5

Randomization Tests

5.1 Introduction

When used to test hypotheses, standard parametric statistics such as analysis of variance (ANOVA) require the samples and data involved to adhere to at least one restrictive assumption. If data fail to meet the conditions laid down in such assumptions, any conclusions drawn from the analyses can be suspect. Randomization methods can also be used to test hypotheses but require fewer assumptions. Given this extra flexibility, it is surprising that there is not a greater awareness of randomization tests. In this chapter we will examine randomization or permutation testing and its potential value to ecological and fisheries research.

5.2 Hypothesis Testing

5.2.1 introduction

Many hypothesis tests involve determining the likelihood that the situation one has observed could have arisen by chance events alone. What this means is that one is testing whether an observed pattern is unusual. But something can only be unusual relative to something else, usually one sample relative to one or more other samples, or relative to a hypothesis of how the sample ought to be. This is tested by determining whether there is a good chance that a perceived pattern between groups could have come about by random variations of whatever components make up the pattern of interest. For example, one could ask whether any differences observed in the size frequencies of a fish species found on the open coast and in coastal bays arose due to chance mixing or if the observed pattern is highly unlikely to be due to random movements between sites.

5.2.2 Standard Significance Testing

When attempting to test a hypothesis, at some stage one invariably calculates a test statistic. Given a value for a particular test statistic (e.g., a t statistic comparing the difference between two means), one then wishes to know if the value one has denotes a significant difference. This is such a common event that it is simple to forget what lies behind the steps leading to the test statistic and the meaning of its test of significance. Three things are needed when testing a hypothesis statistically:

1. A hypothesis ideally should be stated formally and is often translated into a null hypothesis that is the inverse or complement of the hypothesis we wish to test. This inversion is necessary because of the logical asymmetry between disproof of a general statement (possible) and its proof (not possible). Hypotheses are testable theoretical constructs, e.g., average fork lengths of fish from the open coast and coastal bays are the same (are not different).
2. A test statistic, any single-valued (smooth and continuous) function of the data. Commonly used test statistics include the F ratio, the t statistic, χ^2, and r the correlation coefficient.
3. Some means of generating the probability distribution of the test statistic under the assumption that the null hypothesis is true is required to permit a determination of how likely the observed value of the test statistic is if the null hypothesis is true.

The first two requirements are straightforward and need little discussion. The third requirement may not be immediately familiar but is strongly related to the second requirement of using a particular test statistic. The many commonly used parametric statistics are used because their underlying assumptions are known, which means the expected probability density function (e.g., Figure 5.1) of a test statistic under the conditions of a particular situation can be calculated analytically. The value of the test statistic obtained from nature can be compared with the expected values given the circumstances of the particular test (degrees of freedom, etc.).

In this way, assuming the sampled population adheres to the assumptions of the test statistic, it is possible to determine how likely the observed value would be assuming the null hypothesis to be true. For common parametric test statistics the comparison of calculated values with the theoretical probability density function is made simple by its translation into tables relating degrees of freedom to likelihood or significance (Figure 5.1).

By convention, if the value of a test statistic is only expected to occur less than once in twenty replicated samples (i.e., <5% of the time), then usually a "significant" difference is claimed. This is taken to mean that unusual patterns (e.g., four heads in a row when tossing a coin) are not impossible; they are just unlikely chance events relative to the many other possible events

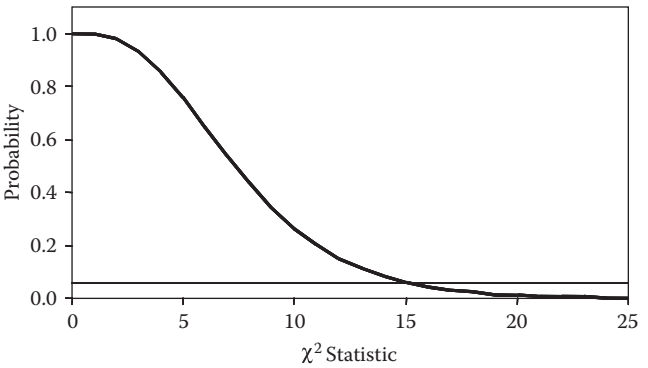

Figure 5.1
A one-tailed probability distribution curve (probability density function (pdf)) for the χ^2 statistic with 8 degrees of freedom. The fine horizontal line indicates a probability of 0.05. One could utilize this curve to determine the likelihood of obtaining a χ^2 value as large as the one observed between two populations, if one assumes they are from the same population. An observed χ^2 statistic greater than 15.507 would be expected to occur less than one time in twenty (i.e., <5% of the time). The shape of the curve varies with the degrees of freedom.

