34. Randomisation

Since introducing statistical hypothesis testing in Chapter 20, this book has steadily introduced statistical tests that can be run for different data types. In all of these statistical tests, the general idea is the same. We calculate some test statistic, then compare this test statistic to a pre-determined null distribution to find the probability (i.e., the p-value) of getting a test statistic as or more extreme than the one we calculated if the null hypothesis were true. This chapter introduces a different approach. Instead of using a pre-determined null distribution, we will build the null distribution using our data. This approach is not one that it is often introduced in introductory statistics texts. It is included here for 3 reasons. First, randomisation presents a different way of thinking statistically without introducing an entirely different philosophical or methodological approach such as likelihood (Edwards, 1972) or Bayesian statistics (Lee, 1997). Second, it helps reinforce the concept of what null hypothesis testing is and what p-values are. Third, it introduces one of many motivations for learning a bit of coding in R (see Chapter 35). Before explaining the randomisation approach, it is useful to summarise the parametric hypothesis tests of introduced in earlier chapters.

34.1. Summary of parametric hypothesis testing

For the parameteric tests introduced in previous chapters, null distributions included the t-distribution, F distribution, and χ^2 distribution. For the t-tests in Chapter 21, the test statistic was the t-statistic, which we compared to a t-distribution. The one-sample t-test compared the mean of some variable ($\bar{x}$) to a specific number (μ_0), the independent samples t-test compared 2 group means, and the paired sample t-test compared the mean difference between 2 paired groups. All of these tests used the t-distribution and calculated some value of t. Very low or high values of t at the extreme ends of the t-distribution are unlikely if the null hypothesis is true, so, in a two-tailed test, these are associated with low p-values that lead us to reject the null hypothesis.

For the analysis of variance (ANOVA) tests of Chapter 23, the relevant test statistic was the F statistic, with the F-distribution being the null distribution expected if 2 variances are equal. The one-way ANOVA used the within and among group variances of 2 or more groups to test the null hypothesis that all group means are equal. The two-way ANOVA of Chapter 26 extended the framework of the one-way ANOVA, allowing for a second variable of groups. This made it possible to simultaneously test whether or not the means of 2 different group types were the same, and whether or not there was

an interaction between group types. All of these ANOVA tests calculated some value of F and compared it to the F distribution with an appropriate degrees of freedom. Sufficiently high F values were associated with a low p-value and therefore the rejection of the null hypothesis.

The Chi-square tests introduced in Chapter 28 were used to test the frequencies of categorical observations and determine if they matched some expected frequencies (Chi-square goodness of fit test) or were associated in some way with the frequencies of another variable (Chi-square test of association test). In these tests, the χ^2 statistic was used and compared to a null χ^2 distribution with an appropriate degrees of freedom. High χ^2 values were associated with low p-values and the rejection of the null hypothesis.

For testing the significance of correlation coefficients (see Chapter 29) and linear regression coefficients (see Chapter 31), a t-distribution was used. And an F distribution was used to test for the overall significance of linear regression models.

For these tests, the approach to hypothesis testing was therefore always to use the t-distribution, F distribution, or χ^2 distribution in some way. These distributions are more formally defined in mathematical statistics (Miller and Miller, 2004), a field of study that uses mathematics to derive the probability distributions that arise from an outcome of random events (e.g., the coin-flipping of Chapter 14.1). The reason that we use these distributions in statistical hypothesis testing is that they are often quite good at describing the outcomes that we expect when we collect a sample from a population. But this is not always the case. Recall that sometimes the assumptions of a particular statistical test were not met. In this case, a non-parametric alternative was introduced. The non-parameteric test used the ranks of data instead of the actual values (e.g., the Wilcoxon, Mann-Whitney U, Kruskall-Wallis H, and Spearman rank correlation coefficient tests). Randomisation uses a different approach.

34.2. Randomisation approach

Randomisation takes a different approach to null hypothesis testing. Instead of assuming a theoretical null distribution against which we compare our test statistic, we ask, 'if the ordering of the data we collected was actually random, then what is the probability of getting a test statistic as or more extreme than the one that we actually did'. Rather than using a null distribution derived from mathematical statistics, we will build the null distribution by randomising our data in some useful way (Manly, 2007). Conceptually, this is often easier to understand because randomisation approaches make it easier to see why the null distribution exists and what it is doing. Unfortunately, these methods are more challenging to implement in practice because using them requires knowing a bit of coding. The best way to get started is with an instructive example.

34.3. Randomisation for hypothesis testing

As in several previous chapters, the data set used here is inspired by the many species of wasps that lay their eggs in the flowers of the Sonoran Desert rock fig (*Ficus petiolaris*). This tree is distributed throughout the Baja peninsula, and in parts of mainland Mexico. Fig trees and the wasps that develop inside of them have a fascinating ecology, but for now we will just focus on the morphologies of two closely related species as an example. The fig wasp below are two unnamed species of the genus *Idarnes*, which can refer to simply as 'Short-ovipositor 1' (SO1) and 'Short-ovipositor 2' (SO2).

