The Spatial Distribution of Ferns

Goals

1. Develop a testable hypothesis about the spatial distribution of ferns.
2. Obtain field data on the spatial distribution.
3. Develop a statistical test to test the hypothesis with the field data obtained.

Ferns

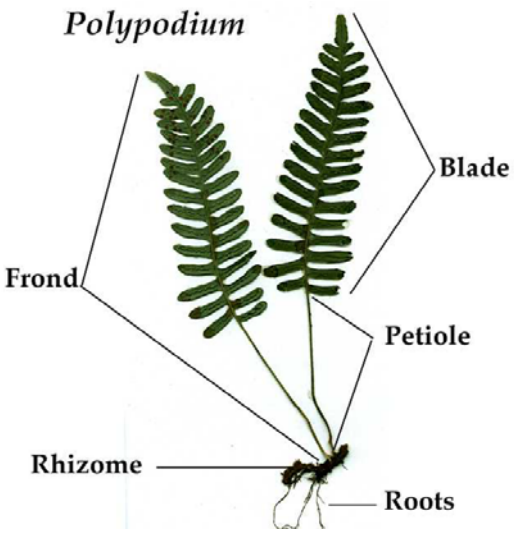


Figure 1: Parts of a fern (Source for Figure 1: http://www.csus.edu/indiv/r/reihmanm/ferns_and_fern_allies.htm)

Ferns are seedless, vascular plants, in the phylum Pteridophyta. This phylum also includes horsetails and clubmosses. Fernlike plants first appeared in the uppermost Lower Devonian, about 380 Million years ago. They are characterized by large compound leaves, called fronds, attached to a stalk that grows out of a rootstock or rhizome. The leaf may be subdivided into leaflets (or pinnae), which may be further subdivided into subleaflets (or pinnules). These may be even further subdivided.

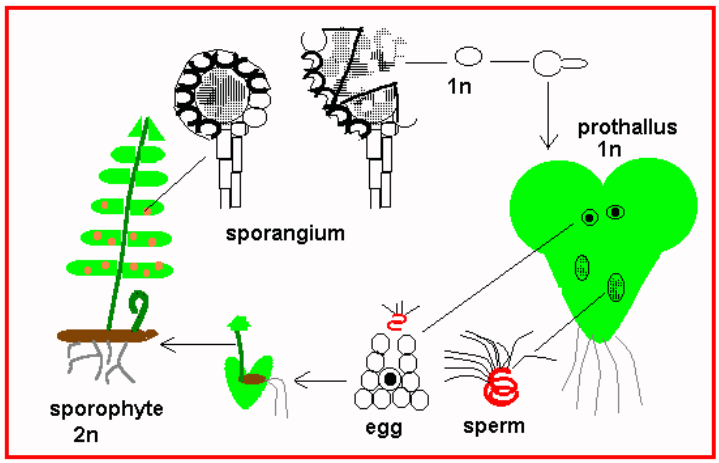


Figure 2: Reproduction (Source of diagram: http://www.hcs.ohio-state.edu/hcs300/svp2.htm)

Ferns reproduce by the means of spores. Spores develop into a haploid gametophyte, with both male and female organs that are microscopic in size. The male sexual organ, the anteridium, releases sperm upon maturity that swim towards the female sexual organ, the archegonium. When the egg in the archegonium is fertilized, the resulting zygote divides immediately and grows into the adult sporophyte, the stage that is easily visible to us. The sporophyte initially receives nutrition from the

gametophyte until it is large enough to produce enough photosynthate to become independent. Fertile leaves develop tiny masses of spore cases, called sporangia, on their undersides. Spores are produced in the sporangia by meiosis and are dispersed in dry weather. A large fern with many spore cases can produce millions of spores.

We will focus on the spatial distribution of one species of fern: the Bracken, *Pteridium aquilinum*.

Figure 3: *P. aquilinum* is a very common fern that lives in many areas but often indicates poor and barren soil. Its leaves are 36” long and 36” wide, triangular, and usually divided into three nearly equal parts. The leaves are almost horizontal to the ground.