EXAMPLE BOX 5.1

Three different test statistics and their respective cumulative density functions indicating the probability that the value of the observed test statistic for the observed pattern being tested could have arisen through chance alone. Examine the Excel help files for descriptions of the functions used. Try varying the test statistic value and the degrees of freedom for each statistic and see the response. For the χ^2 statistic with 1 degree of freedom a value of approximately 3.84 should give a probability of 0.05. Compare the outputs from the Excel functions with published tables.

	A	B	C	D
1	Test statistic	χ^2	**F**	**t**
2	Test statistic value	2.9	2.9	1.98
3	Degrees of freedom 1	1	1	120
4	Degrees of freedom 2		25	
5	Probability	=chidist(B2,B3)	=fdist(C2,C3,C4)	=tdist(D2,D3,2)

that can occur. Because convention has established the levels of significance at 5%, 1% (1 chance in 100), and 0.1% (1 chance in 1,000), values of test statistics at these levels tend to be printed in statistical tables (Example Box 5.1). By comparing the observed value of the test statistic with these published values, one determines whether one has a significant difference and at what level.

5.2.3 Significance Testing by r andomization Test

Unfortunately, the theoretically derived tables of the various test statistics are only valid if the data adhere to the assumptions of the statistical test employed. The important problem is that if the assumptions are not met, then the analytically derived probability density function will not usually be applicable validly and erroneous conclusions can be made. The effect of such failure to match assumptions is an increased chance of rejecting a real difference between groups or of accepting a difference where none exists. This is one reason it behooves an ecologist to know about the effects of data transformations. For a *t* test, the assumptions would be that the two groups being compared are independently and normally distributed random variables with constant means and variances. The only part of this hypothesis that one really wishes to test is the "independently distributed" part (i.e., they come from independent populations), but all the rest is necessary, else the probability density distribution of the test statistic (the *t* distribution) cannot be used validly (though it is, in fact, relatively robust to departures from these assumptions).

We need to know the probability distribution of the selected test statistic to determine the likelihood of the value observed/calculated assuming the null hypothesis to be true. A problem with using a theoretically derived probability density function (pdf), as in the *t* test example above, is that if the observed test statistic value is not significant, we cannot tell without further analyses whether the test failed because the samples are not independently distributed (the thing being tested) or because the samples were not from normally distributed populations (the data or sampled populations failed to conform to the assumptions necessary for the test to be valid). This is where randomization tests exhibit their strength. They are independent of any analytically determined (parametric) probability density function because, during the test, the randomization procedure generates an empirical pdf for the test statistic from the available data (Manly, 1997).

If the hypothesis to be tested claims to explain a particular pattern in a set of data (e.g., inshore fish sizes are smaller on average than those offshore), then the null hypothesis would claim that the observed pattern found in the data is typical of any random allocation of the available data (fish sizes) among the inshore and offshore groups. The original pattern observed is represented in the test by a single value of a test statistic (perhaps a *t* statistic). Obviously, the test statistic should be chosen to be sensitive to the pattern of interest.

Given a null hypothesis, the expected probability density function for the test statistic chosen can be generated by repeatedly randomizing the data with respect to the sample group membership and recalculating the test statistic. In the inshore/offshore example, the membership of inshore and offshore groups is randomized. Fisher (1936) would have suggested that all the fish lengths should be written on separate cards, the cards shuffled, and

then the pack dealt out into the inshore/offshore groups in their original relative frequencies. When this process of randomization and calculation of test statistic is done many times (generally a minimum of one thousand randomizations are used), the frequency of occurrence of different values of the test statistic can be tabulated, and this may be compared with the value observed from the original unrandomized data. If the original value of the test statistic is found to be an unusual event relative to the values generated by the permutations of the data, then the null hypothesis may be rejected. The null hypothesis is, in effect, that the groups being compared are random samples from the same population. Speaking in terms of shuffling cards, Fisher (1936, p. 59) wrote: "Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."