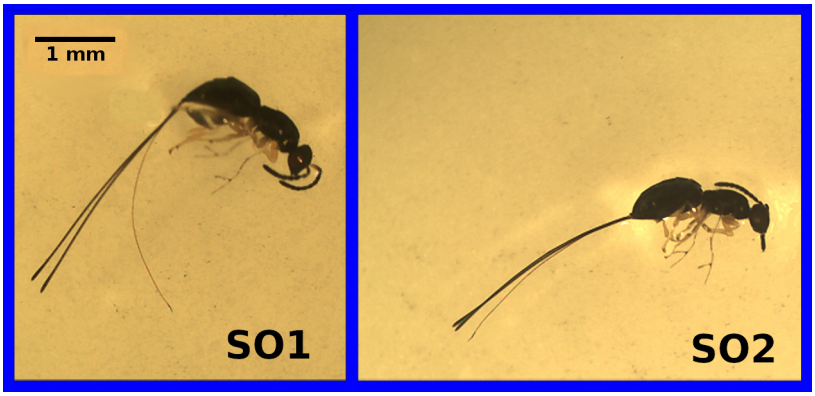

Figure 34.1.: Image of two fig wasp species, roughly 3 mm in length, labelled 'SO1' and 'SO2'

The reason that these two species are called 'SO1' and 'SO2' is that there is actually another species that lays its eggs in *F. petiolaris* flowers, one with an ovipositor that is at least twice as long as the ones above.

Suppose that we have some data on the lengths of the ovipositors from each species. We might want to know whether the mean ovipositor length differs between the 2 species. Below shows histograms of ovipositor lengths collected from 32 fig wasps, 17 of the species 'SO1', and 15 of the species 'SO2'.

To test whether or not mean ovipositor length is different between these 2 fig wasps, our standard approach would be to use an independent samples t-test (see Chapter 21.2). The null hypothesis would be that the 2 means are the same, and the alternative (two-sided) hypothesis would be that the 2 means are not the same. We would need to check the assumption that the data are normally distributed, and that both samples have similar variances. Assuming that the assumption of normality is not violated (in which case we would need to consider a Mann Whitney test), and that both groups had similar variances (if not, we would use the Welch's t-test), we could proceed with calculating our t-statistic for equal sample size,

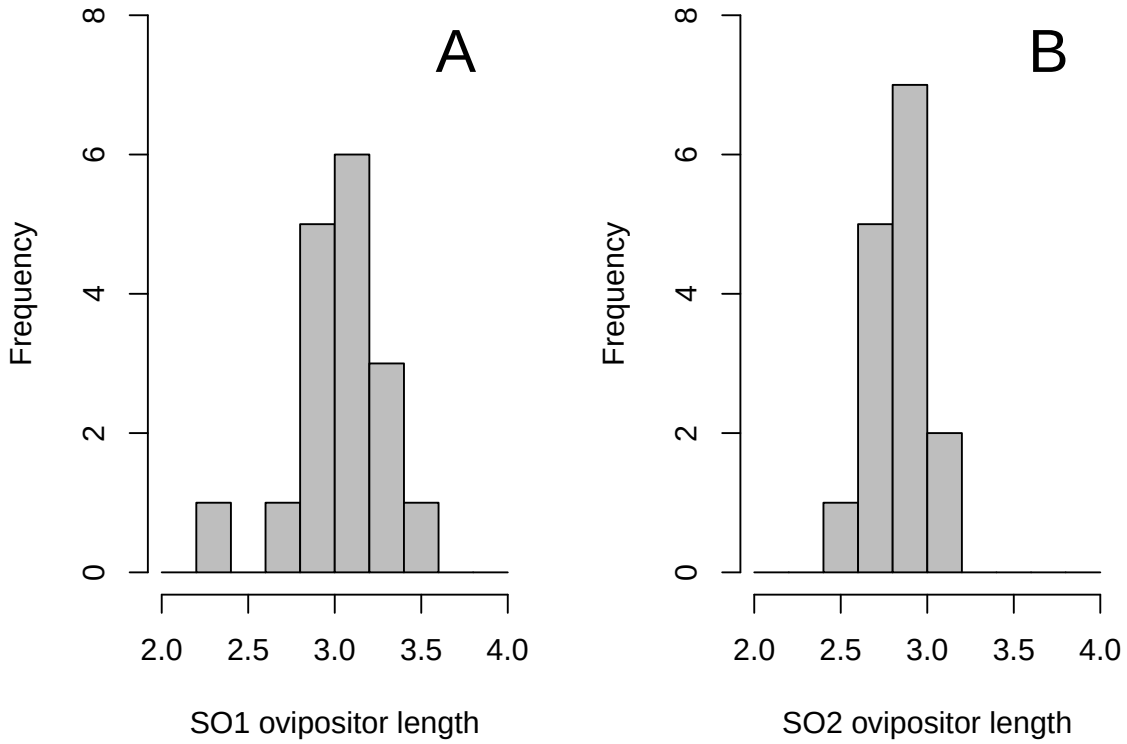

Figure 34.2.: Ovipositor length distributions for two unnamed species of fig wasps SO1 (A) and SO2 (B) collected from Baja, Mexico.

