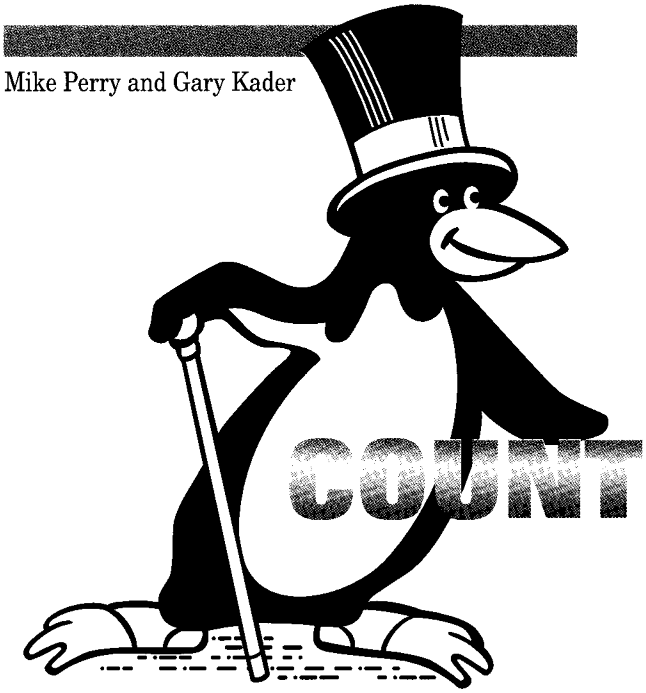
Mike Perry and Gary Kader



COUNTING PENGUINS

THE NCTM'S CURRICULUM STANDARDS for statistics give a specific objective for students in grades 9–12: to "understand sampling and recognize its role in statistical claims" (NCTM 1989, 167). The use of random samples for estimation is a fundamental statistical concept. Random sampling and its consequences can be studied through simulated sampling activities. The nature of sampling variability, the influence of sample size on the quality of estimation, and the role of the underlying population distribution are ideas that can be illustrated with repeated sampling.

Estimating population size is a common type of problem that often requires "spatial" sampling. For instance, estimating the size of an animal population can require selecting samples over a geographic region. Estimating the number of cars in a parking lot requires spatial sampling. Crowd estimation at such events as the Million Man March in Washington, D. C., uses aerial photographs. These types of problems are discussed in Scheaffer, Mendenhall, and Ott (1990).

To obtain estimates of the size of the penguin population of a region in Antarctica, the map of the region is divided into subregions. A random sample of the subregions is selected, and aerial photographs are taken of each selected subregion. Penguins appear as dots on the photographs; thus the photographs produce data for the estimation of the size of the total population.

The activity described herein is a simplification of the actual penguin-counting process but employs the same basic ideas and principles. It is intended for grades 9–12. The level of sophistication of the interpretations can be changed to match the level of the students. The full activity requires two to three class periods, but restricted versions can be completed within a single class. The following materials are required.

Sampling board. A region of Antarctica is represented by the sampling board. The 32-inch-by-32-inch sampling board is laid out in a 10 × 10 grid with margins as shown in **figure 1.** Each 3-inch-by-3-inch square represents a subregion. The borders are numbered 0 through 9 so that each square on the grid can be located by a pair of coordinates. To consider question 3 of the activity, three boards are needed.

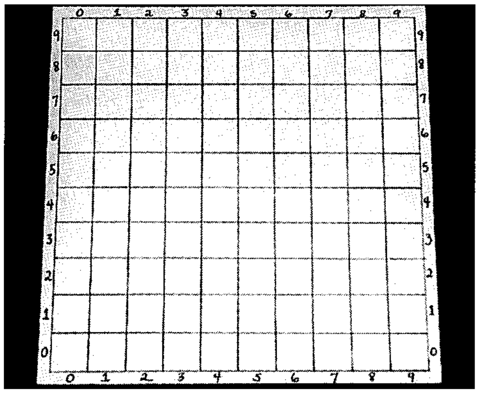


Fig. 1
The sampling board

Mike Perry, perrylm@appstate.edu, and Gary Kader, kadergd@appstate.edu, teach at Appalachian State University, Boone, NC 28608. They are interested in statistics and activity-based instruction.