A test of significance is thus really an attempt to answer whether the observed samples could have been drawn at random from the same population. The answer is a probability, and if it is large, then it is likely that the pattern could have arisen due to chance and one could answer that the samples might have derived from the same population (we could never claim they definitely were from the same population). However, if the probability of the pattern arising through randomly sampling the same population is small, we can be more definite and claim they were not likely to be from the same population. This asymmetry is why one uses a falsifiable null hypothesis.

5.2.4 Mechanics of r andomization Tests

An example will illustrate the ideas that have been discussed and hopefully make the methods clearer (Figure 5.2, Example Box 5.2).

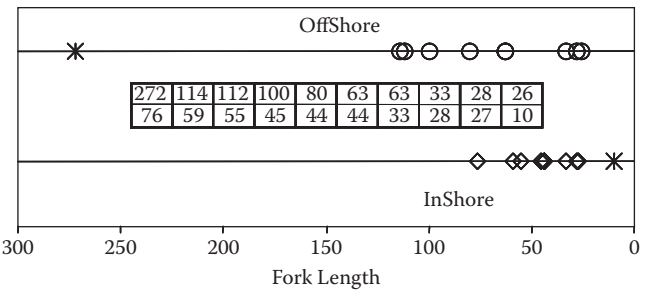

Fig ur e 5.2
Two samples showing the data both numerically and graphically. Note the extreme values specially marked in both groups. Conventional *t* tests of the two groups, both including and excluding the extreme values (and assuming unequal variances), indicate that the groups are not significantly different despite the mean difference being 47. With the extreme values included, $t = 1.99$ and $P = 0.0746$, and without the extreme values, $t = 1.8$ and $P = 0.0993$. The samples are clearly too small in this example.

EXAMPLE BOX 5.2

The mechanics of a randomization procedure to compare two groups of data. The artificial data are from Figure 5.2. A copy of the original data is placed in column A (see Figure 5.2 for all data). Column B starts with a copy from column A; column C is filled with random numbers from the function =rand(). Cells C1 and D1 contain the average of the first and second groups, respectively, while E1 contains the test statistic, the absolute mean difference. The objective is to test whether the observed samples could have arisen by chance from a single population. To preserve a copy, the original values are pasted from column A into B and the observed test statistic pasted into D5. The randomization works by sorting columns B and C relative to the column of random numbers (do not include column A in the sorting). This has the effect of randomly reordering the available data into the two groups (inshore and offshore). The test statistic alters accordingly, and each new value is copied by a macro into column D.

Create a macro using Tools/Macro/Record New Macro menu item. Call it Do_Rand. Make sure the Stop Recording toolbar is visible (View/Toolbars) and start by using absolute references. Press <Shift><F9> to recalculate the sheet. Select and copy A5:A24 and paste their values into B5:B24. Copy E1 and paste its value into D5. Select B5:C24 and sort them on column C. Copy E1. Switch to relative references and paste the values into cell D6. Stop recording the macro. Press <Alt ><F11> and maximize the macro window. The modifications necessary to the macro are shown on the next page. Assign the Do_Rand macro to a button created from the Forms toolbar (not the Control Toolbox). If you change the number of iterations in the macro you will also need to change cell D3. Exclude the extreme values by altering C1 to B6:B14, D1 to B15:B23, and D3 to ">=23.1111", and alter the sort command in the macro.

	A	B	C	D	E
1	Do_Rand		=average(B5:B14)	=average(B15:B24)	=abs(C1-D1)
2					
3			P =	=countif(D5:D1004,">=47")/1000	
4	**Original**	**Values**	**Randomize Rows**		
5	272	272	=rand()		
6	114	114	=rand()		
7	112	112	=rand()		
8	100	100	Copy down		
9	80	80	To row 24		
10	63	63	0.729468		
11	63	63	0.450439		
12	33	33	0.023943		
13	Continue down, data from Figure 5.2				

continued

EXAMPLE BOX 5.2 (continued)