$$t_{\bar{y}_{SO1}-\bar{y}_{SO2}} = \frac{\bar{y}_{SO1} - \bar{y}_{SO2}}{s_p}.$$

The s_p is just being used as a short-hand to indicate the pooled standard deviation. For the 2 species of fig wasps, $\bar{y}_{SO1} = 3.0301176$, $\bar{y}_{SO2} = 2.8448667$, and $s_p = 0.0765801$. We can therefore calculate $t_{\bar{y}_{SO1}-\bar{y}_{SO2}}$,

$$t_{\bar{y}_{SO1}-\bar{y}_{SO2}} = \frac{3.03 - 2.845}{0.077}.$$

After we calculate our t-statistic as $t_{\bar{y}_{SO1}-\bar{y}_{SO2}} = 2.419$, we would use the t-distribution to find the p-value (or, rather, get Jamovi to do this for us). Figure 32.3 shows the t-distribution for 30 degrees of freedom with an arrow pointing at the value of $t_{\bar{y}_{SO1}-\bar{y}_{SO2}} = 2.419$.

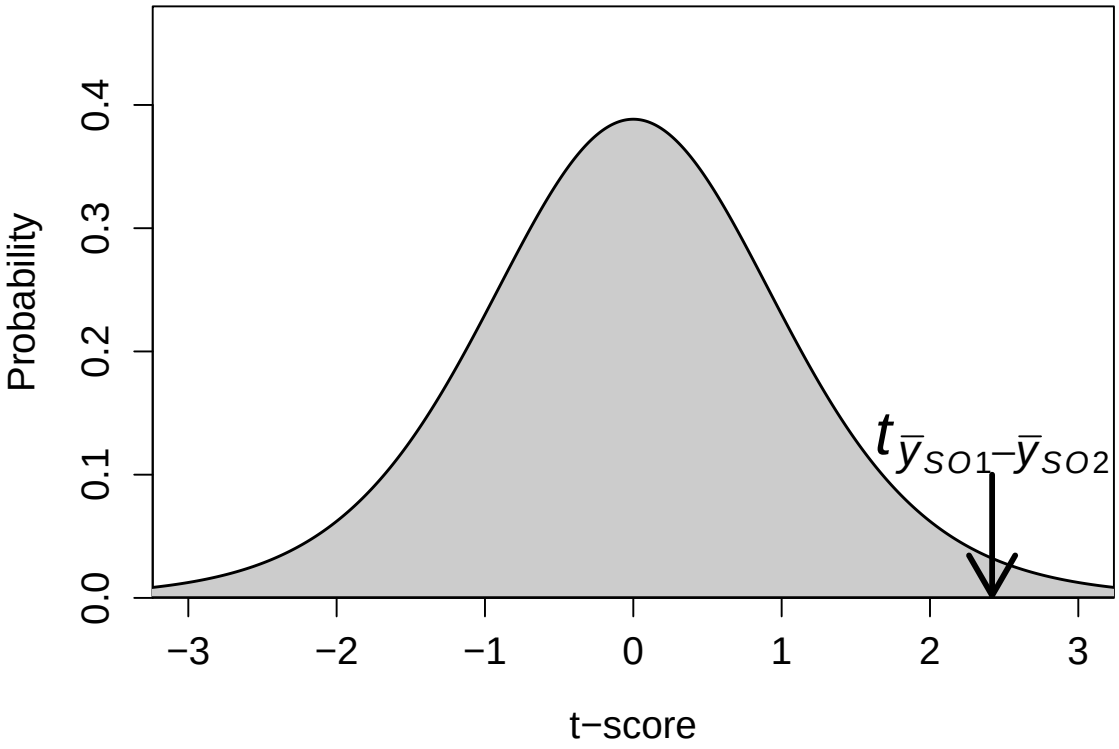

Figure 34.3.: A t-distribution is shown with a calculated t-statistic of 2.419 indicated with a downward arrow.

If the null hypothesis is true, and $\bar{y}_{SO1}$ and $\bar{y}_{SO2}$ are actually sampled from a population with the same mean, then the probability of randomly sampling a value more extreme than 2.419 (i.e., greater than 2.419 or less than -2.419) is $P = 0.0218352$. We therefore reject the null hypothesis because $P < 0.05$, and we conclude that the mean ovipositor lengths of each species are not the same.

Randomisation takes a different approach to the same problem. Instead of using the t-distribution in Figure 34.3, we will build our own null distribution using the fig wasp ovipositor length data. We do it by randomising group identity in the dataset. The logic is that if there really is no difference between group means, then we should be able to randomly shuffle group identities (species) and get a difference between means that is not far off the one we actually get from the data. In other words, what would the difference between group means be if we just mixed up all of the species (so some SO1s become SO2s, some SO2s become SO1s, some stay the same), then calcualted the difference betwen means of the mixed up groups? If we just do this once, then we cannot learn much. But if we randomly shuffle the groups many, many times (say at least 9999), then we could see what the difference between group means would look like just by chance; that is, if ovipositor length really was not different between SO1 and SO2. We could then compare our actual difference between mean ovipositor lengths to this null distribution, in which the difference between groups means really is random (it has to be, we randomised the groups ourselves!).