Penguin cards. One hundred penguin cards are required for each sampling board. Each card is slightly smaller than 3 inches by 3 inches. One side of the card shows a pattern of dots—the penguins. The other side of the card shows a coordinate pair that corresponds to the card's position on the sampling board. **Figure 2** shows a set of penguin cards on the sampling board with the dots showing. **Figure 3** shows the same cards turned over, with the coordinates faceup. To consider question 3, we use three sets of penguin cards, which are described in **table 1.** In other words, board A is set up with no blank cards, one card with a single dot, six cards with two dots, and so forth.

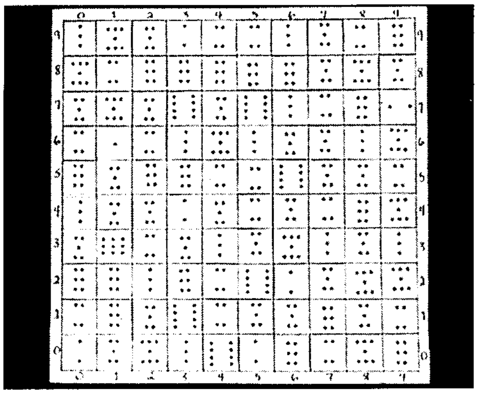
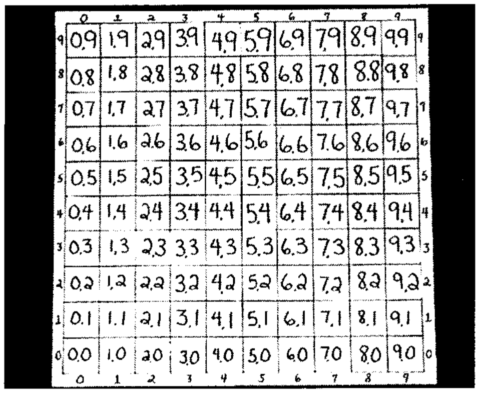


Fig. 2
Penguin cards with dots showing



	0	1	2	3	4	5	6	7	8	9	
9	0.9	1.9	2.9	3.9	4.9	5.9	6.9	7.9	8.9	9.9	9
8	0.8	1.8	2.8	3.8	4.8	5.8	6.8	7.8	8.8	9.8	8
7	0.7	1.7	2.7	3.7	4.7	5.7	6.7	7.7	8.7	9.7	7
6	0.6	1.6	2.6	3.6	4.6	5.6	6.6	7.6	8.6	9.6	6
5	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	5
4	0.4	1.4	2.4	3.4	4.4	5.4	6.4	7.4	8.4	9.4	4
3	0.3	1.3	2.3	3.3	4.3	5.3	6.3	7.3	8.3	9.3	3
2	0.2	1.2	2.2	3.2	4.2	5.2	6.2	7.2	8.2	9.2	2
1	0.1	1.1	2.1	3.1	4.1	5.1	6.1	7.1	8.1	9.1	1
0	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	0
	0	1	2	3	4	5	6	7	8	9	

Fig. 3
Penguin cards with coordinates showing

Random-number generator. We prefer to use a die with ten outcomes (for example, an icosahedron). If these dice are not available, a random-digit table or a random-number generator on a calculator could be used.

Recording sheets. The data sheets in **figure 4** are designed specifically for the data summaries in this activity.

TABLE 1
Distribution of Dots on Cards

Board A		Board B		Board C	
No. of Dots	Frequency	No. of Dots	Frequency	No. of Dots	Frequency
0	0	0	9	0	20
1	1	1	9	1	14
2	6	2	9	2	9
3	12	3	9	3	5
4	18	4	9	4	2
5	26	5	10	5	0
6	18	6	9	6	2
7	12	7	9	7	5
8	6	8	9	8	9
9	1	9	9	9	14
10	0	10	9	10	20
$\mu = 5$ $\sigma \approx 1.65$		$\mu = 5$ $\sigma \approx 3.15$		$\mu = 5$ $\sigma \approx 4.07$	