The macro contents as recorded following the directions in the first part of this example box. You will need to add the lines and changes shown in italics. The only tricky bit is the replacement of the step six rows up from the bottom in the ActiveCell.Offset statement to become the dynamic 4 + i, thereby using the counter variable to identify the pasting location for the results from the randomizations. To conduct the randomizations without the extreme values, alter the calculations for the group averages and alter the select command just prior to the sort to read range("B6:C23").select. This isolates the extreme values and the results should be rather different. Try modifying the macro by omitting commands such as the Application.Screenupdating=False. How many replicates are needed to obtain consistent results?

```
Sub Do_Rand()
' Do_Rand Macro
  Range("A5:A24").Select
  Selection.Copy      ' copy original data
  Range("B5").Select
  Selection.PasteSpecial Paste:=xlPasteValues
  Range("E1").Select
  Selection.Copy      ' store the original test statistic
  Range("D5").Select
  Selection.PasteSpecial Paste:=xlPasteValues
     ' unnecessary defaults removed
  Application.CutCopyMode = False
  Application.ScreenUpdating = False   ' for speed
    For i = 1 To 999  ' plus the original = 1000
      Range("B5:C24").Select
      Selection.Sort Key1:=Range("C5"),
        Order1:=xlAscending, Header:=xlGuess, _
      OrderCustom:=1, MatchCase:=False,
        Orientation:=xlTopToBottom
      Range("E1").Select
      Selection.Copy   ' copy and store test replicates.
      ActiveCell.Offset(4 + i, -1).Range("A1").Select
          ' use counter to identify cells.
      Selection.PasteSpecial Paste:=xlPasteValues
    Next i
  Range("A1").Select  ' returns to top of sheet
  Application.ScreenUpdating = True
  End Sub
```

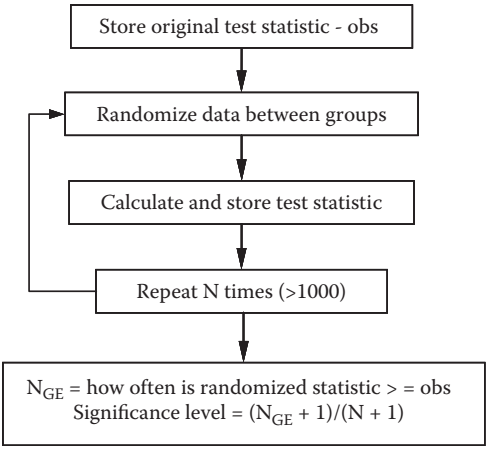

Fig ur e 5.3
Algorithm for conducting a randomization test and for calculating the consequent significance test. One always includes the original test statistic value; this prevents a probability of zero occurring.

The group allocations (inshore/offshore) of the sample observations were randomized one thousand times and the mean difference was recalculated (algorithm shown in Figure 5.3). The randomization test found that the original mean difference only occurred twenty-five times out of one thousand, so evidence of a real difference exists (Figure 5.4). If this analysis is repeated, then a number slightly different from twenty-five might arise, but a significant difference should still be found. It is the weight of evidence that matters, not whether we have a difference significant at 5%, 1%, or whatever. The probability value indicates how likely it was to obtain a value like the observed test statistic. When outliers in the data are ignored (effectively removed from the data set), the randomization test agrees very closely with the conventional *t* test.

When the algorithm in Figure 5.3 is followed, one obtains *N* randomized replicate values. When these are sorted in ascending order, they may be plotted to give a visual representation of the distribution of values (Figure 5.4, Example Box 5.2).

5.2.5 Selection of a Test Statistic

With the release from the requirement of having one's test statistic adhere to a theoretical probability distribution there comes the freedom to select the most appropriate statistic. A test statistic should be chosen because its value is sensitive to the substantive theory being tested (i.e., the hypothetical property being checked or compared). However, some researchers see this freedom as a disadvantage because with the same sample, different test