The idea is easiest to see using an interactive application.

> Click here for an interactive application showing the process of a randomisation test that provides an equivalent test to an independent samples t-test.

The interactive application builds a null distribution of differences between the mean of SO1 and SO2 (click the 'Randomise' button to add a new random difference to the distribution). This null distribution can be compared with the observed difference of $\bar{y}_{SO1} - \bar{y}_{SO2} = 0.185$.

With modern computing power, we do not need to do this randomisation manually. A desktop computer can easily reshuffle the species identities and calculate a difference between means thousands of times in less than a second. The histogram below shows the distribution of the difference between species mean ovipositor length if we were to randomly reshuffle groups 99999 times.

Given this distribution random mean differences between species ovipositor lengths, the observed difference of 0.185 appears to be fairly extreme. We can quantify how extreme by figuring out the proportion of mean differences in the above histogram that are more extreme than our observed difference (i.e., greater than 0.185 or less than -0.185). It turns out that only 2178 out of 9.9999×10^4 random mean differences between ovipositor lengths were as or more extreme than our observed difference of 0.185. To express this as a probability, we can simply take the number of differences as or more extreme than our observed difference (including the observed one itself), divided by the total number of differences (again, including the observed one),

$$P = \frac{2178 + 1}{99999 + 1}.$$

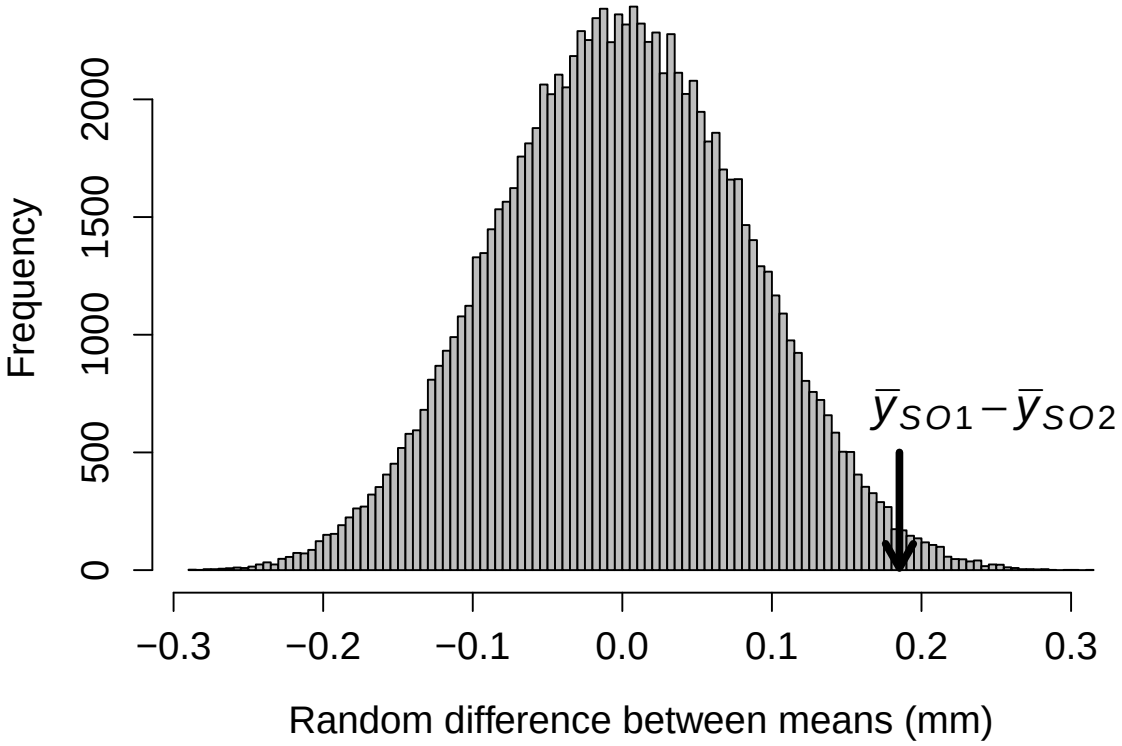

Figure 34.4.: A distribution of the difference between mean species ovipositor lengths in 2 different species of fig wasps when species identity is randomly shuffled. The true difference between sampled mean SO1 and SO2 ovipositor lengths is pointed out with the black arrow at 0.185. Ovipositor lengths were collected from wasps in Baja, Mexico in 2010.

When we calculate the above, we get a value of 0.02179. Notice that the calculated value is assigned with a 'P'. This is because the value **is a p-value** (technically, an unbiased estimate of a p-value). Consider how close it is to the value of $P = 0.0218352$ that we got from the traditional t-test. Conceptually, we are doing the same thing in both cases; we are comparing a test statistic to a null distribution to see how extreme the differences between groups really are.

