"Rafflesia" – stratified sampling

For this exercise you will need the R script file "Rafflesia_stratified_sampling.R".

Many wildlife studies involve spatial sampling: we select a number of plots or transect lines or trap locations or other kinds of sites, and use the data from this sample to make inferences about the whole study area.

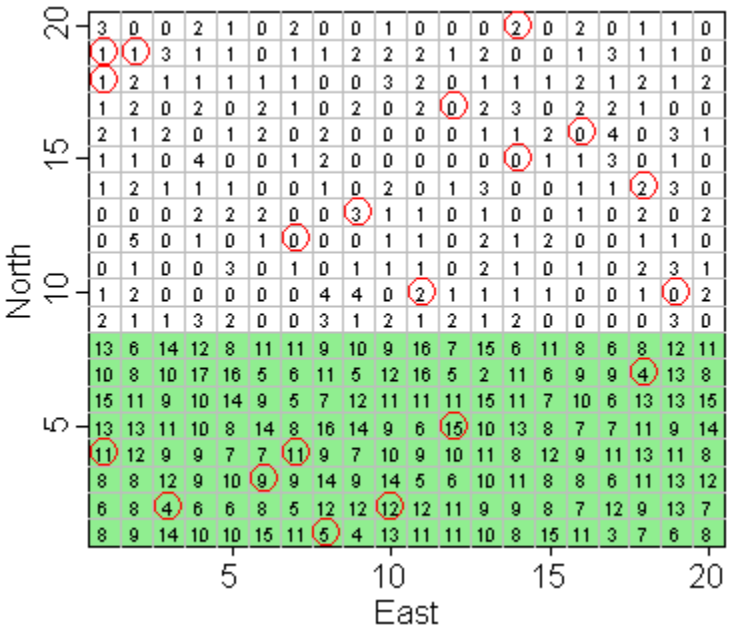
Estimating abundance

In this example we consider a study to find the total population of *Rafflesia* flowers in the study area. The area is divided into two clearly differentiated habitats (eg. dipterocarp forest vs. peat swamp forest, or river plains vs. steep hill sides). We'll experiment with using a stratified survey design and compare the results with an unstratified design.

This is done in R, and the commands are in the R script file "Rafflesia_stratified_sampling.R".

Setting up the population

For this exercise, we'll use a square study area, divided into 400 plots in a 20 x 20 grid. Any other rectangular arrangement would be just as good; an irregular arrangement – corresponding to an actual national park, say – is also possible but a little more complicated to set up. Also, the locations could be transects



The southern section of the study area, comprising 160 plots, is better habitat than the rest, with a higher density of *Rafflesia* flowers, approximately 10 per plot vs. 1 per plot in the poorer habitat on average.

We use the `rpois` function in R to generate true numbers for each of the plots. The example in the diagram on the right has τ = 1805 flowers in all, 1568 in the high density area, 237 in the low density area. (We talk about 'plots' here, but the locations could be transects, with the number being the number of flowers or plants or animals encountered on the transect.)

Unstratified sampling design

We use the sample function to select random samples; the diagram shows a sample of 20 plots taken at random from the whole area, ie. without stratification. The sample mean is 4.15 flowers per plot, so we would estimate the population as $\hat{\tau}$ = 4.15 x 400 = 1660.

We can draw thousands of samples like this from the population and see if the mean of the estimates is close to the true value (ie, if the method is unbiased). We can check the spread of the estimates with histograms and beeswarm plots, and calculate the root mean square error (RMSE), for comparison with other sampling designs.

Stratified designs

Here we select separate samples from the high-density and low-density areas, estimate populations separately, then combine them into a global total. For example, we might try these designs:

A. sub-sample size proportional to area: 8 in smaller stratum, 12 in larger stratum.

B. similar sub-sample sizes, but 13 in smaller, high-density stratum, 7 in the low-density stratum.

C. bigger difference in sub-sample sizes: 17 in smaller stratum, 3 in the larger stratum.

All three designs require the same effort in the field as the unstratified design above, they all involve surveys of 20 plots.

We generate thousands of samples for each design, check for bias, do the histograms and beeswarm plots and calculate RMSE.

Run the R script yourself and see which design gives the most accurate estimates.

Incorrect analysis

It's also interesting to see what happens if we do an incorrect analysis: using a stratified sample design such as B. above, but treat it as a simple random sample for analysis.