcut of finding the proportion of arrangements for which the sum of the ranks in the Experimental Group is as high or higher than the sum here (4 + 5 + 7 + 8) = 24.

First, consider how many ways the 8 values could be divided into two sets of 4. We can apply the formula from the section on *Permutations and Combinations* for the number of combinations of n items taken r at a time (n = the total number of observations; r = the number of observations in the first group) and find that there are 70 ways.

$$C_r^n = \frac{n!}{(n-r)!\,r!} = \frac{8!}{(8-4)!\,4!} = 70$$

Of these 70 ways of dividing the data, how many result in a sum of ranks of 24 or more? Tables 3-5 show three rearrangements that would lead to a rank sum of 24 or larger.

Table 3. Rearrangement of data converted to ranks. Rank sum = 26.

Experimental	Control
6	1
5	2
7	3
8	4

Table 4. Rearrangement of data converted to ranks. Rank sum = 25.

Experimental	Control
4	1
6	2
7	3
8	5

Therefore, the actual data represent 1 arrangement with a rank sum of 24 or more and the 3 arrangements represent three others. Therefore, there are 4 arrangements with a rank sum of 24 or more. This makes the probability equal to 4/70 = 0.057. Since only one direction of difference is considered (Experimental larger than Control), this is a *one-tailed* probability. The *two-tailed* probability is (2)(0.057) = 0.114, since there are 8/70 ways to arrange the data so that the sum of the ranks is either (a) as large or larger or (b) as small or smaller than the sum found for the actual data.

The beginning of this section stated that rank randomization tests were easier to compute than randomization tests because tables are available for rank randomization tests. Table 6 can be used to obtain the critical values for equal sample sizes of 4-10.

For the present data, both n_1 and $n_2 = 4$ so, as can be determined from the table, the rank sum for the Experimental Group must be at least 25 for the difference to be significant at the 0.05 level (one-tailed). Since the sum of ranks equals 24, the probability value is somewhat above 0.05. In fact, by counting the arrangements with the sum of ranks greater than or equal to 24, we found that the probability value is 0.057. Naturally a table can only give the critical value rather than the p value itself. However, with a larger sample size, such as 10 subjects per group, it becomes very time consuming to count all arrangements equalling or exceeding the rank sum of the data. Therefore, for practical reasons, the critical value sometimes suffices.

Table 5. Rearrangement of data converted to ranks. Rank sum = 24.

Experimental	Control
3	1
6	2
7	4
8	5

Table 6. Critical values.

One-Tailed Test Rank Sum for Higher Group							
n_1	n_2	0.20	0.10	0.05	0.025	0.01	0.005
4	4	22	23	25	26	.	.
5	5	33	35	36	38	39	40
6	6	45	48	50	52	54	55
7	7	60	64	66	69	71	73
8	8	77	81	85	87	91	93
9	9	96	101	105	109	112	115
10	10	117	123	128	132	136	139

For larger sample sizes than covered in the tables, you can use the following expression that is approximately normally distributed for moderate to large sample sizes.

$$Z = \frac{W_a - n_a(n_a + n_b + 1)/2}{\sqrt{n_a n_b (n_a + n_b + 1)/12}}$$

where:
W_a is the sum of the ranks for the first group
n_a is the sample size for the first group
n_b is the sample size for the second group
Z is the test statistic

The probability value can be determined from Z using the *normal distribution calculator*.

The data from the *Stereograms Case Study* can be analyzed using this test. For these data, the sum of the ranks for Group 1 (W_a) is 1911, the sample size for Group 1 (n_a) is 43, and the sample size for Group 2 (n_b) is 35. Plugging these values into the formula results in a Z of 2.13, which has a two-tailed p of 0.033.