34.4. Randomisation assumptions

Recall from Chapter 21.4 the assumptions underlying t-tests, which include (1) data are continuous, (2) sample observations are a random sample from the population, and (3) sample means are normally distributed around the true mean. We do not need assume 2 or 3. The randomisation approach is still valid even if the data are not normally distributed. Samples can have different variables. The observations do not even need to be independent. The validity of the randomisation approach even when standard assumptions do not apply can be quite useful.

The downside of the randomisation approach is that the statistical inferences that we make are limited to our sample, not the broader population (recall the difference between a sample and population from Chapter 4). Because the randomisation method does not assume that the data are a random sample from the population of interest (as is the case for the traditional t-test), we cannot formally make an inference about the difference between populations from which the sample was made. This is not necessarily a problem in practice. It is only relevant in terms of the formal assumptions of the model. Once we run our randomisation test, it might be entirely reasonable to argue verbally that the results of our randomisation test can generalise to our population of interest In other words, we can argue that the difference between groups in the sample reflects a difference in the populations from which the sample came (Ludbrook and Dudley, 1998; Ernst, 2004).

34.5. Bootstrapping

We also can use randomisation to calculate confidence intervals for a variable of interest. Remember from Chapter 17 the traditional way to calculate upper and lower confidence intervals using a normal distribution,

$$LCI = \bar{x} - (z \times SE),$$

$$UCI = \bar{x} + (z \times SE).$$

Recall from Chapter 18 that a t-score is substituted for a z-score to account for a finite population size. For review, intervals can be worked out using the interactive application for the t-distribution.

Suppose we want to calculate 95 per cent confidence intervals the the ovipositor lengths of SO1 (Figure 34.2A). There are 17 ovipositor lengths for SO1.

3.256, 3.133, 3.071, 2.299, 2.995, 2.929, 3.291, 2.658, 3.406, 2.976, 2.817, 3.133, 3, 3.027, 3.178, 3.133, 3.21

To get the 95 per cent confidence interval using the method from Chapter 17, we would calculate the mean $\bar{x}_{SO1} = 3.03$, standard deviation $s_{SO1} = 0.26$, and the t-score for $df = N - 1$ (where N = 17), which is $t = 2.120$. Keeping in mind the formula for standard error $s/\sqrt{N}$, we can calculate,

$$LCI = 3.03 - \left(2.120 \times \frac{0.26}{\sqrt{17}}\right),$$

$$UCI = 3.03 + \left(2.120 \times \frac{0.26}{\sqrt{17}}\right).$$

Calculating the above gives us values of LCI = 2.896 and UCI = 3.164.

Again, randomisation uses a different approach to get an estimate of the same confidence intervals. Instead of calculating the standard error and multiplying it by a z score or t score to encompass a particular interval of probability density, we can instead resample the data we have with replacement many times, calculating the mean each time we resample. The general idea is that this process of resampling approximates what would happen if we were to go back and resample new data from our original population many times, thereby giving us the distribution of means from all of these hypothetical resampling events (Manly, 2007). To calculate our 95 per cent confidence intervals, we then only need to rank the calculated means and find the mean closest to the lowest 2.5 per cent and the highest 97.5 per cent.

Remember from Chapter 14.3 that phrase 'resampling with replacement' just means that we are going to randomly sampled some values, but not remove them from the pool of possible values after they are sampled. If we resample the numbers above with replacement, we might therefore sample some values 2 or more times by chance, and other values might not be sampled at all. The numbers below resample from the above with replacement.

2.299, 2.299, 2.976, 3.406, 2.976, 3.406, 3.071, 3.291, 3.133, 3.406, 3.133, 3.133, 2.658, 2.658, 3.178, 3.133, 3.21

Notice that some values appear twice in the data above, while other values that were present in the original data set are no longer present after resampling with replacement. Consequently, this new resampled data has a different mean than the original. The

mean of the original 17 SO1 ovipositor length values was 3.03, and the mean of the values resampled above is 3.022. We can resample another set of numbers to get a new mean.

3.178, 3.291, 3.291, 3.071, 2.929, 2.299, 3.178, 3, 3.071, 3, 3.133, 2.976, 3.291, 3, 3.027, 3.21, 3.406

The mean of the above sample is 3.079. We can continue doing this process until we have a high number of random samples and means. Figure 34.5 shows the distribution of means if we repeat this process of resampling with replacement and calculating the mean 10000 times.