Source: http://botit.botany.wisc.edu/images/401/Pteridophyta_(Polypodiophyta)/Dennstaedtiaceae/Pteridium_aquilinum/Frond_KS.html)

From Observations to Models

The spatial distribution of plants is of intriguing complexity. Look at an untreated lawn in early spring and observe the distribution of dandelions, or fly over a Minnesota forest in the fall and observe the distribution of differently colored trees.

Based on your experience, what kind of spatial patterns do you expect to see in plant populations? (Think about trees in a forest or in an orchard, or flowering plants in a prairie or maize in a field, or …) Can you come up with explanations for such patterns?

Types of Patterns	Explanations

Spatial patterns are important in other contexts as well. Find examples outside of the plant world and explain why we might be interested in the observed spatial patterns.

Example and Explanation

A Model for a Random Spatial Distribution

A spatial distribution of points can be classified according to how clumped or aggregated the distribution is. From an individual point's view, this can be phrased in terms of the likelihood that another point is nearby. If this likelihood is unaffected, then we say that the pattern is **random**. If this likelihood is increased, we say that the pattern is **clumped** or aggregated; and if the likelihood is reduced, we say that the pattern is **regular** or uniform. Regular patterns are also referred to as overdispersed and clumped patterns as underdispersed.

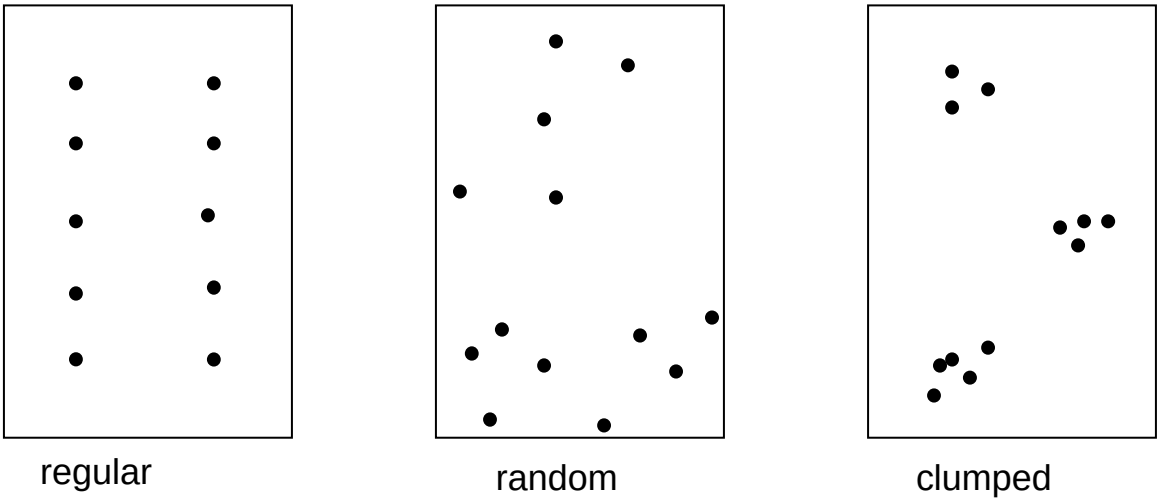


Figure 4: The three basic types of spatial patterns.

Random spatial patterns are easy to simulate on a computer. All we need is a way to generate the two coordinates of a point in a rectangle in a way so that the location of this point is completely random. This can be done using a spreadsheet that has a function that generates "random numbers" between 0 and 1. For instance, if we wanted to generate a random location of a point in a rectangle of length 2 and height 3 (such as in Figure 4), we would generate a random number between 0 and 1, multiply the number by 2 and would call the outcome the x-coordinate of the point. To get the y-coordinate, we would generate another random number between 0 and 1, and multiply the number by 3. (This is how the locations of the points in the middle rectangle in Figure 4 were generated.) Regular patterns can be generated by placing dots on a regular grid. Clumped distributions are much more difficult to generate automatically. In Figure 4, the dots were moved manually to create the clumps.

Assessing the Spatial Pattern

If the area where the organisms or events you want to study are located is very large, you cannot find all spatial locations. Instead, you might want to take a **random sample** and

analyze the spatial pattern based on your sample. A random sample is supposed to reflect the distribution in the entire study area.

Using random samples is a common procedure and is not restricted to biology. For instance, before a presidential election, we might want to know who the front runner is. It is not possible to ask every voter. Instead, a certain number of randomly selected people will be asked.

