Developing your sampling design

In this activity you will develop a sampling design to address a specific hypothesis. You will be using the data preliminary data you collected at Younger Lagoon to determine the sampling scheme. There are two products that you need to develop.

1) The relationship between level of replication and the estimation of the running mean and running standard deviation (an estimate of the variability in your data).
2) The relationship between statistical power (the ability to statistically detect a true difference) and.
   a. Critical p-value
   b. Replication
   c. Effect size

First you should work on #1. Create a data files called Younger Lagoon data (your group name) – an excel file. You should have four columns: Group, Target Variable (e.g. species name, measured attribute, etc), Replicate and Dependent Variable (e.g percent cover, number size, etc). This is the way data fields are usually set up for analyses. Note, there are no empty cells unless the data are truly missing. Empty cells should never be used for zeros or to indicate that the cell is the same as the entry above it.

Note that in the text below the brackets [ ] are there to indicate something should be done and should not be entered as part of formulas etc.

1) First create a new variable in Column E – call it Run_Mean (running mean) in cell E1. Now enter the following in cell E2, [=AVERAGE(D$2:D2)]. This will calculate the mean of values in column D though line 2. Drag this formula to the end of the data range. Notice that the first value D$2 stays locked (this is the purpose of the $), while the second value floats. This creates a running mean.
2) Create a new variable in Column F – call it STD_DEV (standard deviation) in F1. Now enter the following formula in F3 not F2 [=STDEV(D$2:D3)]. Drag this to the bottom of the data set. This creates a running standard deviation. The reason that we could not use cell F2 is that calculation of Standard Deviations needs more than one value.
3) Save your file
4) Now plot the relationship between replication and Run_Mean and STD_DEV. Replication should be on the x axis and the others on the Y axis. You should do this in Excel or export it to JMP and plot it there.
5) Estimate the range in sample sizes that would be best to use in you sampling design from the plot you make.

Estimation of Power

Now you are going to ask a more complex question that relates directly to testing a hypothesis. Lets assume that you are going to set up an experiment to assess the effects of an herbivore (say rabbits) on your species. You suspect that caging may increase the cover – but you also need to estimate the effect size you want to detect. This will allow you to make a good estimate of the number of replicates that you need. As an example, if you want to be able to detect (statistically) a 10% increase you will need many more replicates than if you really are only interested in a 50% or greater increase. Hence, you need to decide what level of effect you want to see. For this activity I want you to estimate the replication you would need to statistically detect the following effect sizes: 10, 20, 30, 40 and 50% increases. Think about what this means for your system.
1) Calculate the mean and standard deviation on your preliminary data. Do this directly in Excel (these will be the last values in the calculation of mean and standard deviation that you did above – that is using the whole data set). Write these down

2) Start using power analysis using the following applet http://www.cs.uiowa.edu/~rlenth/Power/ Assume that you will be doing a two sample t-test to assess your data when you run the experiment. This is the correct procedure for your experiment because your will have two treatments (samples) – one that is a control and the other that is some sort of treatment that you predict will have an effect on your response variable. Therefore run the two-sample model. When you the table opens click options and select graph so that you will get a graph along with other output. The input table has the following variables:
   a. **Sigma** – this is the standard deviation. Both sigma 1 and 2 are requested (the standard deviations for treatment and control). Here you are assuming they are the same.
   b. **N1 and N2** – these are your samples sizes. Again assume that these should be the same
   c. **Alpha** – this is the critical p-value. Traditionally it is set to 0.05 – but this is your choice and it should reflect your consideration of the relative importance of making a Type 1 vs Type 2 error.
   d. **Two – tailed**. Here you need to decide what type of hypothesis you have. If your hypothesis is of the form:
      i. **Treatment mean will be different from the control** – choose two-tailed
      ii. **Treatment mean will be greater than the control**, or separately **Treatment mean will be less than the control** – choose one tailed.
   e. **True difference between means.** Here you need to input the desired difference between means. Hence if you want to pick up a 10% difference, input 10% of your mean.

3) Try graphing the important relationships
   a. **Try graphing**
      i. **Power vs sample size.** Here you want to use the standard deviation you found for your real data, alpha equal to what you have decided, one vs two tailed and a difference between the means equal to the minimum you think you want to detect. Put sample size (n1) as the x variable and power as the y.
      ii. **Power vs effect size (difference between the means).** Here you want to use the standard deviation you found for your real data, the sample size you think you can logistically use (or sample size that you came up with above), alpha equal to what you have decided. Now put difference between the means as the x variable and power as the y.
      iii. Save the data and export to excel. Now plot again using Excel graphics. equal to the minimum you think you want to detect.

4) OK now I want you to look at the relationship between Power (set at 0.8) and effect size, alpha ($\alpha$), and sample size. Make a table with these variables and solve for sample size (remember sample size here is expressed as either N1 or N2. If you are solving for total sample size $N = (N1 + N2)$ make sure this is clearly labeled. Note that in the table below the columns for N are shown for each of the alpha values.
<table>
<thead>
<tr>
<th>Power</th>
<th>Effect Size (%)</th>
<th>Alpha</th>
<th>N (α=0.05)</th>
<th>N (α=0.10)</th>
<th>N (α=0.15)</th>
<th>N (α=0.20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>10</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>20</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>30</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>40</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>50</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>60</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>70</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>80</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>90</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>100</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5) Make a graph of the relationship between Sample size (y-variable) and effect size (x-variable) for all four levels of critical P-value.

Now answer the question: What is your sampling design going to be and why? The answer to why will rely on the interaction between power, effect size and feasibility.