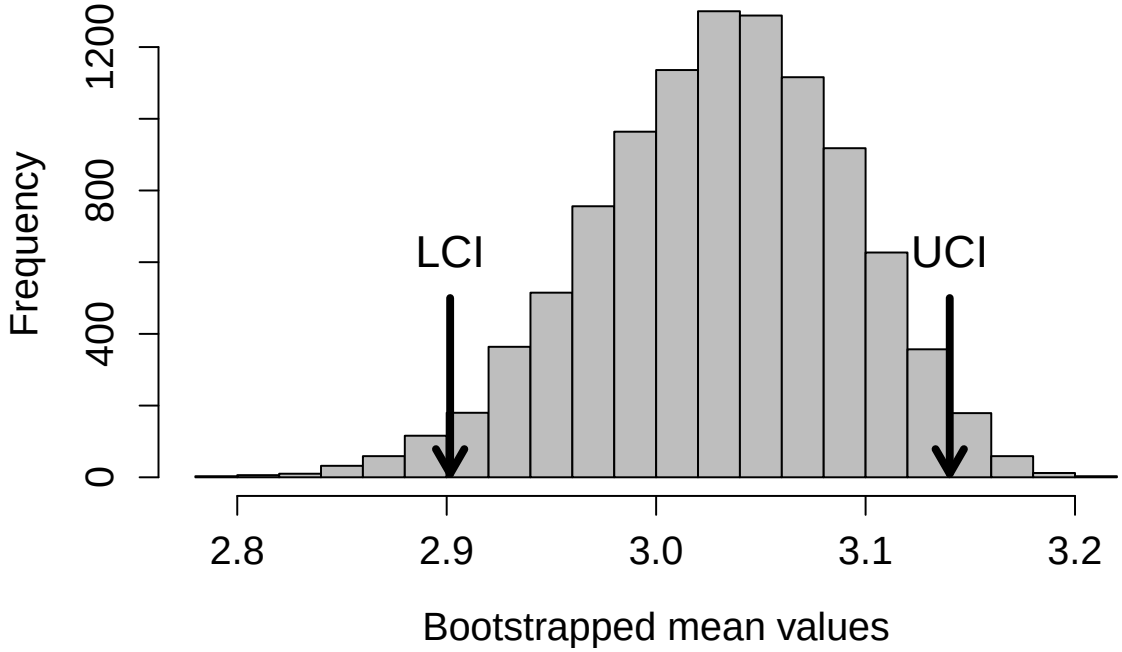

Figure 34.5.: A distribution of bootstrapped values of ovipositor lengths from 17 measurements of a fig wasp species collected from Baja, Mexico. A total of 10000 bootstrapped means are shown, and arrows indicate the location of the 2.5 per cent (LCI) and 97.5 per cent (UCI) ranked boostrapped mean values.

The arrows show the locations of the 2.5 per cent and 97.5 per cent ranked values of the bootstrapped means. Note that it does not matter if the above distribution is not normal (it appears a bit skewed). The boostrap still works. The values using the randomisation approach are LCI = 2.902 and UCI = 3.14. These values are quite similar to those calculated with the traditional approach because we are doing the same thing, conceptually. But instead of finding confidence intervals of the sample means around the true mean using the t-distribution, we are actually simulating the process of resampling from the population and calculating sample means many times.

34.6. Monte Carlo

The previous examples of hypothesis testing and bootstrapping involved resampling from existing data sets to address a statistical question. But we can also use randomisation in cases in which it is impossible to derive the null distribution from the data. In this last example, the goal will be to test whether or not fig trees (Figure 34.5) are randomly distributed across a fixed sampling area.

Figure 34.6.: A tree of the Sonoran Desert Rock Fig from Baja, Mexico.

Locations of these fig trees were collected over the course of many field seasons. In Baja, Mexico, a dataset of 59 trees was collected. The first 6 rows of this dataset with tree locations are shown below.

Table 34.1.: Latitudes, longitudes, and elevations of Sonoran Desert Rock Fig trees collected from a sample site in Baja, Mexico.

Site	Tree	Latitude	Longitude	Elevation
S172	T34	28.29021	-113.1116	718.2859
S172	T01	28.29141	-113.1117	664.8726
S172	T02	28.29130	-113.1118	652.8560
S172	T03	28.29129	-113.1126	663.6709
S172	T04	28.29127	-113.1127	653.3367
S172	T05A	28.29110	-113.1125	676.8889

We can plot the latitude and longitude of each of the 59 trees below. Figure the study plot at the field site, randomly set from latitude 28.289 to 28.2918 and longitude -113.1145 to -113.1095 (Figure 34.7).