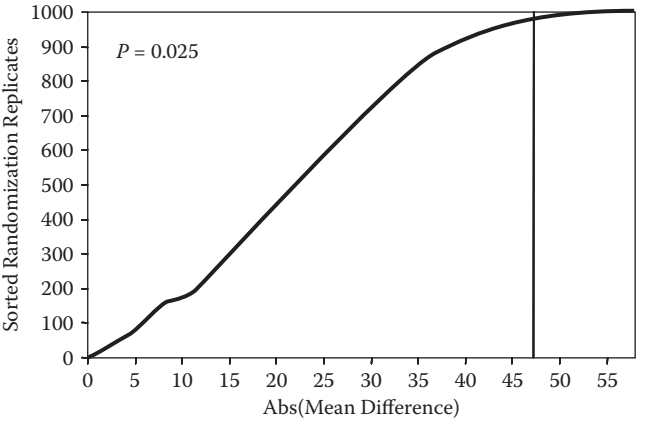

Figure 5.4
One thousand sorted randomization replicates of the absolute mean difference between the groups in Figure 5.2, using the algorithm in Figure 5.3. The outer vertical line is at 47, the original mean difference. Including the original 47, there were twenty-five replicates that ≥47 (i.e., lie on or outside 47); therefore, $P = 0.025$. With the extreme values excluded the P value tends to be more like 0.094 (similar to the parametric t test). More replicates would lead to a smoother curve.

statistics can give different levels of significance. They are mistaken in thinking this invalidates the process. Rather, it emphasizes that one should be careful when selecting and using a nonstandard test statistic.

Basu (1980) was aiming to criticize Fisher and took as an example a simple data set of five values with which it was desired to determine whether the expected population value was significantly different from zero. The data were a set of five numbers, the values of which could have been either positive or negative but were, in fact, (1, 2, 3, 4, 7), which have a mean of 3.4. Basu (1980) argued that the question of whether this is significantly different from zero can be tested by a randomization test. He states that the significance of this test (is the sample mean > 0?) can be determined by comparing the mean with all the means possible from the thirty-two possible unique combinations of (±1, ±2, ±3, ±4, ±7). It is quickly shown that a mean of 3.4 is the largest value out of all possible values with the data available, and so is significant at the 1/32 level ($P = 0.0313$). However, the median is 3, and when used as the test statistic, there are four combinations, namely, (±1, ±2, 3, 4, 7), with a median as high as 3, implying a significance of 4/32 = 1/8 = 0.125, which is not generally recognized as being significant. This disparity between the mean and median works for any set of five positive numbers tested in this way. In the end, this does not imply that the randomization procedure is not workable, but instead that this particular test statistic, the median, can be unstable and should not be used for such purposes with such limited data (Example Box 5.3).

EXAMPLE BOX 5.3

An example (from Basu, 1980) of a randomization procedure to compare two test statistics. Column B is filled with the random number generator =rand(). Cells C1 and D1 contain the average and median of the top 5 cells (A5:A9). The objective is to test whether the collection shown is significantly different from zero when the values can be (±1, ±2, ±3, ±4, ±7). The observed values of the two test statistics are copied as values into C5 and D5. The randomization works by sorting columns A and B on the column of random numbers (keep a copy of the raw data and its order somewhere safe on the worksheet). This has the effect of randomly reordering the available data into the two groups (used and unused). The test statistics alter accordingly, and these new values are copied down below the starting observed values. After sufficient replicates we can count the number that are the same as or greater than the original observed values. Create a macro using Tools/Macro/Record New Macro menu item. Call it Do_Randz and start by using absolute references. Select C2:D2 and delete the contents. Press <Shift><F9>. Select A5:B14 and sort them on column B. Copy C1:D1. Switch to relative references and paste the values into cell C6 (and D6). Switch to absolute references. Select C2 and type =countif(C5:C1004,">=3.4"), and select D2 and type =countif(D5:D1004,">=3"). Stop recording the macro. Press <Alt><F11> and maximize the macro window. The modifications necessary to the macro are shown on the next page. Assign the Do-Randz macro to a button created from the Forms toolbar. If you change the number of iterations, you will also need to change cells C3 and D3.

	A	B	C	D	E
1			=average(A5:A9)	=median(A5:A9)	
2		**Do_Rand**	1	10	
3			=C2/1000	=D2/1000	
4	**Values**	**Reorder**			
5	7	=rand()	3.4	3	
6	4	=rand()			
7	3	=rand()			
8	2	Copy Down			
9	1	To Row 14			
10	–1	0.729468			
11	–2	0.450439			
12	–3	0.023943			
13	–4	0.97243			
14	–7	0.312233			

continued

EXAMPLE BOX 5.3 (continued)

