Workshop 9: Bootstrap resampling methods

This workshop will illustrate some of the principles of data analysis using resampling.

Methods summary

Here is a quick summary of the two methods, the bootstrap and the permutation test, applied in today's workshop.

Bootstrap estimation

The bootstrap is is a computer-intensive method used to calculate standard errors for an estimate and to obtain approximate confidence intervals for a population parameter. It works well in most situations except when sample size is small. The basic procedure is:

1. Take a random sample of the data, sampling with replacement.
2. Compute the statistic on the bootstrap replicate obtained in (1).
3. Repeat steps (1) and (2) many times (e.g., $B = 10^4$), saving the result each time.
4. Plot the B bootstrap replicate estimates to visualize the shape of the sampling distribution.
5. Calculate the standard deviation of the B boostrap replicate estimates to obtain the standard error of the estimate.
6. It is also possible to use the bootstrap replicates to obtain a confidence interval for the parameter using a percentile method and the $\mathrm{BC_a}$ method.

Permutation test

The permutation test is a computer-intensive method for hypothesis testing. It is used to generate a null distribution for a test statistic measuring association between two variables (or differences between groups). When used to test differences between two or more groups in a numerical variable the method assumes that the distributions have equal shape. However, the method is fairly robust to departures from this assumption. The steps are:

1. Create a randomized data set in which the values for one of two variables are randomly reassigned to subjects (without replacement).

2. Calculate the test statistic measuring association between the variables (or difference between groups) on the randomized sample obtained in (1).

3. Repeat steps (1) and (2) many times ($B = 10^4$), saving the result each time.

4. Plot the B values of the test statistic generated in (3) to visualize the shape of the null distribution.

5. Using the distribution of the test statistic from (3), calculate the tail probability for the observed value of the test statistic (calculated on the original data). Multiply this number by 2 if test is two-sided.

Caribbean bird immigration

Birds of the Caribbean islands of the Lesser Antilles are descended from rare immigrants from larger islands and the nearby mainland. The data here are the approximate dates of immigration, in millions of years, of each of 37 bird species now present on the Lesser Antilles (Ricklefs and Bermingham 2001, *Science* 294: 1522-1524). The dates were calculated from the difference in mitochondrial DNA sequences between each of the species and its closest living relative on larger islands or the mainland. The data can be downloaded here.

1. What shape is the frequency distribution of estimated immigration dates? Use a graph to display it.

2. What are the mean and median dates of immigration, in millions of years? Why are the two values so different?

3. Obtain a single bootstrap replicate of the immigration dates and plot the results. How different is the frequency distribution from that of the data?

4. Write a short loop or use an apply statement to generate 10^4 bootstrap replicate estimates for the sample median immigration date. It is a good idea to begin by using only 10 replicates until your code is tested and found to be working smoothly. Store the resulting medians in a vector.

5. Plot the frequency distribution of your results from (4). What does this frequency distribution estimate?

6. Using your results in (4) to calculate a standard error for the sample median.

7. Most of the familiar estimators of population parameters, such as sample mean and variance, are unbiased, which means that the mean of its sampling distribution equals the parameter value being estimated. For example, the mean of the sampling distribution of the sample mean is μ, the very parameter being estimated. The sample mean is an unbiased estimator. However, some estimators are biased, and the bootstrap is often used to estimate bias. Is the median of immigration dates biased? Calculate the mean of the bootstrap replicate estimates of the median immigration date to estimate the bias.

8. Use the percentile method (check the `quantile()` function) to generate an approximate 95% confidence for the median.

9. Use the `boot` library to generate a more accurate boostrap confidence interval for the median using the BC_a method.

Trillium fragmentation

Logging in western North America impacts populations of western trillium (*Trillium ovatum*), a long-lived perennial inhabiting conifer forests. Jules and Rathcke (1999, *Conservation Biology* 13:784-793) measured attributes of eight local populations of western trillium confined to forest patches of varying size created by logging in southwestern Oregon. A subset of their data are here. The variables included are

- population
- forest fragment size (ha)
- distance between local population and forest edge (m)
- years since isolated
- number of plants in local population
- proportion of plants consumed by deer in of the years of study
- recruitment rate, the density of new plants produced in each population per year

1. Plot recruitment against fragment size.

2. From visual inspection of your plot in (1), choose an appropriate transformation of one or both variables to meet (roughly) the assumptions of linear regression.

