Session 9

Random Sampling and Estimation

Key Terms for This Session

Previously Introduced

- box plot
- distribution
- interval
- population
- random assignment
- sample
- stem and leaf plot

New in This Session

- sample mean
- sample size
- sampling with replacement
- sampling without replacement

Introduction

In this session, you will learn how to use results from a random sample to estimate characteristics of an entire population. To predict the accuracy of your estimates, you will investigate the variation in estimates based on repeated random samples from that population. **[See Note 1]**

Learning Objectives

In this session, you will estimate population quantities from a random sample. You will learn how to do the following:

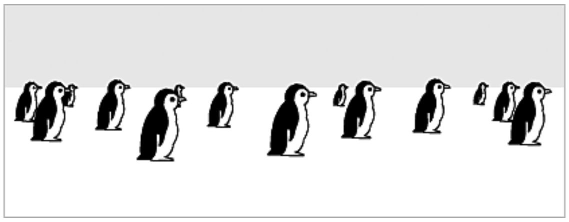
- Define an estimate based on sample data
- Select a random sample
- Describe sample-to-sample variation
- Predict the accuracy of an estimate
- Assess the effect of sample size on the accuracy of an estimate

Note 1. This session considers the use of random sampling for estimating characteristics of an entire population. Random sampling leads to random variation in estimates, and this variation can be described by a probability distribution. The normal curve approximation, which some statistics learners may be familiar with, is described only briefly. A stem and leaf plot of typical results of independent sample estimates is used for the investigations, which removes a level of abstraction from the description of sampling concepts.

Part A: Random Samples (15 minutes)

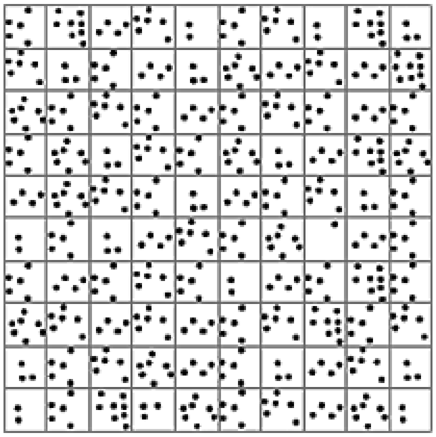
Counting Penguins

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. **[See Note 2]**

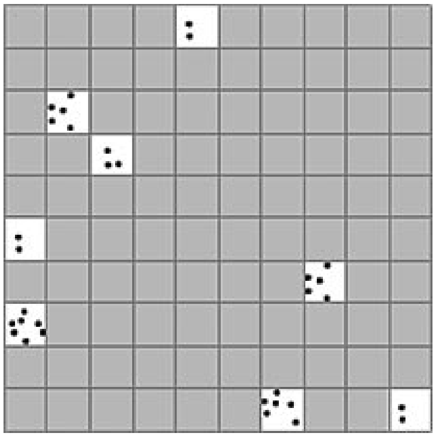
Note 2. Working with a spatial representation of a population offers several advantages for the introduction of sampling ideas. You can picture the population, and, more importantly, you can view samples in relation to the population.

You are asked to think about how you might use the information in a sample to estimate the total number of penguins in the entire region. Some textbook presentations of sampling and estimation skip this question; the text gives the definition of the estimator and proceeds from there. It is important to understand that the estimator is a human invention, and that you can choose your own method for estimating a total number based on a sample.

Part A, cont'd.

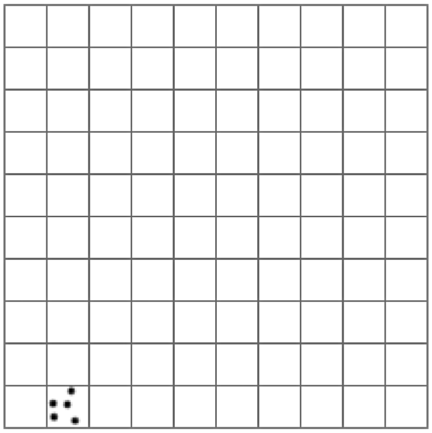
Making Estimates

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.

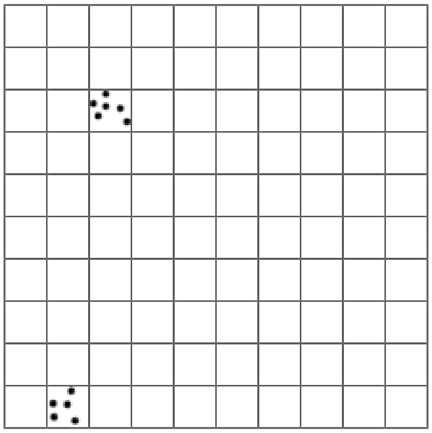


Problem A1. Suppose you had access to three samples: one with a single photo of one of the 100 sub-regions, one with photos of two sub-regions, and one with photos of three sub-regions. Use the results from each of these samples (pictured below) to make an estimate of the total number of penguins in the entire region (i.e., all 100 sub-regions).

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



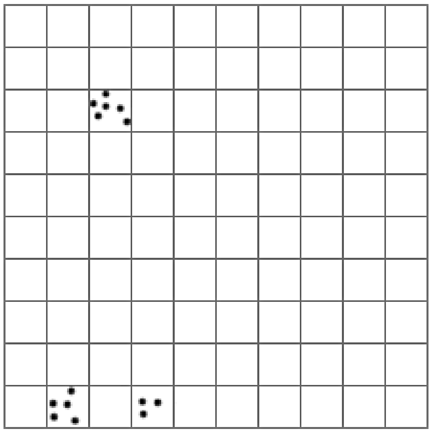
Sample B: *n* = 2 (sample size = 2)



Part A, cont'd.