Rank Randomization: Two or More Conditions (Kruskal-Wallis)

by David M. Lane

Prerequisites

- Chapter 17: Chi Square Distribution
- Chapter 18: Randomization Test for Two or More Conditions
- Chapter 18: Rand Randomization (Two Groups)

Learning Objectives

1. Compute the Kruskal-Wallis test

The Kruskal-Wallis test is a rank-randomization test that extends the Wilcoxon test to designs with more than two groups. It tests for differences in central tendency in designs with one between-subjects variable. The test is based on a statistic H that is approximately distributed as Chi Square. The formula for H is shown below:

$$H = -3(N+1) + \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i}$$

where

N is the total number of observations,
T_i is the sum of ranks for the i^{th} group,
n_i is the sample size for the i^{th} group,
k is the number of groups.

The first step is to convert the data to ranks (ignoring group membership) and then find the sum of the ranks for each group. Then, compute H using the formula above. Finally, the significance test is done using a Chi Square distribution with k-1 degrees of freedom.

For the "Smiles and Leniency" case study, the sum of the ranks for the four conditions are:

False:	2732.0
Felt:	2385.5

```
Miserable: 2424.5
Neutral:   1776.0
```

Note that since there are "ties" in the data, the mean rank of the ties is used. For example, there were 10 scores of 2.5 which tied for ranks 4-13. The average of the numbers 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 is 8.5. Therefore, all values of 2.5 were assigned ranks of 8.5.

The sample size for each group is 34.

$$H = -3(136+1) + \frac{12}{(136)(137)}\left(\frac{2732^2}{34} + \frac{2385.5^2}{34} + \frac{2424.5^2}{34} + \frac{1776^2}{34}\right) = 9.28$$

Using the *Chi Square Calculator* (external link; requires Java) for Chi Square = 9.28 with 4-1 = 3 df results in a p value of 0.028. Thus the null hypothesis of no leniency effect can be rejected.

Rank Randomization for Association (Spearman's ρ)

by David M. Lane

Prerequisites

- Chapter 4: Values of Pearson's r
- Chapter 18: Randomization Test for Pearson's r

Learning Objectives

1. Compute Spearman's ρ
2. Test Spearman's ρ for significance

The rank randomization test for association is equivalent to the *randomization test for Pearson's r* except that the numbers are converted to ranks before the analysis is done. Table 1 shows 5 values of X and Y. Table 2 shows these same data converted to ranks (separately for X and Y).

Table 1. Example data.

X	Y
1.0	1.0
2.4	2.0
3.8	2.3
4.0	3.7
11.0	2.5

Table 2. Ranked data.

X	Y
1	1
2	2
3	3
4	5
5	4

The approach is to consider the X variable fixed and compare the correlation obtained in the actual ranked data to the correlations that could be obtained by rearranging the Y variable. For the data shown in Table 2, the correlation between X and Y is 0.90. The correlation of ranks is called “Spearman's ϱ.”

There is only one arrangement of Y that produces a higher correlation than 0.90: A correlation of 1.0 results if the fourth and fifth observations' Y values are switched (see Table 3). There are also three other arrangements that produce an r of 0.90 (see Tables 4, 5, and 6). Therefore, there are five arrangements of Y that lead to correlations as high or higher than the actual ranked data (Tables 2 through 6).

The next step is to calculate the number of possible arrangements of Y. The number is simply N!, where N is the number of pairs of scores. Here, the number of arrangements is 5! = 120. Therefore, the probability value is 5/120 = 0.042. Note that this is a one-tailed probability since it is the proportion of arrangements that give a correlation as large or larger. The two-tailed probability is 0.084.

Since it is hard to count up all the possibilities when the sample size is even moderately large, it is convenient to have a table of critical values.

From the table shown below, you can see that the critical value for a one-tailed test with 5 observations at the 0.05 level is 0.90. Since the correlation for the sample data is 0.90, the association is significant at the 0.05 level (one-tailed). As shown above, the probability value is 0.042. Since the critical value for a two-tailed test is 1.0, Spearman's ϱ is not significant in a two-tailed test.

