SNPs and Hypothesis Testing

Goals

1. Explore a data set on SNPs.
2. Develop a mathematical model for the distribution of SNPs and simulate it.
3. Assess spatial patterns.
4. Develop a testable hypothesis about the distribution of SNPs.
5. Use data set to explore the hypothesis.
6. Develop a statistical test and test the hypothesis by simulating the null hypothesis.

SNPs

The genetic make-up of an organism is stored in DNA (deoxyribonucleic acid), a polymer composed of nucleotides with a backbone of sugar and phosphate molecules. While the DNA sequences of individuals of the same species are quite similar, there is variation that allows for distinguishing two individuals. The most common form of segregating variation in individuals at the molecular level is variation in single nucleotides, most of which are caused by mutations at a single nucleotide that result in changes in the base. Data on single nucleotide polymorphism (SNP) are important in, for instance, detection of disease loci or developing genetic tests for pathogenicity. Data on SNPs are now published for many organisms. A good place to start is the National Center for Biotechnology Information (NCBI) whose web site is at http://www.ncbi.nlm.nih.gov/About/index.html. This web site has a primer on SNPs (http://www.ncbi.nlm.nih.gov/About/primer/snps.html) and a link to the public SNP database (http://www.ncbi.nlm.nih.gov/SNP/).

A Data Set

The following is from http://www.ncbi.nlm.nih.gov/genomes/static/Salmonella_SNPS.html (In the accompanying EXCEL spreadsheet, you will find the table of locations of SNPs under the “Raw Data” tab.)

“Researchers at the Egg Safety and Quality Research Unit (ESQRU) of the U. S. Department of Agriculture, Agricultural Research Service (USDA-ARS) have used microarray technology to identify a set of potential single nucleotide polymorphisms (SNPs) that distinguish two PT13a strains of S. Enteritidis that are genetically related but phenotypically divergent. Using the S. Enteritidis PT4 genomic sequence from the Sanger Institute as a starting point, a set of overlapping primers were generated and used in conjunction with Nimblegen Systems technology to detect SNPs that distinguish S. Enteritidis PT13a strain 21027 (bf), which is biofilm-forming but does not contaminate eggs, from S. Enteritidis PT13a strain 21046 (wt), which contaminates eggs but does not form biofilm. BLAST comparisons of the SNPs with other Salmonella enterica annotated databases revealed potentially relevant phenotypic traits matched to genetic loci. Since this data is considered preliminary it should be used with caution.”
The table lists the location of 195 SNPs and the probable change in amino acid sequence if any. Of these 195 confirmed SNPs, 3 occurred in the virulence plasmid, 102 altered predicted amino acid sequences, and 33 occurred in either non-coding or ribosomal DNA regions. Two deletions were detected in wt strain 21046, which does not form biofilm but retains the ability to contaminate eggs. Three SNPs in bf strain 21027, which forms biofilm but does not contaminate eggs, removed stop codons, whereas two single base pair SNPs in the wt strain introduced stop codons. ZipA of wt strain 21046 accumulated numerous SNPs that altered amino acid sequence.

**Task 1: Explore the data set**

Use the second sheet under the tab Data Analysis to extract the locations of the Chromosomal SNPs (see Figure 1 below):

Cell A1: Enter the name “SNP Locations”

Cells A2-A193: copy the SNP locations of the chromosomal DNA that are in cells A10:A201 under the tab Raw Data. Note that Ctrl+C copies a highlighted area and Ctrl+V pastes the values in the designated area.

(a) Use a Line Chart to plot the locations of the SNPs. To do this, highlight the cells with the SNP locations, click on the Chart Wizard, and select the “Line” Chart Type. Follow the instructions on the window that pops up. Make sure you label the axes appropriately. What does the slope of this curve tell you about the density of the SNPs along the chromosome?

(b) Using bins of size 125,000 nucleotides, construct a histogram of the number of SNPs per bin along the chromosome. Enter D1, D2, E1, E2, E3, and E4 as shown below. In Column D, enter the Bin numbers 1-40 (only the first eight bin numbers are shown and only the first 38 bins will be needed if the bins are of length 125,000). In Column E enter the right boundary of each of the bins. For instance, in Cell E3, enter “=D3*$E$1” (without the quotation marks). To count the cumulative number of points in each bin, use the COUNTIF function. For instance, in cell F3, enter “=COUNTIF($A$2:$A$193,"<"&E3)” Column G has the actual numbers of SNPs in each bin by subtracting consecutive numbers in Column F. Use the Chart Wizard to graph the resulting histogram. Does your answer in (a) agree with the histogram? Change the bin size and discuss what you see.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNP Locations</td>
<td>Bin Size</td>
<td>125000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20572</td>
<td>Bin</td>
<td>125000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>26369</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>69867</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>83428</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>114120</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>134979</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>172931</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>214228</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>229683</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Snapshot of the worksheet
Mathematical Model for Distribution of Mutations along Chromosomes

To analyze the data further, we will build a simple mathematical model and then test the data against the model.

