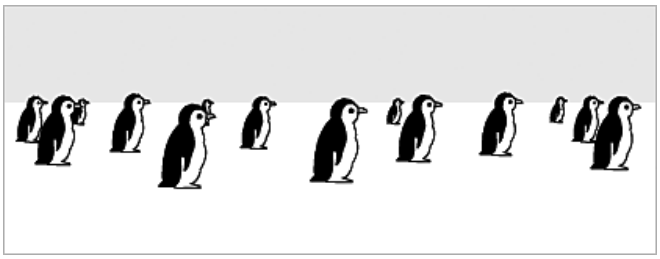
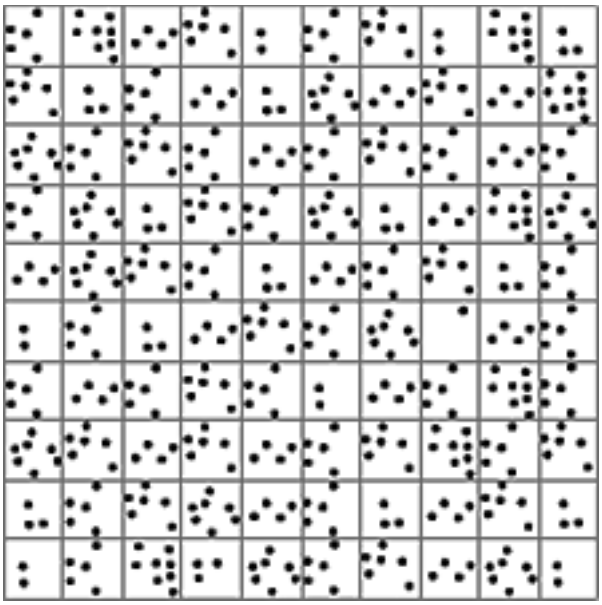
Counting Penguins



Statisticians often use a random sample to estimate characteristics of a population when the population is very large and they cannot obtain data on every individual in the population. Statistical estimation asks the fundamental questions "What can I say about a whole population based on information from a random sample of that population?" and "To what degree can I say that my estimate is accurate?" Let's put random sampling into action to answer a question about demographics: "How many penguins are there on a particular ice floe in the Antarctic?"

Counting a penguin population can be tricky. Penguins tend to move around and swim off, and it's cold! So scientists use aerial photographs and statistical sampling to estimate population size. Some of the techniques they use are quite sophisticated, but we can look at a simplified version of their approach to examine the basic ideas of random sampling and estimation. Imagine a large, snow-covered, square region of the Antarctic that is inhabited by penguins. From above, it would look like a white square sprinkled with black dots:



If you had access to such an aerial view, you could count the dots to determine the number of penguins in this region. But suppose the region was too large to see in one photo. You might instead take 100 photographs of the 100 smaller square sub-regions, count the penguins in each sub-region, and total these to obtain a count for the entire region. However, this might take too long and be too expensive. So here's another alternative: You can select a representative sample of the sub-regions, obtain photos of only these, and use the counts from these sub-regions to estimate the total number of penguins in the entire region.

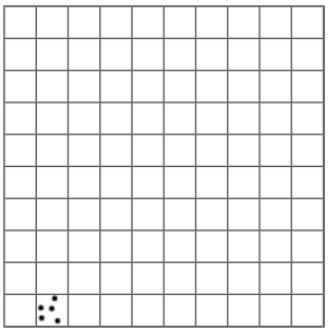
Making Estimates

A possible sample might look like the one below. Let's explore how we might use the information in this sample to estimate the total number of penguins in the entire region.

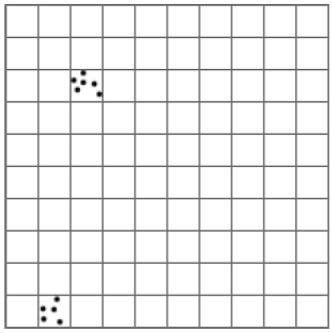
Problem A

Suppose you had access to three samples: one with a single photo of one of the 100 sub-regions, one with photos of two sub-regions, and one with photos of three sub-regions. Use the results from each of these samples (pictured below) to make an estimate of the total number of penguins in the entire region (i.e., all 100 sub-regions).