N	.05 2-tail	.01 2-tail	.05 1-tail	.01 1-tail
5	1.000		0.900	1.000
6	0.886	1.000	0.829	0.943
7	0.786	0.929	0.714	0.893
8	0.738	0.881	0.643	0.833
9	0.700	0.833	0.600	0.783
10	0.648	0.794	0.564	0.745
11	0.618	0.755	0.536	0.709
12	0.587	0.727	0.503	0.671
13	0.560	0.703	0.484	0.648
14	0.538	0.675	0.464	0.622
15	0.521	0.654	0.443	0.604
16	0.503	0.635	0.429	0.582
17	0.485	0.615	0.414	0.566
18	0.472	0.600	0.401	0.550
19	0.460	0.584	0.391	0.535

20	0.447	0.570	0.380	0.520
21	0.435	0.556	0.370	0.508
22	0.425	0.544	0.361	0.496
23	0.415	0.532	0.353	0.486
24	0.406	0.521	0.344	0.476
25	0.398	0.511	0.337	0.466
26	0.390	0.501	0.331	0.457
27	0.382	0.491	0.324	0.448
28	0.375	0.483	0.317	0.440
29	0.368	0.475	0.312	0.433
30	0.362	0.467	0.306	0.425
31	0.356	0.459	0.301	0.418
32	0.350	0.452	0.296	0.412
33	0.345	0.446	0.291	0.405
34	0.340	0.439	0.287	0.399
35	0.335	0.433	0.283	0.394
36	0.330	0.427	0.279	0.388
37	0.325	0.421	0.275	0.383
38	0.321	0.415	0.271	0.378
39	0.317	0.410	0.267	0.373
40	0.313	0.405	0.264	0.368
41	0.309	0.400	0.261	0.364
42	0.305	0.395	0.257	0.359
43	0.301	0.391	0.254	0.355
44	0.298	0.386	0.251	0.351
45	0.294	0.382	0.248	0.347
46	0.291	0.378	0.246	0.343
47	0.288	0.374	0.243	0.340
48	0.285	0.370	0.240	0.336
49	0.282	0.366	0.238	0.333
50	0.279	0.363	0.235	0.329

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 1: Levels of Measurement
- Chapter 18: Benefits
- Chapter 18: Rank Randomization for Two Conditions,

Cardiac troponins are markers of myocardial damage. The levels of troponin in subjects with and without signs of right ventricular strain in the electrocardiogram were compared in the experiment described here.

The Wilcoxon rank sum test was used to test for significance. The troponin concentration in patients with signs of right ventricular strain was higher (median = 0.03 ng/ml) than in patients without right ventricular strain (median < 0.01 ng/ml), $p<0.001$.

What do you think?

Why might the authors have used the Wilcoxon test rather than a t test? Do you think the conclusions would have been different?

> Perhaps the distributions were very non-normal. Typically a transformation can be done to make a distribution more normal but that is not always the case. It is almost certain the same conclusion would have been reached, although it would have been described in terms of mean differences instead of median differences.

Exercises

Prerequisites

All of this chapter

1. For the following data, how many ways could the data be arranged (including the original arrangement) so that the advantage of the Experimental Group mean over the Control Group mean is as large or larger then the original arrangement.

Experimental	Control
5	1
10	2
15	3
16	4
17	9

2. For the data in Problem 1, how many ways can the data be rearranged?

3. What is the one-tailed probability for a test of the difference.

4. For the following data, how many ways can the data be rearranged?

T1	T2	Control
7	14	0
8	19	2
11	21	5

5. In general, are rank randomization tests or randomization tests more powerful?

6. What is the advantage of rank randomization tests over randomization tests?

7. Test whether the differences among conditions for the data in Problem 1 is significant (one tailed) at the .01 level using a rank randomizaton test.

Questions from Case Studies

SAT and GPA (SG) case study

8. (SG) Compute Spearman's ϱ for the relationship between UGPA and SAT.

Stereograms (S) case study

9. (S) Test the difference in central tendency between the two conditions using a rank-randomization test (with the normal approximation) with a one-tailed test. Give the Z and the p.

Smiles and Leniency (SL) case study

10. (SL) Test the difference in central tendency between the four conditions using a rank-randomization test (with the normal approximation). Give the Chi Square and the p.