Problem A1, cont'd.

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



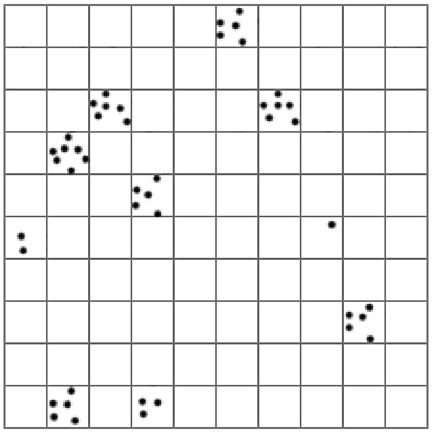
Record your counts and estimates in this table:

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

[See Tip A1, page 285]

In Problem A1, 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). **[See Note 3]**

Problem A2. 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:



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.

Note 3. Some people will come up with a workable idea for estimating the total number of penguins immediately, while others may need some direction. It helps to start with a sample of one sub-region, as directed in this part. Some people will suggest multiplying the number of dots in the one sub-region by 100. Next, a sample of two and then three sub-regions can evolve into the idea of averaging the number of dots in the sub-regions before multiplying by 100.

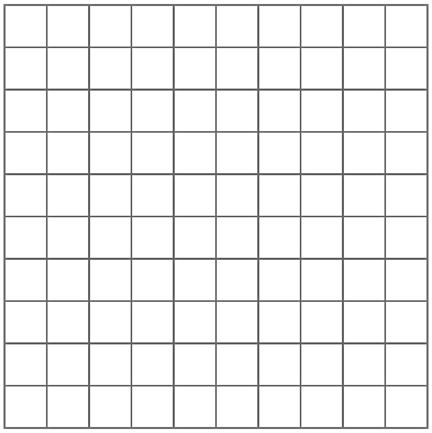
Part B: Selecting the Sample (30 minutes)

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? **[See Note 4]**

Problem B1. How might you select 10 sub-regions from the 100 total sub-regions so that you would be most likely to have a "representative" sample for estimating the size of the penguin population in the entire region? You can use the empty chart below to explore your ideas. **[See Tip B1, page 285]**



Video Segment (approximate times: 6:26-8:06): You can find this segment on the session video approximately 6 minutes and 26 seconds after the Annenberg/CPB logo.

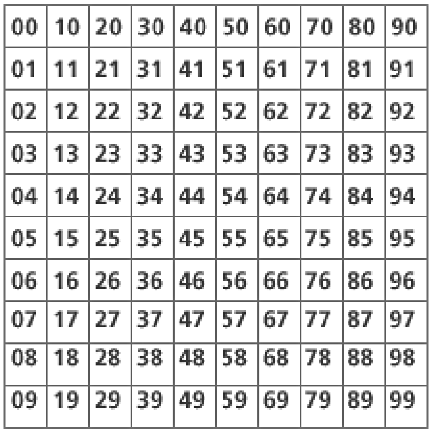
In this video segment, groups of participants devise methods for collecting a random sample of penguins. Watch this segment after you have completed Problem B1 and compare your method with that of the onscreen participants. Do these methods ensure that the samples will be random?

Note 4. Take time to develop your own ideas. There are many different ways to randomly select 10 sub-regions. Developing a method of selection will help you clarify the concept as well as provide a tool for the practice of sampling. After you have considered your own methods, you can then investigate the specific methods introduced in Part B.

Part B, cont'd.

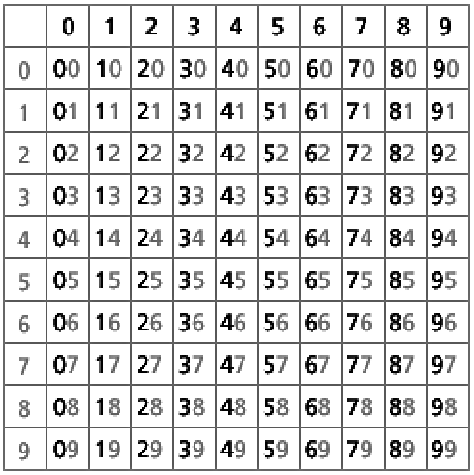
A Fair Sampling Method

There are many different ways to randomly select 10 sub-regions. Many of these methods involve initially numbering the 100 sub-regions. In this section, we will use the numbering system below, which numbers the sub-regions from 00 through 99:



00	10	20	30	40	50	60	70	80	90
01	11	21	31	41	51	61	71	81	91
02	12	22	32	42	52	62	72	82	92
03	13	23	33	43	53	63	73	83	93
04	14	24	34	44	54	64	74	84	94
05	15	25	35	45	55	65	75	85	95
06	16	26	36	46	56	66	76	86	96
07	17	27	37	47	57	67	77	87	97
08	18	28	38	48	58	68	78	88	98
09	19	29	39	49	59	69	79	89	99

Locating number positions is easier if we put digits on the outside borders as shown. Each number in the grid corresponds to a black and gray number combination; the black number is the first digit, and the gray number is the second digit:



	0	1	2	3	4	5	6	7	8	9
0	00	10	20	30	40	50	60	70	80	90
1	01	11	21	31	41	51	61	71	81	91
2	02	12	22	32	42	52	62	72	82	92
3	03	13	23	33	43	53	63	73	83	93
4	04	14	24	34	44	54	64	74	84	94
5	05	15	25	35	45	55	65	75	85	95
6	06	16	26	36	46	56	66	76	86	96
7	07	17	27	37	47	57	67	77	87	97
8	08	18	28	38	48	58	68	78	88	98
9	09	19	29	39	49	59	69	79	89	99

Problem B2. Think of a way to pick 10 numbers between 00 and 99 at random. (You may prefer to select each digit individually, or to select the entire two-digit number at once.) Then use your method to generate the 10 random numbers. **[See Tip B2, page 285]**

Part B, cont'd.