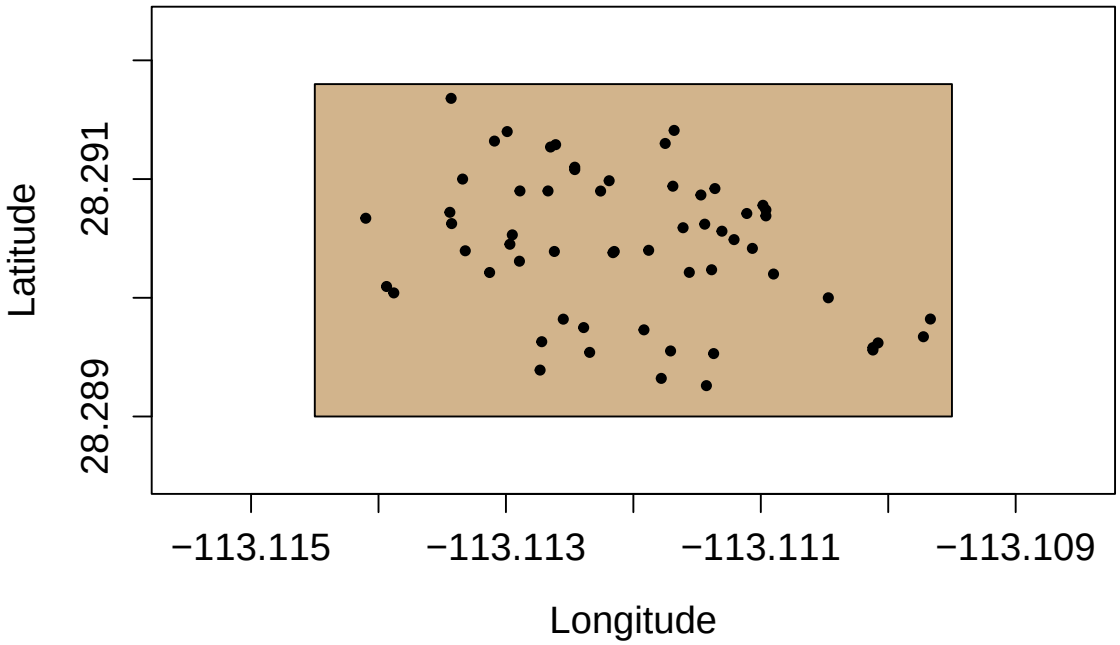

Figure 34.7.: Latitude and longitude locations of Sonoran Desert Rock Fig trees from Baja, Mexico.

The focal question is whether or not these trees are randomly distributed within the brown box. Alternatively, the trees might be more clustered together than we would expect by chance. This is not a question that can be answered by randomising the latitude and longitude coordinates of trees. What we need to do instead is compare the distribution of the trees above with that of trees with randomly placed latitude and longitude coordinates within the box. This is a job for a Monte Carlo test, which compares an observed test statistic with that derived from a theoretical null model (Manly, 2007). In this case, the test statistic we will use is distance to the nearest neighbour (i.e., for a focal tree, how close the nearest member of the same species). The null model that we will assume is a set of randomly sampled latitude and longitude coordinates within the fixed study area.

More formally, our null hypothesis will be that the mean nearest neighour distance for a focal tree in the observed data will not differ significantly from the mean nearest neighour distance obtained form the same number of trees randomly distributed within the sampling area (brown box). To test this null hypothesis, we can randomly place 59 trees within the sampling area and calculate the mean nearest neighbour distance for the randomly placed trees. If we repeat this procedure a large number of times, then we can build a null distribution of nearest neighbour distances for trees that are randomly

placed within the study are (i.e., random latitude and longitude coordinates).

Given the random placement of these trees, find the mean distance to the nearest neighbouring tree, calculated across all of the randomly placed trees. This mean nearest neighbour distance is then stored so that a null distribution can be built. We can see what these randomly placed trees look like by plotting some of the iterations in Figure 34.8.

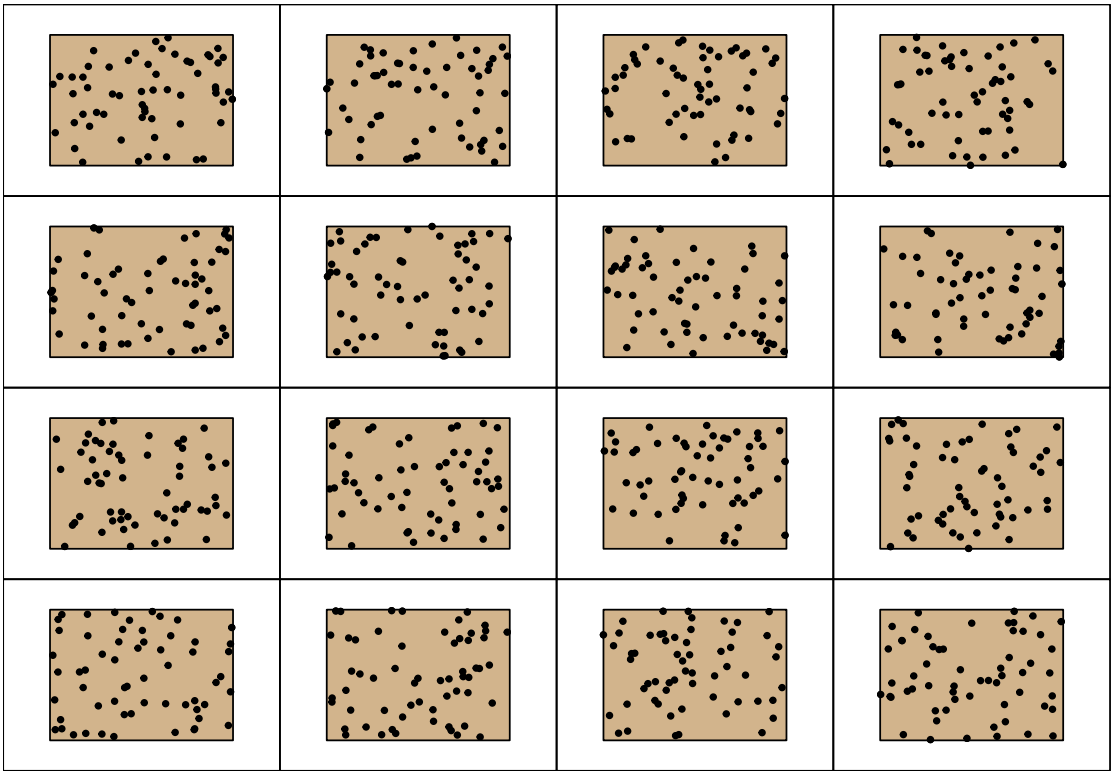

Figure 34.8.: Sixteen separate random allocations of latitude and longitude locations within a sampling Sonoran Desert Rock Fig trees from Baja, Mexico. There are 59 total locations randomly allocated for each panel, each representing an individual fig tree.