To get a better understanding of what constitutes a random sample, discuss which of the following samples could be considered random if you wanted to know who the front runner is in an upcoming election:

1. 50 students selected from the incoming freshman class.
2. 50 people selected from the middle of pages 100, 110, 120, 130, and 140 of the St. Paul phone book.
3. The first 50 people registering for lunch at the Democratic Convention.

Back to spatial patterns: Statisticians have found ways to quantify spatial patterns based on random samples. The idea is to count the number of "points" in a random sample of non-overlapping rectangles of equal area. To define the quantity that will help us to quantify spatial patterns, we need two other quantities: The **mean** is the average number of points per rectangle over all the rectangles that constitute the sample; the **variance** is a measure of how much individual values for each rectangle deviate from the mean. A measure of spatial aggregation is then the ratio of variance to mean. This quantity is called the **index of dispersion**.

$$\text{Index of Dispersion} = \frac{\text{Variance}}{\text{Mean}}$$

A mathematical model can be formulated that allows us to compute the index of dispersion under the assumption that the points are randomly distributed. We will simulate this kind of distribution later. For a random distribution, the index of dispersion is 1; for a clumped distribution, it is greater than 1, and for a regular distribution less than 1. We will see below how to estimate the mean and the variance, and hence the index of dispersion, for a given data set of spatial locations.

A Primer on Hypothesis Testing

Karl Popper (philosopher, 1902-1994) asserted that a hypothesis or theory can only be considered scientific if it is falsifiable. We need to keep this in mind when we propose a hypothesis. In statistics, hypotheses are often formulated to describe a situation of "no effect," and are referred to as a null hypothesis (H_0). For instance, the mathematical model of fern distribution assumes that the location has no effect on occurrence of ferns or that ferns do not have any spatial preferences. This hypothesis is testable since we can

simulate the distribution of ferns under this hypothesis and compare the actual data to our simulation.

H_0: Ferns do not have any spatial preferences.

The mathematical model we formulated above allows us to define a random variable X, which counts the number of ferns in a given area. By sampling in the woods, we can obtain a random sample of these counts. This assumes that disjoint sampling regions in the woods are independent, an assumption we made implicitly in the model.

We can treat simulated data just like real data. The data consists of the number of ferns in non-overlapping rectangles of equal size. Suppose the number of rectangles is n and x_i denotes the number of ferns in rectangle i, then the data is given by the vector $(x_1, x_2, \ldots, x_n)$.

A test statistic that is appropriate for testing the hypothesis is the index of dispersion

$$T(X) = \frac{\mathrm{var}(X)}{EX}$$

We choose a *significance level* α, for instance α=0.05, and define a *critical region* R_c such that

$$P(T \in R_c \mid H_0) = \alpha$$

We then reject H_0 if the observation falls into R_c. The critical region R_c can be one- or two-sided. If we wanted to reject H_0 whenever the test statistics is too large, we would determine t_0 so that

$$P(T > t_0 \mid H_0) = \alpha$$

that is, the critical region would be the interval $R_c = (t_0, \infty)$. If we wanted to reject H_0 whenever the test statistics is either too small or too large, we would determine t_1 and t_2 so that

$$P(T < t_1 \mid H_0) = \frac{\alpha}{2} \quad \text{and} \quad P(T > t_2 \mid H_0) = \frac{\alpha}{2}$$

that is, the critical region would be the interval $R_c = [0, t_1) \cup (t_2, \infty)$.

The significance level is also called type I error. It is the probability of rejecting H_0 even though it is assumed to be true.

When the test statistic does not fall into the critical region, we do *not* accept the hypothesis. Why not?

Before You Go Out Into the Field

We will form three groups. Each group will establish three rectangular plots with length 2m and width 1.5m. These plots are placed at random locations and are non-overlapping. Each plot can be constructed using a rope with length segments 3-4-5-4-3, as shown in Figure 5. A segment of length 1 measures 0.5m. (Why does this work?) Use the provided rope and the yard stick to mark the segments.