One possible method for solving Problem B2 is to use two 10-sided dice, one red and one blue. The sub-region can then be determined by the two dice (in the order red, and then blue).

You might notice that the random selection process will sometimes produce duplicates. There is a greater than one-third chance that 10 numbers picked at random between 00 and 99 will produce at least one duplicate, and almost a 90% chance that 20 such numbers will produce at least one duplicate.

For instance, you might find that seven tosses of the dice produced these sub-region choices:

19 22 39 50 34 05 39

If we do not want duplicates, we can skip them until we get 10 distinct numbers, for example:

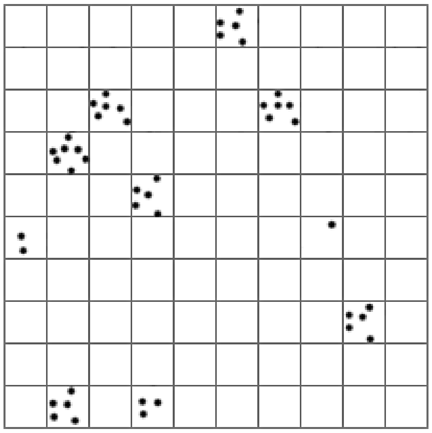
19 22 39 50 34 05 75 62 87 13

This is called sampling without replacement, since each time we choose a sub-region we remove it from the list of sub-regions we can choose on the next toss of the dice. In some experiments, it may be impractical or impossible to exclude duplicates from the random selection process. If duplicates are allowed, it is called sampling with replacement.

The 10 distinct numbers (19, 22, 39, 50, 34, 05, 75, 62, 87, 13) correspond to these 10 sub-regions:

	0	1	2	3	4	5	6	7	8	9
0						50				
1										
2			22				62			
3		13								
4				34						
5	05							75		
6										
7									87	
8										
9		19		39						

Here is a look at the number of penguins in each of the 10 sub-regions we selected:

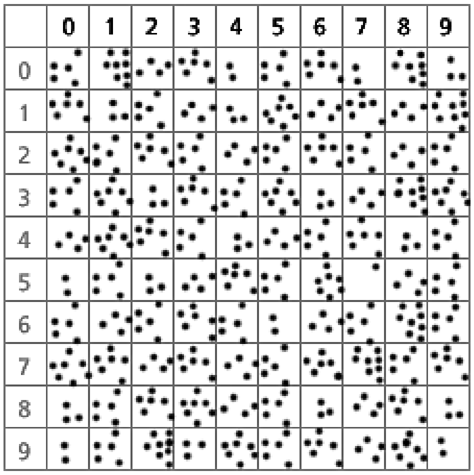


The estimate of the total number of penguins for the entire region based on this random sample of 10 sub-regions is as follows:

100 x [(5 + 6 + 6 + 7 + 5 + 2 + 1+ 5 + 5 + 3)/10] = 100 x (45/10) = 450

Part B, cont'd.

Problem B3. Use the random sample you found in Problem B2 to estimate the total number of penguins in the region. Find your 10 random sub-regions in the chart below:



Write and Reflect

Problem B4. Did you expect your estimate from Problem B3 to equal your estimate from Problem B2? Why or why not? What explains this variation? If the sample size were increased to 20 sub-regions, would you expect the variation in the estimates to increase or decrease? Why?

Part B, cont'd.

Variation in Estimates

A computer can perform random sampling and estimation much more quickly than you can by hand. In this problem, you will obtain three more samples of 10 sub-regions and then calculate the total number of penguins in the region.

Problem B5. Use the following three sets of random samples that were generated by a computer.

	Sample 1	Sample 2	Sample 3
Sub-region 1	6	3	5
Sub-region 2	7	3	4
Sub-region 3	6	5	4
Sub-region 4	5	6	5
Sub-region 5	8	7	2
Sub-region 6	7	3	4
Sub-region 7	7	4	3
Sub-region 8	8	7	7
Sub-region 9	5	2	4
Sub-region 10	8	4	5
Sample Total	67	44	43
Sample Average	67 / 10 = 6.7	44 / 10 = 4.4	43 / 10 = 4.3
Total Estimate	6.7 x 100 = 670	4.4 x 100 = 440	4.3 x 100 = 430

Record your estimates in the following table:

Sample	Estimate
First Sample	450
Problem B3 Sample	
Computer-Generated Sample 1	
Computer-Generated Sample 2	
Computer-Generated Sample 3	

Based on these five estimates, how many penguins do you think there are in the entire region?

Try It Online! **www.learner.org**

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 9, Part B, Problem B5.