In Figure 34.9, the distribution of mean distance to the nearest neighbour is plotted for the 9999 randomly generated tree study areas. The arrow shows the actual observed mean distance between nearest neighbours, as calculated form the original data set.

It appears, from the position of the mean tree nearest neighbour distance in the observed data, that the *F. petiolaris* trees are no more or less spatially aggregated than would be expected by chance.

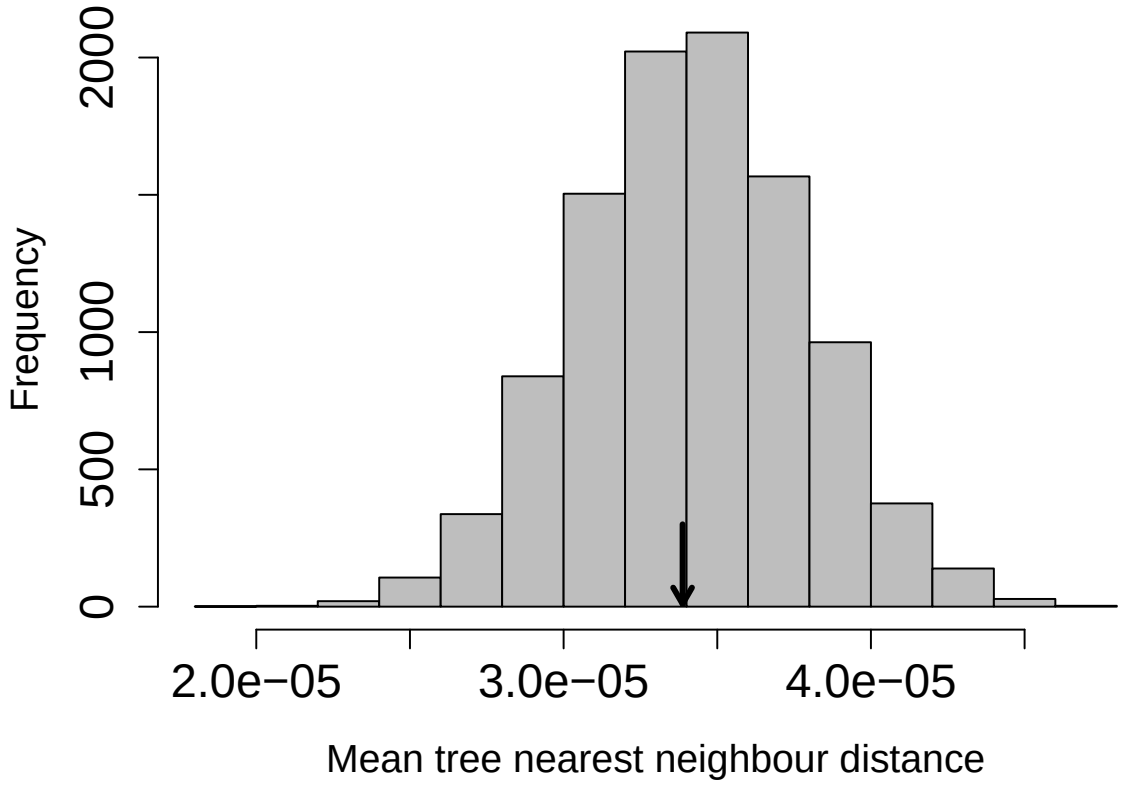

Figure 34.9.: The distribution of mean neareast neighbour distance from 9999 random simulations of 59 latitude and longitidue locations within a field site in Baja, Mexico. The black arrow indicates mean nearest neighbour distance for 59 observed trees from the field site.

34.7. Randomisation conclusions

The examples demonstrated in this chapter are just a small set of what is possible using randomisation approaches. Randomisation, bootstrapping, and Monte Carlo tests are highly flexible tools that can be used in a variety of ways (Manly, 2007). They could be used as substitutes (or complements) to any of the hypothesis tests that we have used in this book. For example, to use a randomisation approach for testing whether or not a correlation coefficient is significantly different from 0, we could randomly shuffle the values of one variable 9999 times. After each shuffle, we would calculate the Pearson product moment correlation coefficient, then used the 9999 randomised correlations as a null distribution to compare against the observed correlation coefficient. This would give us a plot like the one shown in Figure 34.4, but showing the null distribution of the correlation coefficient instead of the difference between group means. Similar approaches could be used for ANOVA or linear regression (Manly, 2007).