The macro contents as recorded following the directions in the first part of this example box. You will need to add the lines and changes shown in italics. You can either delete the struck-out text or leave it, whichever you prefer. If you decide on more than one thousand iterations, then you will need to alter the COUNTIF statements to reflect your choice. The only tricky bit is the replacement of the step ten rows up from the bottom in the ActiveCell.Offset statement to become the dynamic 4 + i, thereby using the counter variable to identify the pasting location for the results from the randomizations. Try running the macro and comparing the significance of the comparison with a mean of zero when using the average and when using the median. There is a clear difference. In the text, complete evaluations of all possible combinations were discussed. In small discrete cases like this, complete evaluation is a reasonable option, but with larger numbers of observations a complete evaluation becomes onerous and a random sampling provides sufficient resolution (Manly, 1997).

```
Sub Do_Randz()
' Do_Randz Macro
Dim i As Integer                    ' Not strictly needed
'
Application.ScreenUpdating = False  ' vital for sanity
  Range("C2:D2").Select
  Selection.ClearContents           ' saves recalculating
For i = 1 To 1000                   ' usually >=1000
  ActiveSheet.Calculate             ' new random numbers
  Range("A5:B14").Select
  Selection.Sort Key1:=Range("B5"), Order1:=xlDescending,
    Header:=xlGuess, _
  OrderCustom:=1, MatchCase:=False,
    Orientation:=xlTopToBottom
  Range("C1:D1").Select
  Selection.Copy
  ' Replace 5 with 4 + i so it reads:
  ActiveCell.Offset(4 + i, 0).Range("A1").Select
  Selection.PasteSpecial Paste:=xlPasteValues,
    Operation:=xlNone, SkipBlanks:= _
    False, Transpose:=False     ' defaults can be deleted
Next i
  Range("C2").Select
  Application.CutCopyMode = False
```

continued

EXAMPLE BOX 5.3 (continued)

```
    ActiveCell.FormulaR1C1 = "=COUNTIF(R[3]C:R[1004]
      C,""">=3.4""")"
    Range("D2").Select
    ActiveCell.FormulaR1C1 =
      "=COUNTIF(R[3]C:R[1004]C,""">=3""")".
Range("A1").Select              ' returns to top of sheet
Application.ScreenUpdating = True
End Sub
```

Examples of nonstandard test statistics could include the differences between parameter estimates of two growth curves ($K_1 - K_2$, or $L_{\infty 1} - L_{\infty 2}$, etc., although a randomization test to compare growth curves should not focus on single parameters and would have other complications; see later example in Chapter 9). Other examples could include the comparison of parameters from other models, such as stock recruitment relationships. Such direct comparisons would not be possible with traditional statistics. One can investigate questions that were previously untestable. For example, Lento et al. (1997) examined haplotype diversity across the geographical range of southern fur seal species using a randomization test for testing the reality of apparent genetic structure between populations. This tests whether the H_{st} value (a measure of the diversity and evenness of genetic variation between populations) provides evidence of geographical structure.

Multivariate comparisons are also possible (Anderson et al., 2008) and are limited only by imagination. Having said that, one must be extremely careful to determine exactly which hypothesis is being tested by the test statistic developed; when in doubt, be conservative.

5.2.6 ideal Test Statistics

Ideally one should select the test statistic that has the greatest statistical power for the situation being studied. This relates to the two types of error (Table 5.1). Making decisions under circumstances of uncertainty is very much concerned with the different types of statistical inference error it is possible to make. The implications of making each type of error in a particular situation should be made explicit before deciding on a test statistic.

If the test had been about the effects of fishing and the null hypothesis that fishing had negligible effects was incorrectly accepted, this would be a type II error that could have dangerous implications. No mitigating management actions would be implemented in the false belief that the stock was healthy. Conversely, if it were incorrectly concluded that stock damage was accruing when it was not (type I error), then managers may unjustifiably reduce the potential earnings of fishers (null hypothesis would be that no

TABLe 5.1

Type I and Type II Error Types

	No Differences Exist Null True	**Differences Exist Null False**
Null Accepted	OK	Type II error
Null Rejected	Type I error	OK

Source: After Sokal and Rohlf, 1995.

stock damage was detectable and this is mistakenly deemed false). It is hard to escape the conclusion that in the past more type II errors than type I errors have been made in fisheries management.

The *significance* of a test is the probability of making a type I error. Because we would expect a particular test statistic value to arise about one time in twenty at a probability of 0.05, accepting a significance level of 5% implies that one time in twenty we are likely to be claiming to have found a significant difference between groups where one does not exist (type I error). This is why a difference that is significant at the 5% level is less convincing than one at the 0.1% level.