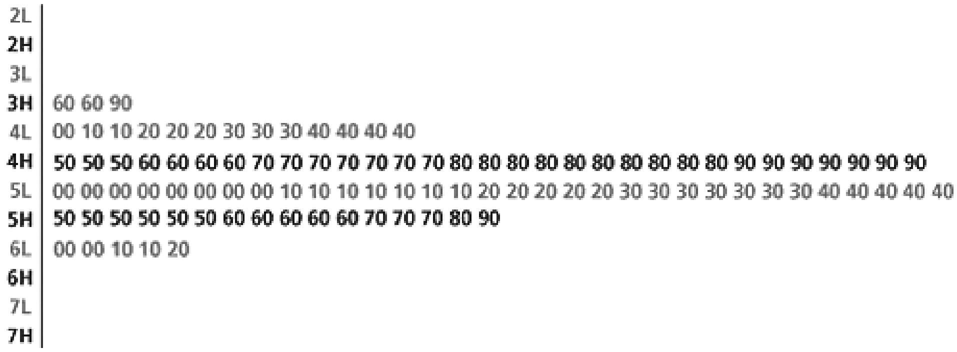
Part C: Investigating Variation in Estimates (45 minutes)

Using a Stem and Leaf Plot

In Part B, you obtained several different estimates for the total number of penguins in this region based on the different samples you chose. **[See Note 5]**

We can use a stem and leaf plot to help us organize the estimates and to determine any patterns that exist in the distribution of the estimates.

In the manner you used to generate your own estimates, 100 estimates of the penguin population count were produced from independently selected random samples of size 10. Here are these 100 estimates in a stem and leaf plot, where the intervals are of size 50:



2L	
2H	
3L	
3H	60 60 90
4L	00 10 10 20 20 20 30 30 30 40 40 40 40
4H	**50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H	**50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90**
6L	00 00 10 10 20
6H	
7L	
7H	

Note that in this stem and leaf plot, the spacing on the stems is 50. For example, the stem marked "3L" displays all the estimates between 300 and 349, while the stem marked "3H" displays all the estimates between 350 and 399. Also, since the samples are of size 10, all the estimates are multiples of 10.

Problem C1.

a. Based on the stem and leaf plot of these 100 estimates, make a guess for the actual number of penguins in the region.

b. Give an interval of values in which you are fairly certain the actual number of penguins in the region lies. (This interval should include the guess you made in the question above!) **[See Tip C1(b), page 285]**

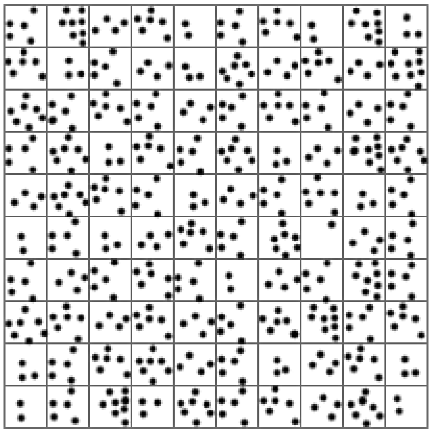
Note 5. The presentation used here is based on 100 estimates of the size of the penguin population, which were produced from independently selected random samples of size 10. These are "typical" of what you would expect to get if another 100 samples of size 10 were selected: You would obtain a similar (but not exactly the same) pattern exhibited in the stem and leaf plot of estimates.

Though statistics textbooks might be more likely to use a "continuous" model to illustrate the idea of sampling distributions, this is a somewhat more concrete and accessible way to demonstrate the same concepts.

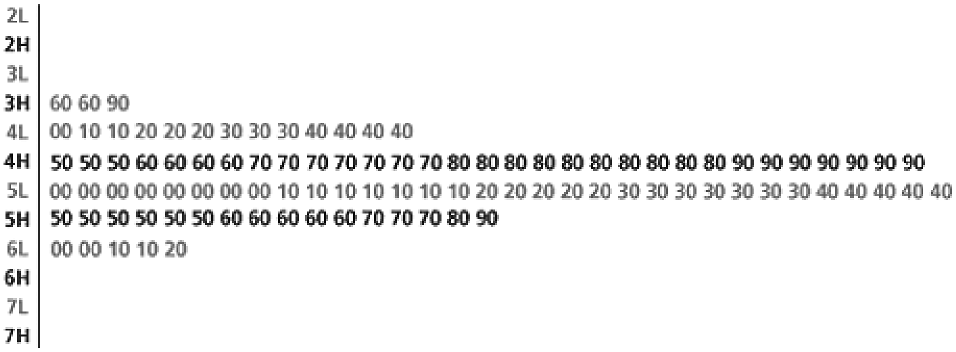
Part C, cont'd.

Judging the Quality of Estimates

Here is the entire region we've been studying. If you actually count all the penguins, you'd find that there are exactly 500. Notice that the number of penguins varies from sub-region to sub-region. Some of the squares in the grid contain as few as one penguin, and some contain as many as nine. On average, each of the 100 sub-regions contains five penguins.



Problem C2. Now that you know the actual total number of penguins, let's examine the stem and leaf plot of the 100 estimates from sample size 10:



2L	
2H	
3L	
3H	60 60 90
4L	00 10 10 20 20 20 30 30 30 40 40 40 40
4H	50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H	50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L	00 00 10 10 20
6H	
7L	
7H	

a. What is the best estimate? For how many samples did this estimate occur?

b. What are the six worst estimates?

c. What percentage of the estimates are 50 or fewer penguins away from the actual total?