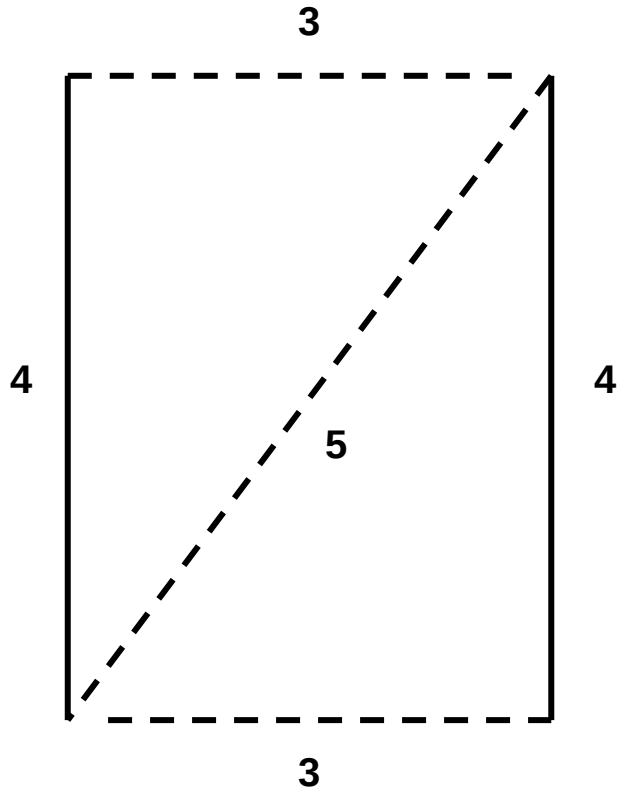


Figure 5: Using a 3-4-5-4-3 rope to stake out a 3 by 4 rectangle.

Each group will use a spreadsheet to find random locations for the three plots along a path. The plots should not overlap.

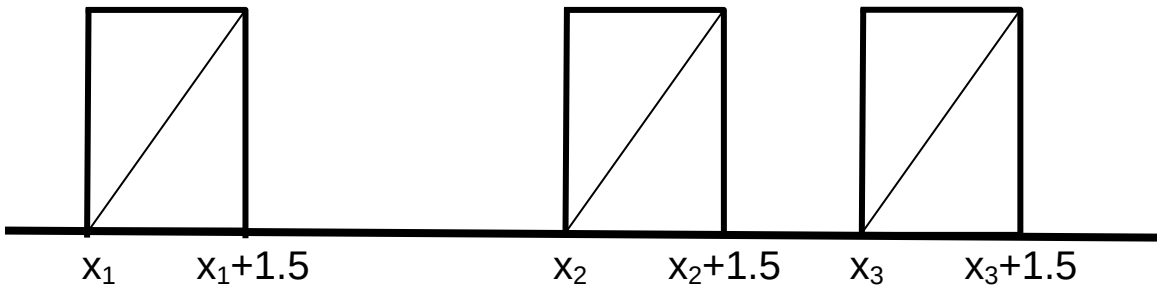


Figure 6: A configuration of the three plots along a path.

We will use a spreadsheet to establish the locations of the three plots, each 1.5x2 m^2. You can find the spreadsheet under the *Spatial Locations* tab in your workbook. Choose as the transect a 20 m segment and establish the locations of the left corners of each of the rectangles as shown in Figure 6. Below is an example of a spreadsheet with all the required fields filled in (your numbers will be different):

	A	B	C
1	**Establishing the Plots**		
2			
3		**Length**	
4		20	
5			
6	**PlotID**	**x**	**x+1.5**
7	**1**	5.4	6.9
8	**2**	3.2	4.7
9	**3**	17.4	18.9

You enter the length of the path into the cells B4. To generate random locations, you use the EXCEL function RAND(). This function generates a random number between 0 and 1. To find the x-coordinate of the first plot, you type "=B4*RAND()"in cell B7 (without the quotation marks). To see where the first plot ends, add 1.5 to the value of the x-coordinate in cell C7 (type: "=B7+1.5"). Find the x-coordinates of the other two plots. Make sure that the rectangles do not overlap. Enter the x coordinates into the table below.

PlotID	**x**
1	
2	
3	

Collecting your data

We will do one plot together: Number of ferns in common plot _____

We will then split into groups, and each group will establish the three plots in the designated study area. Count the number of ferns in each of the three plots.

