CHAPTER 1

Randomization

1.1 The Idea of a Randomization Test

Many hypotheses of interest in science can be regarded as alternatives to null hypotheses of randomness. That is, the hypothesis under investigation suggests that there will be a tendency for a certain type of pattern to appear in data, whereas the null hypothesis says that if this pattern is present, then this is a purely chance effect of observations in a random order.

Randomization testing is a way of determining whether the null hypothesis is reasonable in this type of situation. A statistic S is chosen to measure the extent to which data show the pattern in question. The value s of S for the observed data is then compared with the distribution of S that is obtained by randomly reordering the data. The argument made is that if the null hypothesis is true, then all possible orders for the data were equally likely to have occurred. The observed data order is then just one of the equally likely orders, and s should appear as a typical value from the randomization distribution of S. If this does not seem to be the case (so that s is significant), then the null hypothesis is discredited to some extent and, by implication, the alternative hypothesis is considered more reasonable.

The significance level of s is the proportion or percentage of values that are as extreme or more extreme than this value in the randomization distribution. This can be interpreted in the same way as for conventional tests of significance. If s is less than 5%, then this provides some evidence that the null hypothesis is not true; if s is less than 1%, then it provides strong evidence that the null hypothesis is not true; and if s is less than 0.1%, then it provides very strong evidence that the null hypothesis is not true. To avoid the characterization of belonging to "that group of people whose aim in life is to be wrong 5% of the time" (Kempthorne and Doerfler, 1969), it is better to regard the level of significance as a measure of the strength of evidence

against the null hypothesis, rather than showing whether the data are significant at a certain level.

In comparison with more standard statistical methods, randomization tests have two main advantages. First, they are valid even without random samples. Second, it is often relatively easy to take into account the peculiarities of the situation of interest and use nonstandard test statistics.

There is a disadvantage with randomization tests that may appear at first sight to be severe: it is not necessarily possible to generalize the conclusions from a randomization test to a population of interest. What a randomization test tells us is that a certain pattern in data is or is not likely to have arisen by chance. This is completely specific to the data at hand. The concept of a population from which other samples could be taken is not needed, which is why random sampling is not required.

Some statisticians argue that the lack of a theory for generalizing the results of randomization tests to populations means that these tests have very little value, if any, in comparison to more standard tests for which well-developed methods of statistical inference exist. Others, however, suggest that in reality samples are often not really random at all, but simply consist of items that happen to be readily available. The generalization of results then rests on the assumption that the sample obtained is effectively the same as a random sample. This nonstatistical judgment is similar to the type of judgment that is made when deciding that the result of a randomization test is what can generally be expected for data collected in a particular way.

As an example, suppose that a physiologist wishes to see whether drinking alcohol in moderation has an effect on reaction times of subjects aged 20. Rather than take a random sample of all possible subjects of this age (which leads to considerable difficulties about the definition of the population, and is in any case impossible), he uses all the 20-year-old students in a university class in physiology. These are divided at random into two groups: one has reaction times measured after taking a drink with a small amount of alcohol, and the other has reaction times measured after taking an alcohol-free drink.

Various methods can be used to analyze the results of an experiment of this type. For example, if a mean difference between the test scores for the two groups is of interest, then a conventional t-test can be used to determine whether the observed difference is significantly different from zero. However, whatever the outcome of such a test is, using it to draw conclusions about the effect of alcohol on all 20-year-olds is only valid on the assumption that the 20-year-olds in the

physiology class are equivalent to a random sample of all 20-year-olds with respect to the measurement of reaction times that is used. Hence, any such generalization has to be questionable, and requires a judgment as to whether the same type of result is likely to occur again if a different group of subjects is tested.

It seems clear that one experiment of this type will not give a definitive result, no matter how many subjects are used. However, if the experiment is repeated on other groups (law students, factory workers, office workers, etc.) and the results always come out about the same, then most people would believe that the effect (or lack of an effect) seen is common to all 20-year-olds. In other words, in the absence of truly random samples, convincing evidence of an effect requires it to be demonstrated consistently at different times in different places. This is a nonstatistical type of inference that works equally well with conventional and randomization tests.

Another point is that in many situations either the concept of a population is irrelevant or the data can be considered as representing the whole population. Thus, an example that is considered in the next section concerns the relationship between the world distribution of earwigs and the positions of the continents. There is only one distribution of earwigs that exists, and one set of continents, so the data are not samples from populations except in a most unrealistic sense. Another example in the next section addresses the question of whether there is a cycle in the times of mass extinctions of animals and plants in the geologic past. Here only one extinction record exists, and the concept of this being a random record from a population of possible records is again artificial.

Although random samples are not necessarily required to justify randomization tests, there are times when they do provide the justification. For example, in the reaction time experiment there would be no need to divide the subjects at random into two groups if initially there were two random samples available from the population of 20-year-olds. In that case, either group could be the one given the alcohol, and a valid comparison between the test scores of the two groups to examine the effect of alcohol could be made using a randomization test or a more conventional alternative.

Randomization tests are most easily justified if either the samples being analyzed are random or the experimental design itself justifies randomization testing. This has led some authors (e.g., Kempthorne and Doerfler, 1969) to use the description permutation tests for situations where random samples justify the calculations, and randomization tests for situations where the experimental design provides

the justification. Here both of these descriptions will be used for any situation where randomly reordering observations is used to determine the significance level of a test statistic.

1.2 Examples of Randomization Tests

To clarify the procedures and principles that are used with randomization testing, it will be helpful to consider some detailed examples at this point.

Example 1.1 Mandible Lengths of Male and Female Golden Jackals

The data shown below are mandible lengths in millimeters for male and female golden jackals (*Canis aureus*) for 10 of each sex in the collection in the British Museum of Natural History in London:

Males	120	107	110	116	114	111	113	117	114	112
Females	110	111	107	108	110	105	107	106	111	111

The lengths were measured as part of a study by Higham et al. (1980) on the relationship between prehistoric canid bones from Thailand and similar bones from modern species. For the present example, the question addressed is whether there is any evidence of a difference in the mean lengths for the two sexes.

Data like the above are often collected in the belief that there will be a difference in the results for the two groups. In fact, it is a reasonable supposition that male jackals will tend to be larger than females. The result expected before collecting the data was therefore that the male mean would be higher than the female mean. This can be tested indirectly by setting up a null hypothesis that says that any difference between the two sample means is purely due to chance. If this null hypothesis is consistent with the data, then there is no reason to reject this in favor of the alternative hypothesis — that males have a higher mean.

It may seem strange to test the hypothesis of interest by setting up a null hypothesis and seeing how the data compare with this. However, there is frequently little choice in the matter. It is possible to work out probabilities of different sample results or to generate possible sample results using the null hypothesis. To do this for the hypothesis that is