Data

Board ____________ Sample No. ____________

Observation	Random Coordinates	Number of Penguins
1	____________	____________
2	____________	____________
3	____________	____________
4	____________	____________
5	____________	____________
6	____________	____________
7	____________	____________
8	____________	____________
9	____________	____________
10	____________	____________

Sum : ____________
Mean : ____________
Estimate of N (100 × Mean) : ____________

Board ____________ Sample No. ____________

Observation	Random Coordinates	Number of Penguins
1	____________	____________
2	____________	____________
3	____________	____________
4	____________	____________
5	____________	____________
6	____________	____________
7	____________	____________
8	____________	____________
9	____________	____________
10	____________	____________

Sum : ____________
Mean : ____________
Estimate of N (100 × Mean) : ____________

Combine two samples; sample size $n = 20$

Sum 1 + Sum 2 : ____________
Mean : ____________
Estimate of N (100 × Mean) : ____________

Fig. 4
Recording sheets

THE ESTIMATOR

From a random sample of size n, we can estimate the average, or mean, number of dots per square on the entire board with the average, or mean, number of dots per square in our sample. We denote this sample mean by $\overline{X}$. An estimate for the total number of dots N on the board is given by $\hat{N} = 100\overline{X}$.

THE QUESTIONS

1. Is this method a "good" way to estimate the number of dots on the board?
2. How does sample size affect this estimation procedure?
3. How does the distribution of dots on the board affect the estimation procedure?

Random selection tends to produce similar results

To study these questions, two criteria for a good estimation procedure are considered:

- Different samples should give similar results, at least most of the time.
- The estimator should be unbiased. It should neither systematically underestimate nor systematically overestimate the number of dots on the board.

In other words, a large number of independent samples should produce estimates that *(a)* do not vary too much and *(b)* give, on the average, the correct result.

We can study these two properties with repeated sampling, that is, by considering a large number of independent samples. A good rule of thumb for a study of this type is to have about one hundred samples so that reasonable inferences might be made from the data.

THE DATA

Each sampling board is set up with the coordinates showing as in **figure 3.** Each student selects at least two samples of size $n = 10$ from each of the boards A, B, and C. In a class of about twenty-five students, each student should select four samples from each board for a total of twelve samples per student. This quantity can be adjusted according to class size to get about one hundred samples from each board, but the number of samples should always be a multiple of 2 because we later combine samples to get samples of size $n = 20$.

To select a sample of size $n = 10$, roll the die two times to select randomly a coordinate pair. In this way, select ten pairs for each sample. These pairs should be recorded on the data sheet. For each pair, select the corresponding penguin card, turn it over, and record the number of dots on the data sheet. **Figure 5** shows a sample of size $n = 10$ on the sampling board.

A card can be used more than once within each sample. If a repetition of random coordinate pairs occurs for a particular sample, the repetition is used in the sample. This procedure is often referred to as sampling "with replacement" because when a card is selected, it is placed back on the board before the next card is chosen.

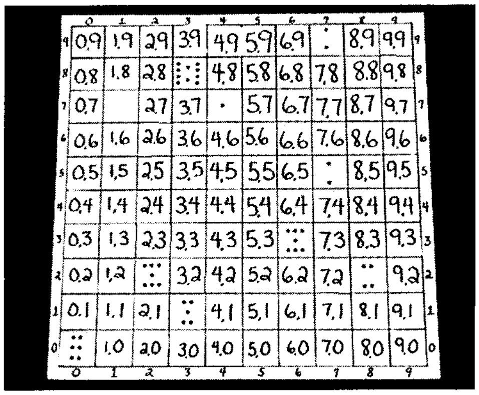


Fig. 5
Penguin cards with sample showing

ANALYSIS

We illustrate several types of analyses, which can be varied to match the needs of a given class. The data used were collected by forty-seven teachers who attended a summer institute during July 1995. Each participant selected two samples from each board, giving a total of ninety-four samples from each. These samples were pooled. **Figure 6** gives stem-and-leaf plots of estimates of N for samples of size $n = 10$. Each data sheet contains two samples of size $n = 10$. These samples are combined to give estimates of N that are based on samples of size $n = 20$ as outlined at the bottom of each data sheet. In the case of the data from the institute participants, forty-seven estimates were given. **Figure 7** gives stem-and-leaf plots for these results. In **table 2,** means and quartile summaries are presented for each of the six cases, and the box plots are shown for comparison in **figure 8.** In **figures 6** and **7,** the stems are split into h, higher, and l, lower, sets.