Number of Ferns	1	2	3
Bracken (*Pteridium aquilinum*)			

Come back and pool your data with the other groups. Enter the data below (Plot 1 is the common plot)

	1	2	3	4	5	6	7	8	9	10
Bracken										

You are now ready to test your hypothesis.

Testing the Hypothesis

Use a spreadsheet to test the hypothesis of random distribution. Compute the arithmetic average, also called the mean, of the pooled data:

$$\hat{x} = \frac{1}{10}\left(x_1 + x_2 + \ldots + x_{10}\right)$$

You will also need the sample variance:

$$\hat{s}^2 = \frac{1}{9}\left((x_1 - \hat{x})^2 + (x_2 - \hat{x})^2 + \ldots + (x_{10} - \hat{x})^2\right)$$

The ratio $\frac{\hat{s}^2}{\hat{x}}$, called the estimated index of dispersion, tells you something about how clumped the distribution is, as explained above.

Use the spreadsheet with the *Data* tab to do the calculations. Enter the Plot ID and the Number of Ferns as shown below. EXCEL has functions that compute the mean and the variance of a sample. To compute the mean in cell B13, enter

=AVERAGE(B2:B11)

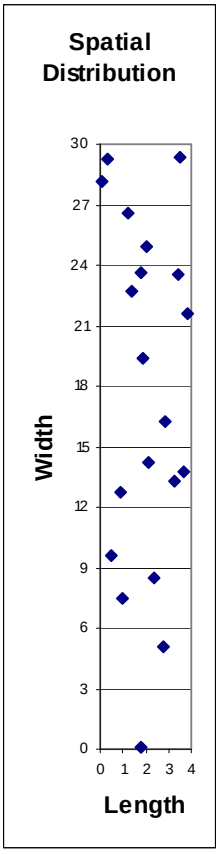
To find the variance of the sample in cell B14, enter

=VARA(B2:B11)

The ratio of variance over mean in cell B15 is computer as “=B14/B13”

	A	B
1	Plot ID	Number of Ferns
2	1	5
3	2	4
4	3	3
5	4	3
6	5	1
7	6	3
8	7	2
9	8	5
10	9	3
11	10	4
12		
13	Mean	3.3
14	Variance	1.56667
15	Ratio	0.47475

Simulating data



An example of a spreadsheet that was used to simulate the data is shown below. You can find the setup for the worksheet under the *Simulation* tab.

Here is how to proceed: Find the total number of ferns in your sample. Call this number *N*. Generate a total of *N* random points in a rectangle of size 4x30. (The workbook has room for 33 random numbers. Adjust that number accordingly.) Think of this large rectangle as a stack of ten rectangles, each of size 4 units by by 3 units, as shown in the figure at the left. The x-coordinates of each of the points are between 0 and 4. The y-coordinates are between 0 and 30. By looking at the y-coordinates, we can assign each random point to one of the ten rectangles. EXCEL has a function that can count how many points have y-coordinate greater than a given number. The function is called COUNTIF. The syntax is “COUNTIF(range,criteria)”. The range is a block of cells, and the criterion in our case is “>0”, “>3”, “>6”, etc. This will give us the cumulative number of points in the rectangles. If the y-values are in the cells C7 to C39, for instance, then range would be “C7:C39”. (The range for your data set will likely be different.) This is done in

cells E7-E16 (an example of the spreadsheet is given on the last page). For instance, in cell E7, we enter

=COUNTIF(C7:C39,">0")

In cell C8, we enter

=COUNTIF(C7:C39,">3")

and so on until we get to cell E 16 where we enter

=COUNTIF(C7:C39,">27")

Take successive differences to determine the number of points in each of the ten rectangles. This is done in the cells that are called PlotID and Number. The ten plots are listed under PlotID (cells G7 to G16). The number of points in each of the ten plots is computed in cells H7 to H16. For instance, in cell H7, we enter

=E7-E8

In cell H8, we enter

=E8-E9

and so on until we reach cell H15 where we enter "=E15-E16" and finally in cell H16, we enter "=E16"

The numbers in cells H7 to H16 are a simulated sample. We can compute the mean and the variance as before, and this is done in cells H18 and H19. The index of dispersion is computed in cell H20.