The *power* of a test is the complement of the probability of making a type II error. To make a type II error is to claim no differences when differences exist; therefore, the complement of this probability is that of deciding when differences exist (the complement of δ is $1 - \delta$). In short, the power of a test is the probability of making the correct decision.

In practice, one would fix the significance level as small or large as is acceptable and then choose a statistic that maximizes the power of the test. A test is said to be *unbiased* if in using the test one is more likely to reject a false hypothesis than a true one (Good, 1994).

5.3 Randomization of Structured Data

5.3.1 introduction

The fact that randomization tests are restricted to tests that are basically comparisons between groups may be considered a major limitation; for example, they cannot be used for parameter estimation. On the other hand, they are very good at comparison tests.

Not every comparison is one of means; sometimes we would wish to compare variation between samples. It turns out that simple randomization tests of a mean difference between two samples can be influenced by differences in variation, particularly if this is related to sample size. For comparisons of variation, Manly (1997) recommends that one should randomize

the residuals of the two samples instead of the original data values. That effectively standardizes the means to zero and concentrates the test upon the variation within each sample.

Many comparisons are of more than two groups, and the conventional approach is to analyze these using ANOVA. Randomization tests have been used in conjunction with ANOVA for a few decades. They are best used when the assumptions of ANOVA fail badly. The basic rule is that with unbalanced and highly nonnormal data, one should use a randomization procedure to determine the significance of one's analyses.

With more complex ANOVA models, such as three-way, perhaps with nested factors, or repeated measures, there is some debate in the literature over the correct procedure to use in a randomization test. The question at issue is whether one should just randomize one's data across all categories and treatments (Manly (1997) says yes, Edgington (1987) emphatically says no), or whether there should be some restrictions placed upon what is randomized (randomize the treatments and categories separately). Manly also claims that one can test for interactions as one would normally, while Edgington claims the contrary and says one cannot test interactions. Edgington (1995) continues the discussion and provides many insights into the analysis of structured data. Investigations into the most efficient method of conducting randomization tests on structured data are an open field in need of further research (Anderson and Legendre, 1999). The question that tends to be addressed when considering structured data is whether one should randomize the raw data, subsets of the raw data, or residuals from the underlying model. There is no simple answer to this, as the debate between Manly and Edgington demonstrates.

What this all means is that when one's data are structured or nonlinear (often the case with fisheries models and relationships), care needs to be taken in deciding what components should be randomized during a test. Anderson and Legendre (1999, p. 302) capture the present situation when they conclude the following:

> Obtaining empirical measures of type I error or power allows direct practical comparisons of permutation methods. Current theoretical comparisons of the methods cannot provide us with complete information on how the methods will compare in different situations in practice.

5.3.2 More Complex examples

Consider the problem of comparing growth curves (this will be considered in detail in Chapter 9, on growth). If one wanted to compare age length data from two populations of fish, then one might fit two von Bertalanffy curves and wish to determine whether any of the parameters of the two curves are significantly different. At present one would generally use a likelihood ratio test (Kimura, 1980; see Chapter 9), but to avoid some of the assumptions

concerning underlying distributions, one could utilize a randomization test. However, it is not immediately clear what one should randomize.

One could randomize the available data pairs (ages plus associated lengths) between the two populations, keeping the number of observation pairs in each sample the same as in the original data. However, by chance many of the older animals may be collected in one population and many of the younger in the other; this would obviously distort any growth curve fitted to these randomized data. On the other hand, this randomization design could be used when testing whether the proportional age composition or length composition of the different data sets was significantly different.

If one wanted to test the difference between the average growth curves for the two samples, then one requires a slightly more complicated scheme of randomization. The original data sets need to be stratified, in this case into discrete ages, and then the data should be randomized between populations but within age strata, thereby maintaining the original numbers of individuals within each age class for each population. In this way the underlying age structures of the two populations could be maintained while the dynamics of growth are compared. In this case, it would be important to make sure that every age class was represented in each data set, though the numbers in each age class would not need to be the same. An alternative approach, which should produce equivalent results, would be to randomize ages between populations but within length classes. This example will be considered explicitly when we consider the comparison of growth curves in Chapter 9.