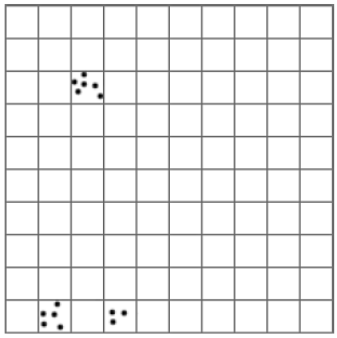
Sample A: $n = 1$ (sample size = 1)



Sample B: $n = 2$ (sample size = 2)



Sample C: $n = 3$ (sample size = 3)



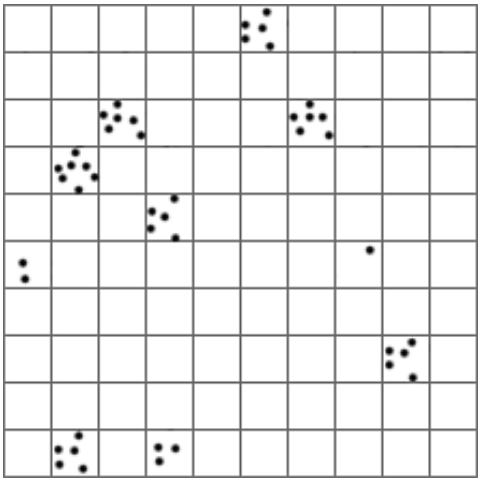
Record your counts and estimates in the table below.

Sample	Photo 1	Photo 2	Photo 3	Estimate of Total
A		N/A	N/A	
B			N/A	
C				

You may have determined a general rule for estimating the number of penguins in the entire population. One useful method is to find the mean of the counts in the sample and then multiply the mean by 100 (the number of sub-regions).

In making estimates by sampling, there is a balancing act in selecting the sample size. A larger sample size may cost more money or be more difficult to generate, but it should provide a more accurate estimate of the population characteristic you are studying. On the other hand, a sample size that is too small may not be accurate enough for you to be certain of your results.

Below is a sample of 10 sub-regions. Based on the number of penguins in this sample, make an estimate of the number of penguins in the entire region:



Fair Sampling

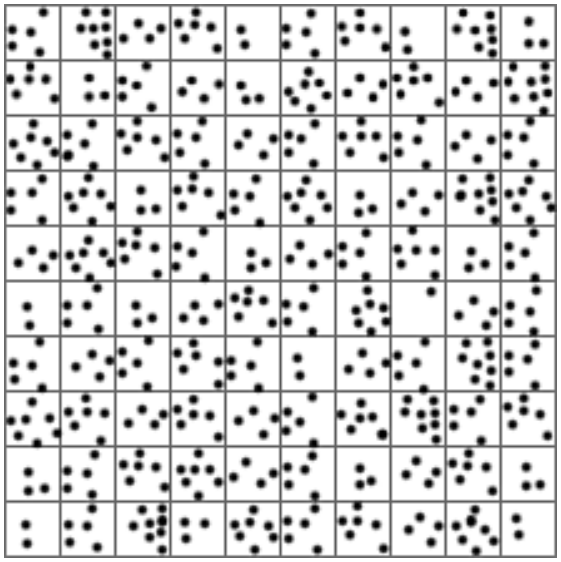
You may have noticed that your estimates for the total penguin population vary quite a bit based on both the sample size and which sub-regions were sampled. The decision about how to select a sample, accordingly, is a critical one in statistics. It is important that each part of the population be treated fairly. If you are fair in the selection, then you should obtain a representative sample and thus a more fair estimation procedure. In earlier sessions, you looked at notions of fairness and randomness and noticed that people have a difficult time being fair or random. So what methods can you use to accomplish fair sampling?

For this next section, you are going to sample the penguin population using different sampling methods. A picture of the penguins has been included for each sampling method so you can hi-light the sub-region that has been selected. It may be helpful to number the top and left sides of each picture so that you can quickly find your sub-region of penguins. Below is the table that you will use to record all of your information about the sample.