Manual Simulation

Record the value of the index of dispersion in column K in your spreadsheet and repeat this 100 times by pressing the F9 key repeatedly or by hitting the "ENTER" key every time after entering data. Take turns so that each member of the group contributes to the simulated data. If you can no longer see the simulated data and the cell where you enter the data in the same screen, enter the data in a neighboring column and then copy the values into column K once you have enough data points. Make sure that after you complete the simulations all your simulated data is in one column.

EXCEL allows you to arrange values in ascending (from smallest to largest) or descending (from largest to smallest) order. Order the simulated data from smallest to largest by first highlighting the values, then click on the pulldown menu "Data" (see bar on top of spreadsheet). Click on "Sort." A window pops up that asks for "Sort by" and whether you want it in ascending or descending order. Compare the index of dispersion

you found in the field to the simulated data. Where is your empirical value compared to the simulated data? Discuss what this means and decide whether you can reject your hypothesis? Can you suggest ways to improve this method?

Automated Simulation: Using Macros

Instead of entering simulated data by hand, you can define a Macro. Macros allow you to store repeated key strokes and to recall them using a designated shortcut.

Office 2003
First, change the mode of calculation to manual by selecting Tools|Options|Calculations. Open the Macro function on the Tools menu and select Record New Macro. Give the macro a name and choose shortcut, for instance, Ctrl+a. The sequence of key strokes is

- F9
- Select the cell where the ratio is stored and then open Edit|Copy
- Select the cell where you want the ratio to be stored and open Edit|Find. Leave the Find What box empty and use Search: By Columns. Select Next and Close.
- Open Edit|Paste Special|Paste Values. Click OK.
- Stop recording by opening Tools|Macro|Stop Recording

Office 2007
Under EXCEL Options>Formulas check “Manual” under the “Calculation Options.” To record a Macro, open the Tab “Developer.” Click on “Record Macro” in the “code” group. Give the macro a name and choose shortcut, for instance, Ctrl+a. The sequence of key strokes is

- F9
- Select the cell where the ratio is stored and select Copy (Home tab, Clipboard group)
- Select the cell where you want the ratio to be stored and Paste Values
- Select Insert in the “Cells” group, click “Insert Cells,” a window appears. Select “Shift cells down.” Click OK.
- Go to the Developer tab and click on “Stop Recording” in the “Code” group

Every time you press the shortcut, a new run is produced and the outcome (index of dispersion) is recorded. Repeat this 500 times.

Discussion Points and Decision

Report the results of your discussion and your findings to the group.

Spreadsheet for Simulation:

	A	B	C	D	E	F	G	H
1								
2		**Length**	**Width**					
3		4	3					
4								
5								
6	**N**	**x**	**y**		**Cumulative**		**PlotID**	**Number**
7	1	0.106879	22.75871		33		1	3
8	2	0.123193	23.02479		30		2	0
9	3	2.720773	25.5938		30		3	3
10	4	1.552516	20.17506		27		4	5
11	5	0.477106	0.349642		22		5	1
12	6	2.499836	27.02719		21		6	3
13	7	3.737013	15.26634		18		7	3
14	8	2.05383	10.91528		15		8	5
15	9	3.507987	24.73725		10		9	5
16	10	0.781924	24.10329		5		10	5
17	11	0.118978	7.394131					
18	12	1.329531	22.85521				**Mean**	3.3
19	13	2.171359	13.00545				**Variance**	3.122222
20	14	1.758451	0.452771				**Ratio**	0.946128
21	15	2.593692	24.54374					
22	16	1.49157	7.568293					
23	17	1.553832	28.18565					
24	18	1.026137	11.13653					
25	19	2.534118	1.897444					
26	20	0.417807	15.39207					
27	21	3.009593	10.95195					
28	22	2.054654	24.34626					
29	23	3.604348	10.84289					
30	24	3.143299	27.82247					
31	25	3.184348	8.985998					
32	26	2.32413	28.7448					
33	27	2.12135	20.44114					
34	28	2.622062	28.61156					
35	29	0.119892	23.74629					
36	30	2.308892	19.90606					
37	31	2.553017	23.95617					
38	32	2.029437	16.89722					
39	33	3.944014	11.39197					