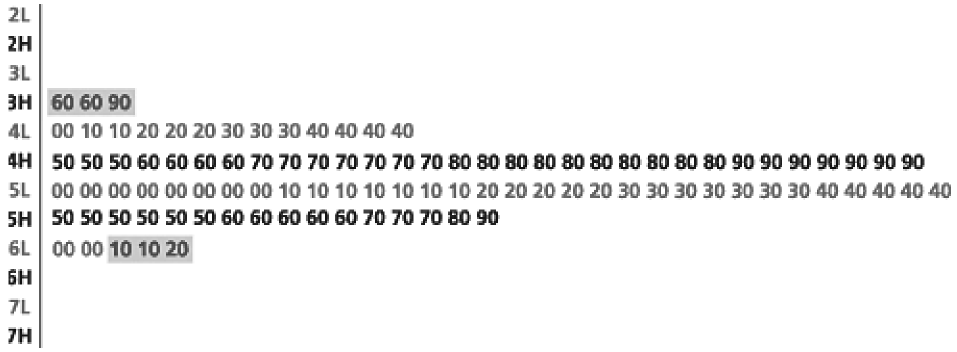
d. What percentage of the estimates are 100 or fewer penguins away from the actual total?

[See Tip C2, page 285]

Part C, cont'd.

Intervals

The six worst estimates are shown in the gray boxes on the stem and leaf plot:



```
2L |
2H |
3L |
3H | 60 60 90
4L | 00 10 10 20 20 20 30 30 30 40 40 40 40
4H | 50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H | 50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L | 00 00 10 10 20
6H |
7L |
7H |
```

These six estimates are the most different from 500. Specifically, they differ from 500 by more than 100; they are either less than 400 or greater than 600.

The other 94 estimates differ from 500 by 100 or less. These 94 estimates fall between 400 and 600 (inclusive).

We'll refer to these inclusive ranges of values as intervals, and use such intervals to classify the estimates:

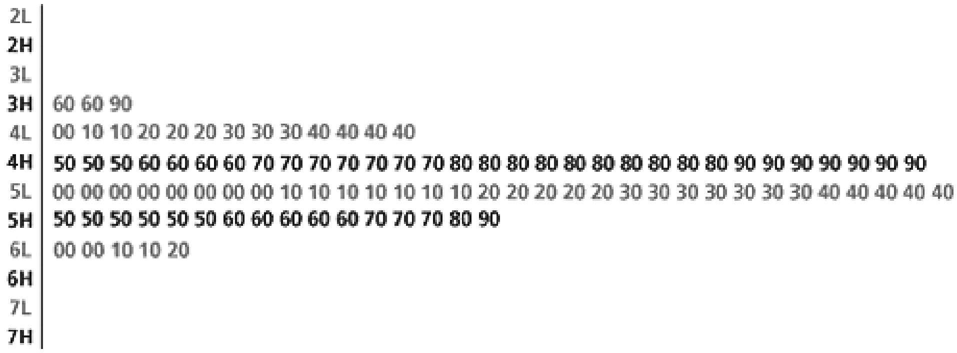
- Ninety-four of 100 (94/100) estimates fall between 400 and 600 (inclusive).
- Six of 100 (6/100) estimates fall outside of this interval.

Problem C3.

a. What proportion of the estimates are 75 or fewer penguins away from the actual value?

b. What proportion of the estimates are more than 75 penguins away from the actual value?

Problem C4.



```
2L |
2H |
3L |
3H | 60 60 90
4L | 00 10 10 20 20 20 30 30 30 40 40 40 40
4H | 50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H | 50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L | 00 00 10 10 20
6H |
7L |
7H |
```

Part C, cont'd.

Problem C4, cont'd.

Use the stem and leaf plot on the previous page to complete this table:

Interval (Inclusive)	Proportion of Estimates in Interval	Proportion of Estimates Not in Interval
350-650		
375-625		
400-600	94/100	6/100
425-575	84/100	16/100
450-550		
475-525		

Describing Intervals

These six intervals provide a description of how widely the estimates vary from sample to sample, and how close the estimates are to the actual value of 500: **[See Note 6]**

Interval	Interval Range	Interval Radius	Proportion of Estimates in Interval
350-650	300	150	100/100
375-625	250	125	98/100
400-600	200	100	94/100
425-575	150	75	84/100
450-550	100	50	69/100
475-525	50	25	37/100

The interval from 350 to 650 is the largest interval in the table above; its interval range is 300. This tells us two things:

- All (100/100) of the estimates are between 350 and 650, a range of 300.
- These estimates fall within 150 (the interval radius) of 500.

The interval 475 to 525 is the smallest interval in the table. This tells us two things:

- Fewer than half (37/100) of the estimates are between 475 and 525, a range of 50.
- These estimates fall within 25 (the interval radius) of 500.

Problem C5.

a. Explain why it is useful for the proportion of estimates in an interval to be high.

b. Explain why it is useful for the interval range to be small.

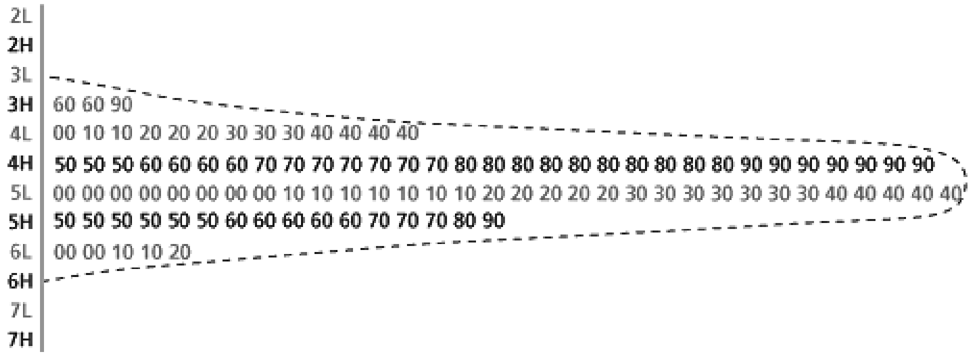
c. What happens to the proportion of estimates in the interval as the interval range decreases?