Of course, when another class does the activity, the data will be different. However, random selection will tend to produce similar results with some differences resulting from random variation.

INTERPRETATION

The cards on each of boards A, B, and C have a total of $N = 500$ dots. The estimator $\hat{N} = 100\overline{X}$ is an unbiased estimator of $N = 500$. The average of the estimates after repeated sampling should be close to 500. Note the means in **table 2.** The averages, or means, of the estimates based on ninety-four samples of size $n = 10$ from boards A, B, and C are 485,

Board A

```
1l  |
1h  |
2l  |  30
2h  |
3l  |
3h  |  70,70,80,90
4l  |  00,00,00,10,10,20,20,20,20,30,30,30,40,40,40,40
4h  |  50,50,50,50,50,50,50,60,60,60,70,70,70,70,70,70,80,80,80,80,80,80,80,80,80,90,90,90,90,90,90
5l  |  00,00,00,00,00,00,00,00,10,10,10,10,10,10,10,20,20,20,30,30,30,30,30,30,30,40,40,40,40
5h  |  50,50,50,60,60,80,80,80,80,80
6l  |  00,00
6h  |  70
7l  |
7h  |
8l  |
8h  |
```

Board B

```
1l  |
1h  |  80
2l  |
2h  |  60,90
3l  |  20,40
3h  |  50,50,60,60,60,70,90,90
4l  |  00,10,10,20,20,20,30,30,30,30,40
4h  |  50,50,60,60,60,60,60,60,60,80,80,80,80,80,80,80,80,90,90,90,90
5l  |  00,00,00,00,00,00,00,10,10,10,10,10,10,10,20,20,20,20,30,40,40,40,40,40
5h  |  50,50,50,60,60,60,60,70,70,70,80,80,80,90,90,90,90,90
6l  |  00,20
6h  |  50,50,60,70
7l  |  10
7h  |
8l  |
8h  |
```

Board C

```
1l  |
1h  |  70
2l  |  00
2h  |  70
3l  |  00,10,10,20
3h  |  50,50,60,60,60,60,70,70,80,90
4l  |  00,00,00,00,00,00,10,10,10,10,20,30,30,40,40
4h  |  50,50,50,50,50,50,60,60,60,60,60,70,70,70,70,80,90,90,90
5l  |  00,00,10,10,10,10,30,30,30,40,40
5h  |  50,60,60,70,70,80,80,90,90,90
6l  |  00,10,10,10,30,30
6h  |  50,60,60,60,60,80,80,90,90,90
7l  |  20,40,40
7h  |  50,70
8l  |
8h  |  50
```

Fig. 6
Stem-and-leaf plots of estimates, n = 10

491, and 498. The averages for the forty-seven samples of size n = 20 are 480, 481, and 514.

The box plots in **figure 8** suggest some answers to all three questions and are useful for comparing results for the two sample sizes and for comparing results from the three sampling boards. The corresponding numbers are in **table 2.** Estimates falling either 1.5 (IQR) below the lower quartile or 1.5 (IQR) above the upper quartile have been designated as outliers, indicated by asterisks in the box plots. A thorough discussion of outliers is given by Landwehr and Watkins (1994).

First consider the effect of sample size. In the first two box plots, which summarize estimates from board A, we see an apparent reduction in the variation of estimates when the sample size is increased from n = 10 to n = 20. The next four box plots show similar comparisons for the results from boards B and C. The reduction is most pronounced for board C, with an interquartile range of 180 for samples of size n = 10 and an interquartile range of 115 for n = 20.

Next consider the differences in results among the three boards. In the three box plots for samples