Sampling Method	Mean Penguins per sub-region	Estimate of total penguins
Simple Random Sample		
Vertical Strata Sample		
Horizontal Strata Sample		
Systematic Sample		
Clustered Sample		

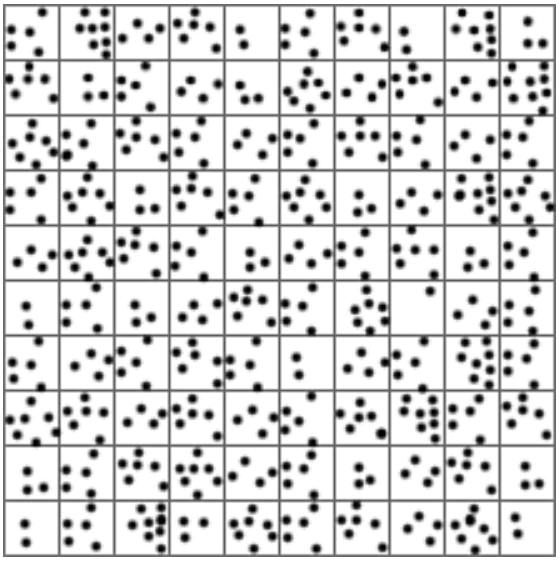
Simple Random Sample

Use your calculator or a random number table to choose 10 sub-regions of penguins. Hi-lite them below and record your calculations in the table above.



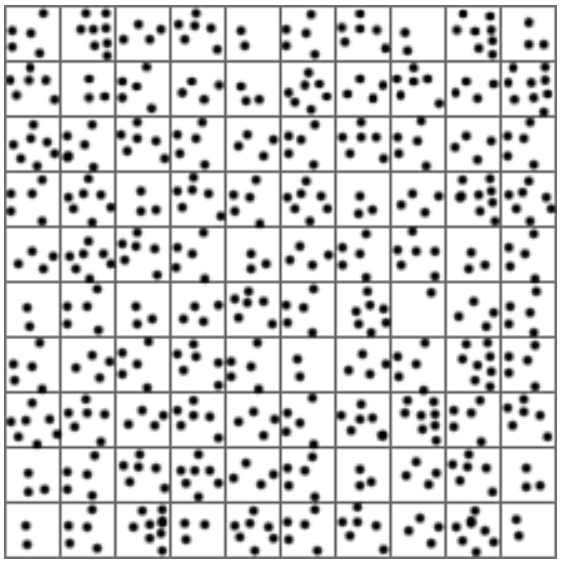
Vertical Strata Sample

Use your calculator or a random number table to choose 1 sub-region of penguins for each vertical column. Hi-lite them below and record your calculations in the table above.



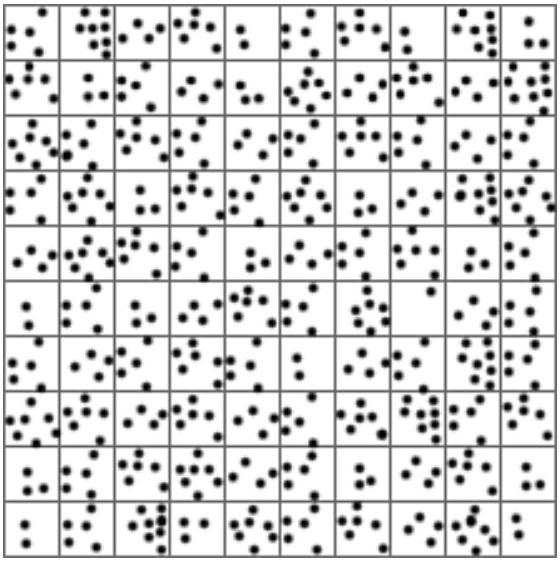
Horizontal Strata Sample

Use your calculator or a random number table to choose 1 sub-region of penguins for each horizontal row. Hi-lite them below and record your calculations in the table above.



Systematic Sample

Create a systematic sample of 10 random sub-groups. Start by choosing a random sub-group and then apply your system. State where you started and what you used for your system.



Clustered Sample

Create a clustered sample of 9 random sub-groups (we will do 9 instead of 10 because 9 will make an perfect square). Start by choosing a random sub-group and then determine if that will be the center of your square, the top-left, top-right, etc. State where you started and how you drew your square for your cluster.