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.

Random spatial patterns are easy to simulate on a computer. This can be done in any dimension. All we need is a way to generate the coordinates of a point on a line, in a rectangle, or a cube 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 the interval (2,5), we would generate a random number between 0 and 1, multiply the number by 3, and then add 2. The resulting number would then be the x-coordinate of the point. Regular patterns are also easy to generate, simply space points regularly on a line. Clumped distributions are much more difficult to generate and we will not provide a recipe.

To apply this to mutations along a DNA sequence, assume that the DNA sequence is of length $N$ and that mutations occur at random along this sequence. This might be the case if all mutations were neutral. Given that there are $K$ mutations, we would choose $K$ out of the $N$ sites at random. These sites would be locations of SNPs. This is easy to simulate. We can use simulated data to compare them to observed data from actual DNA sequences to test whether SNPs are randomly distributed.

Task 2: Simulate the mathematical model

Simulate the number of SNPs in the first 2,500,000 nucleotides of the data in the Simulation sheet. To do this, find the number of SNPs in this region and then simulate the locations of the SNPs. We will make the simplifying assumption that multiple hits at a site are allowed. [The RAND() function generates pseudorandom numbers in the interval (0,1). The INT function rounds a number down to the closest integer. The F9 key allows you to generate different realizations of this simulation.] The simulated data can be treated just like real data, and you should repeat Task 1, Part (b) to get a better feel for the variation in the different realizations.

A snapshot of your spreadsheet is included below in Figure 2. Note that since this is a stochastic simulation, your values of the simulated data will vary from the sheet below.

Cell A1: Enter the number 2500000
Cells B1, D1,E1,D2,E2,F2,G2: Enter what is displayed in Figure 2.
Cells A2 and following: Enter the numbers 1, 2, ..., up to the number of SNPs in the first 2,500,000 nucleotides of the Salmonella genome (you can get the number from the analysis you did in Task 1).
Cell B2: Enter “=INT($A$1*RAND())” (without the quotation marks) and then drag the cell down to generate the locations of the other SNPs.

To generate the histogram, proceed as in Task 1, part (b).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2500000</td>
<td>SNP Locations</td>
<td>Bin Size</td>
<td>125000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>580666</td>
<td>Bin</td>
<td>Bin Border</td>
<td>Cumulative</td>
<td>#</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2250551</td>
<td>1</td>
<td>125000</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>628186</td>
<td>2</td>
<td>250000</td>
<td>10</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2398056</td>
<td>3</td>
<td>375000</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1682151</td>
<td>4</td>
<td>500000</td>
<td>19</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>1578569</td>
<td>5</td>
<td>625000</td>
<td>23</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>2217142</td>
<td>6</td>
<td>750000</td>
<td>28</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>2325087</td>
<td>7</td>
<td>875000</td>
<td>32</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>88756</td>
<td>8</td>
<td>1000000</td>
<td>36</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Snapshot of the worksheet

**Assessing Spatial Patterns**

While looking at different realizations will give you a better sense of what it means to be “randomly distributed,” it does not allow you to assess whether the real data deviates from this distribution. To do this, we define a quantity that will allow us to do this comparison.

Statisticians have found ways to quantify spatial patterns. To define the quantity that will help us to quantify the spatial pattern of points on a line, we need two other quantities: The **mean** is the average number of points in a bin/interval; the **variance** is a measure of how much individual values for each interval deviate from the mean. A measure of spatial aggregation is the ratio of variance to mean. This quantity is called the **index of dispersion**.

\[
\text{Index of Dispersion} = \frac{\text{Variance}}{\text{Mean}}
\]

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.

While the index of dispersion is simply a number that depends on the underlying distribution, if we simulated SNPs so that the simulated data resembles the experimental design of the data we collected, we would find that the index of dispersion based on the simulated data (this is a sample index of dispersion) varies from simulation run to simulation run. We will see below how
to compute the sample mean and the sample variance based on simulations or collected data, and hence the sample index of dispersion, for a given data set of spatial locations.

**Formulating a Question**

Observations help us to develop questions. This may guide us in identifying the type of data we need to collect to find answers to our questions. Thinking about spatial patterns and what we have learned so far, what kind of questions could we ask? What kind of information would we need to answer these questions?