3. Using a linear model on your transformed data, estimate the slope of the linear regression of recruitment on fragment size. Inspect the diagnostic plots. These will reveal that although transformation improved matters, there might still be some issues regarding the assumptions of equal variance of residuals. A small sample size makes it difficult to pursue the issue much further. Let's take a conservative approach for the purposes of this exercise and use a nonparametric approach, the permutation test, to test the null hypothesis of zero slope. We'll use the transformed data because the transformation rendered the relationship more linear (it is the slope of a line that we wish to test) and it eliminated the problem of one or two data points having excessive influence.

4. Using the permutation test, create a null distribution for the slope of the linear regression.

5. Plot the null distribution. Compare the distribution to the value of the slope you estimated in (3). Is your slope near the middle or toward one of the tails of the null distribution?

6. Calculate the tail probability using the results from your analysis of the data in (3) and the null distribution.

7. Use your results in (6) to calculate an approximate P-value for the test of the null hypothesis of zero slope.

Note: As we've discussed in class, the analysis of data shouldn't end with a P-value because it tells us nothing about magnitudes of effects. A drawback of the permutation test is that it is entirely focused on the P-value and provides no estimates of parameters. The bootstrap is more informative, but sample size here is probably too small for a reliable outcome.

Vampire attack

The vampire bat, *Desmodus rotundus*, commonly feeds on the blood of domestic cattle. It prefers cows to bulls, which suggested that the bats might be responding to a hormonal cue. To explore this, the frequency of vampire bat attacks on cows in estrous was compared with attacks on anestrous female cows. The following data are from one night of observation (D. C. Turner 1975, The vampire bat: A field study in behavior and ecology. Johns Hopkins Press, Baltimore).

	Female cows in estrous	Anestrous female cows
Bitten by vampire bat	15	6
Not bitten	7	322
Total	22	328

1. To begin the analysis, create two variables (estrous and bitten) each with two states (yes/no) representing the above measurements for the 350 cows. 15 of the cows should have "yes" for both variables, 7 should have "yes" for estrous and "no" for bitten, and so on. You may find the`rep` command useful here.

2. Create a 2×2 contingency table to verify the results of (1).

3. Use a conventional χ^2 contingency test to test the null hypothesis of no difference in the frequency of cows bitten by vampire bats between estrous and anestrous cows. Use the `chisq.test` command in R (type `?chisq.test` for help). Store the results in an object (e.g., `out.chi`). If you execute the command correctly you should receive a warning. What is the source of the problem? To determine this, examine the expected frequencies (they are stored in the results object, e.g.,that you named `out.chi`).

4. Use the permutation test instead to test the null hypothesis. HINT: the `chisq.test` can do this for you, if you specify the options correctly.

5. Calculate the odds ratio using the variables you created in (1). The odds ratio is a commonly used measure of association in 2×2 tables. It is the ratio of the odds of success in a treatment group compared to the odds of success in a control group. Here, the ratio will be odds of estrous cows being bitten / odds of anestrous cows being bitten. The quickest way to calculate odds ratio from a table of frequencies is

$$OR = \frac{a/c}{b/d}$$

where

	treatment	control
success	a	b
failure	c	d

A problem arises if one of the four table frequencies is 0, which can create an odds ratio of 0 or ∞. Although not optimal, a standard fix is to add 0.5 to each cell of the table before calculating OR.

6. Use the bootstrap to obtain a standard error for the estimate of odds ratio. By chance your resampling might occasionally produce a zero for the number of anestrous cows bitten, which will result in a division by zero and a value of `Inf` for the odds ratio or relative risk. One solution is to add 0.5 to each cell of the 2×2 table before calculating the odds ratio. Otherwise, just remove the `Inf` values from the vector of results using

```
x <- x[is.finite(x)]
```

7. Use the percentile method to obtain a 95% confidence interval for the odds ratio.