Note 6. The use of intervals demonstrated in this session is a very important statistical idea. It is the conceptual basis for the Confidence Interval Estimation. More advanced texts will use continuous models, such as the normal distribution, as approximate descriptions of sampling distributions, and then develop interval ideas based on these models. The intent here is to provide an understanding of the concepts in a less formal and perhaps more readily understandable setting.

Part C, cont'd.

Probabilities

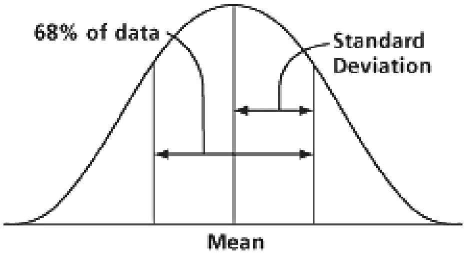
We have worked with a stem and leaf plot of the distribution of estimates of a population based on 100 random samples of size 10. The display is reasonably bell-shaped, with estimates occurring on both sides of 500 (the actual total number of penguins). There is a concentration of estimates around 500, with fewer estimates occurring as you move farther away from 500. **[See Note 7]**



We can think of these estimates as "typical" of what you would get if you were to select another 100 samples of size 10. That is, you would generate a similar (but not exactly the same) distribution. The stem and leaf plot would also be similar, and you would expect about the same proportions of estimates to fall into the intervals we identified earlier.

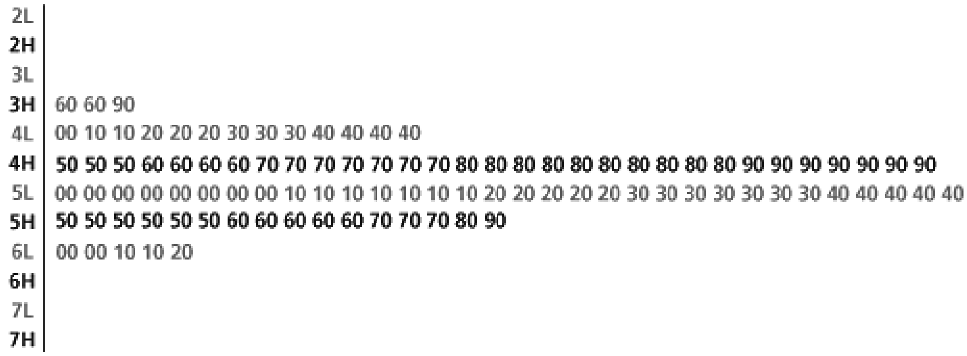
Under normal circumstances, if you were asked to estimate the size of a population, you wouldn't already know the population size—otherwise, you wouldn't need to estimate it! Also, you would not repeatedly select samples as we did in this session. In practice, you take only one sample to make your estimate based on the results in your sample.

Note 7. The normal distribution curve is symmetric and bell-shaped. It is characterized by the mean and the standard deviation (see below). The mean is located at the center of the distribution curve, and the standard deviation determines the width of that curve. Approximately 68% of the data values fall within one standard deviation of the mean, and 95% of the data values fall within two standard deviations of the mean.



Part C, cont’d.

How can you predict how accurate that one sample is likely to be? For our problem of counting penguins, we can use probability to make that prediction, using the “typical” distribution we found for the 100 samples:



2L	
2H	
3L	
3H	60 60 90
4L	00 10 10 20 20 20 30 30 30 40 40 40 40
4H	50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H	50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L	00 00 10 10 20
6H	
7L	
7H	

Let’s say that the one sample you found yielded 360 for your estimate. This is not a very good estimate, since the actual population size is 500. But since only two of our samples produced this estimate, the probability of coming up with that estimate is only about 2/100.

On the other hand, your sample might generate an estimate of 500, right on target! Your probability for this is approximately 8/100, because eight of the samples produced an estimate of 500.

Problem C6. Here is the table of intervals from Problem C5:

Interval	Interval Range	Interval Radius	Proportion of Estimates in Interval
350-650	300	150	100/100
375-625	250	125	98/100
400-600	200	100	94/100
425-575	150	75	84/100
450-550	100	50	69/100
475-525	50	25	37/100

a. What is the probability that an estimate will fall in the interval from 425 to 575?

b. What is the probability that the estimate will fall in the smallest interval listed above?

c. According to the table, how likely is it that one sample of 10 sub-regions will give an answer within 100 penguins of the actual number?

Part D: The Effect of Sample Size (30 minutes)

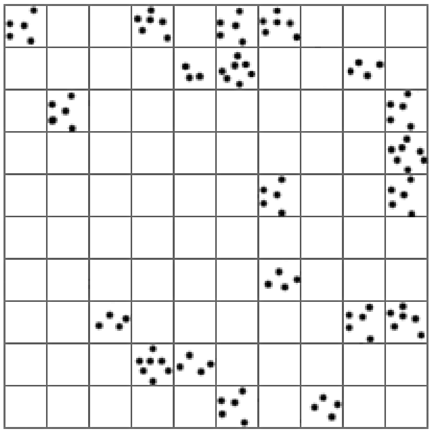
Sample Size 20

All of our estimates thus far have been based on a sample size of 10 randomly selected sub-regions out of 100. In this part, we will examine the effects of changing the sample size to 20 sub-regions.