The way this is often done is to formulate a statement that can be tested using a mathematical model. The statement is called a null hypothesis.

Discuss and formulate a null hypothesis regarding spatial patterns of SNPs along a DNA sequence:

<table>
<thead>
<tr>
<th>Null Hypothesis:</th>
</tr>
</thead>
</table>

Under this hypothesis, we expect a particular spatial pattern. We can use the computer to generate patterns that are consistent with this hypothesis and then compare the data to the predictions.

**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$_0$). We would hypothesize that the location of nucleotides has no effect on the occurrence of a mutation. This is reflected in the mathematical model of SNP distribution where we assumed that SNPs are randomly distributed along the chromosome. Our hypothesis is testable since we can simulate the distribution of SNPs under this hypothesis and compare the actual data to our simulation.

H$_0$: All nucleotides along the chromosome are equally likely to be the location of a SNP.

The mathematical model we formulated above allows us to define a random variable $X$, which counts the number of SNPs in a bin/interval of a given length. By sampling along the chromosome, we can obtain a random sample of these counts. This assumes that disjoint
segments along the chromosome 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 SNPs in non-overlapping bins of equal length. Suppose the number of bins is \( n \) and \( x_i \) denotes the number of SNPs in bin \( i \), 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

\[
T(X) = \frac{\text{var}(X)}{EX}
\]

We choose a significance level \( \alpha \), for instance \( \alpha = 0.05 \), and define a critical region \( R_c \) such that

\[
P(T \in R_c | H_0) = \alpha
\]

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

\[
P(T > t_0 | H_0) = \alpha
\]

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

\[
P(T < t_1 | H_0) = \frac{\alpha}{2} \quad \text{and} \quad P(T > t_2 | H_0) = \frac{\alpha}{2}
\]

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 because 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?

**Testing the Hypothesis**

Use a spreadsheet to test the hypothesis of random distribution. Assume that the sample size is \( n=20 \) (namely the number of bins). Compute the sample mean, also called the arithmetic average, of the bin counts:

\[
\hat{x} = \frac{1}{n} x_1 + x_2 + \ldots + x_{20}
\]

You will also need the sample variance:
The ratio \( \frac{s^2}{\bar{x}} \), called the estimated or sample index of dispersion, tells you something about how clumped the distribution is, as explained above.

Use the spreadsheet with the *Data Analysis* tab to do the calculations. EXCEL has functions that compute the mean and the variance of a sample. To compute the mean of the first twenty bins, enter into cell J3 (see Figure 3):

\[
=\text{AVERAGE(G3:G22)}
\]

To find the variance of the sample, enter in cell J4

\[
=\text{VAR(G3:G22)}
\]

The ratio of variance over mean in cell J5 is computed as “\( =J4/J3 \)”

<table>
<thead>
<tr>
<th></th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3.7</td>
</tr>
<tr>
<td>Variance</td>
<td>4.326316</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.169275</td>
</tr>
</tbody>
</table>

*Figure 3: Columns I and J of the Data Analysis sheet*

**Task 3**

We will now repeatedly simulate locations of SNPs on the first 2,500,000 nucleotides along the chromosomes and calculate the index of dispersion for each simulation run. Enter cells I3:J5 of the *Data Analysis* sheet into the *Simulation* sheet (see Figure 3). Each time you press the F9 key, you will see a new run of the simulation. To keep track of the simulations, enter the numbers 1 to 500 into cells L3:L502. Press the F9 key, copy the content of cell J5, and use Paste As Values to paste the index of dispersion into cell M3. You could repeat this now many times. But this would be quite tedious. For repetitive tasks on EXCEL, you can use Macros. Macros allow you to store repeated key strokes and to recall them using a designated shortcut.

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+y. 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.
Every time you press the shortcut, a new run is produced and the outcome (index of dispersion) is recorded. Test the hypothesis now by simulating the coefficient of dispersion 500 times under the null hypothesis for the first 2,500,000 nucleotides. Order the simulated values from smallest to largest.

Change the mode of calculation back to Automatic. Produce a histogram for the index of dispersion. Use a total of ten bins. To find the appropriate bin size, calculate the maximum value of the simulated index of dispersion (the function MAX(number 1, number 2,...) can find the maximum value among numbers), and then divide the maximum value by 10.

Explore ways that would allow you to assess whether the data deviates from the null hypothesis.

In our analysis, we stopped just short of the bin that had a fairly high number of SNPs. Repeat your analysis with this bin included or with a different segment of the DNA that includes this bin.