Here is a sequence of 20 random numbers selected by sampling without replacement:

81 48 66 94 87 60 51 30 92 97 00 41 27 12 38 64 93 79 50 59

Here is the corresponding sample of 20 sub-regions:



As before, we estimate the total number of penguins in the region by finding the mean of our samples, and then multiplying by 100 (the number of regions):

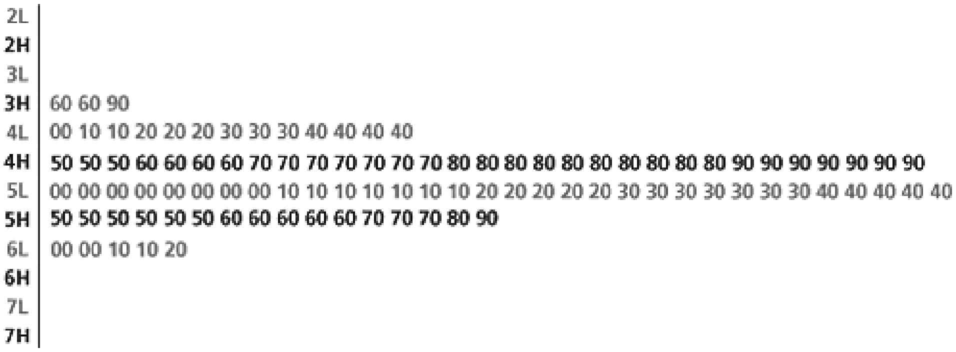
100 x [(5 + 6 + 5 + 6 + 3 + 7 + 4 + 5 + 5 + 7 + 5 + 5 + 4 + 4 + 5 + 6 + 7 + 4 + 5 + 4)/20] = 510

This estimate is very accurate (it is within 10 of the actual number of penguins). Let's now investigate the effect that increasing the sample size has on the accuracy of our estimation procedure.

Comparing Sample Sizes 10 and 20

In order to investigate whether samples of 20 sub-regions are more likely to produce better estimates than samples of 10 sub-regions, you will need to consider repeated sampling results for samples of size 20.

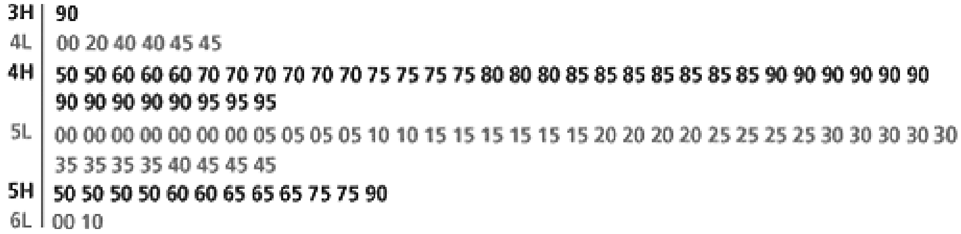
Here is the stem and leaf plot for 100 estimates of sample size 10:



2L	
2H	
3L	
3H	60 60 90
4L	00 10 10 20 20 20 30 30 30 40 40 40 40
4H	50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H	50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L	00 00 10 10 20
6H	
7L	
7H	

Part D, cont'd.

Here is the stem and leaf plot for 100 estimates of sample size 20:



3H	90
4L	00 20 40 40 45 45
4H	50 50 60 60 60 70 70 70 70 70 70 75 75 75 75 80 80 80 85 85 85 85 85 85 85 90 90 90 90 90 90 90 90 90 90 90 95 95 95
5L	00 00 00 00 00 00 00 05 05 05 05 10 10 15 15 15 15 15 15 20 20 20 20 25 25 25 25 30 30 30 30 30 35 35 35 35 40 45 45 45
5H	50 50 50 50 60 60 65 65 65 75 75 90
6L	00 10

Problem D1. Compare the two distributions above. In particular, look at how many estimates for each fall in the interval 450 to < 550 (i.e., the 4H and 5L stems). What does this suggest about the effect of sample size on the accuracy of estimation? **[See Tip D1, page 285]**

Problem D2. Now let's revisit our table of intervals.

a. Use the 100 estimates from samples of size 20 to determine the proportion of estimates in each of the intervals.

		Proportion of Estimates in Interval	
Interval	Interval Length	Sample Size 10	Sample Size 20
350-650	300	100/100	
375-625	250	98/100	
400-600	200	94/100	
425-575	150	84/100	
450-550	100	69/100	
475-525	50	37/100	

b. Compare the proportions within the six intervals for the two different sample sizes. What does this suggest about the effect of sample size on the accuracy of the estimation procedure?

In summary, as the sample size increases, the distribution of the estimates becomes more concentrated. Consequently, a larger sample size generally improves the accuracy of the estimation procedure.

Box Plot Comparisons

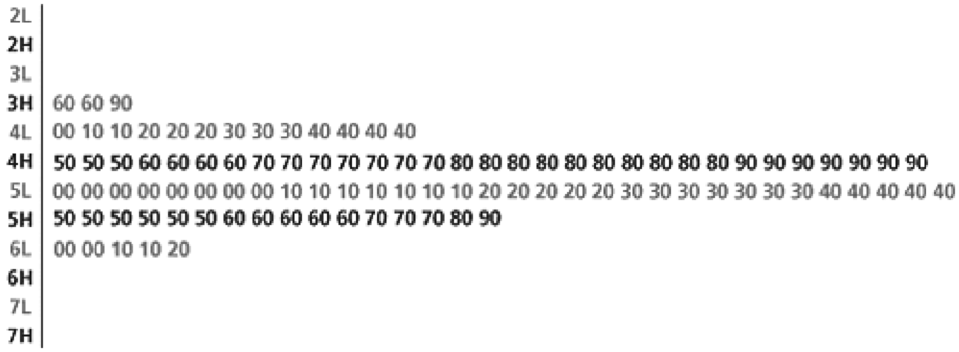
In the previous discussion, you investigated how increasing the sample size does two things:

- Decreases the sample-to-sample variation in the estimates
- Produces a higher proportion of estimates closer to the actual population size

We can also use another familiar method to explore this phenomenon: the Five-Number Summary and box plot.

Part D, cont'd.

Problem D3. Here is the stem and leaf plot for the 100 estimates from samples of size 10:



2L	
2H	
3L	
3H	60 60 90
4L	00 10 10 20 20 20 30 30 30 40 40 40 40
4H	50 50 50 60 60 60 60 70 70 70 70 70 70 70 80 80 80 80 80 80 80 80 80 80 90 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 20 20 30 30 30 30 30 30 30 40 40 40 40 40
5H	50 50 50 50 50 50 60 60 60 60 60 70 70 70 80 90
6L	00 00 10 10 20
6H	
7L	
7H	

Use the stem and leaf plot to determine the Five-Number Summary for these estimates. These questions may help you along:

a. What is the position of the median, and which two values are used to calculate it?

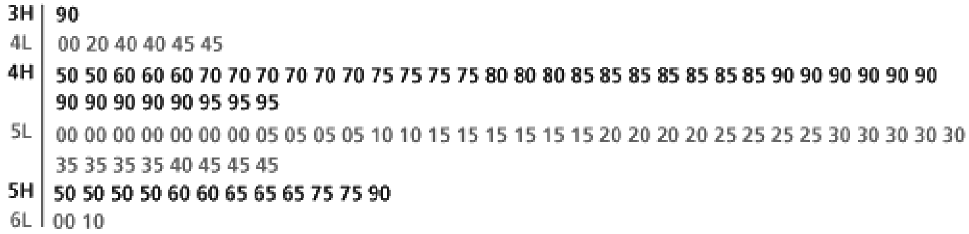
b. If there are 50 values in each half, how are the quartiles calculated?

c. Complete the Five-Number Summary table:

	Sample Size 10
Maximum	
Upper Quartile (Q3)	
Median	
Lower Quartile (Q1)	
Minimum	

Part D, cont'd.

Problem D4. Generate the Five-Number Summary for this stem and leaf plot of the 100 estimates based on samples of size 20:



```
3H | 90
4L | 00 20 40 40 45 45
4H | 50 50 60 60 60 70 70 70 70 70 70 75 75 75 75 80 80 80 85 85 85 85 85 85 85 90 90 90 90 90 90
   | 90 90 90 90 90 95 95 95
5L | 00 00 00 00 00 00 00 05 05 05 05 10 10 15 15 15 15 15 15 20 20 20 20 25 25 25 25 30 30 30 30 30
   | 35 35 35 35 40 45 45 45
5H | 50 50 50 50 60 60 65 65 65 75 75 90
6L | 00 10
```

	Sample Size 20
Maximum	
Upper Quartile	
Median	
Lower Quartile	
Minimum	

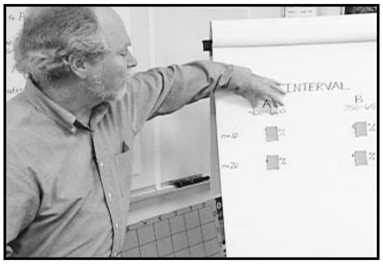
[See Tip D4, page 285]

Problem D5. Create two box plots for the Five-Number Summaries you generated in Problems D3 and D4, placing them side by side on the same scale to make them easier to compare.

Problem D6. What do the box plots suggest about the effect of sample size on the accuracy of the estimates? In particular, how do the box plots illustrate the following:

a. How much the estimates vary from sample to sample

b. How close the estimates are to the actual value of 500



Video Segment (approximate times: 16:02-19:27): You can find this segment on the session video approximately 16 minutes and 2 seconds after the Annenberg/CPB logo.

In this video segment, the participants discuss what percentages of their data fell in particular interval ranges for samples of size 10 and 20. Professor Kader then introduces the Central Limit Theorem to further discuss the connection between probability and statistics. What is the give-and-take between selecting an interval range and sample size when designing a statistical investigation? How would you use this information to plan a statistical investigation? How can you be more precise when taking a sample size? How can you be more accurate?

Homework

These homework problems will take you through the process of statistical estimation using numerical data. You will investigate the quality of the estimation procedure based on different sample sizes.

Gather numerical data for 100 different people. Gather data that has a significant amount of variation; for example, the age of 100 fourth-grade students would not be good numerical data for these purposes. Other than that, the data can be very simple, such as height in centimeters or age in years.

These 100 people will represent your overall population for the seven homework problems. Your goal is to investigate your sample mean as an estimate of your population mean and to explore the accuracy of your estimates.

Solutions are not provided for these homework problems, since answers will vary according to the data you have gathered.

Problem H1. Find the average (mean) value for all 100 people. Computer software may make this process easier.

Problem H2.

a. Use the random process you developed in Part B to generate a random sample of 10 people from the population. Sample without replacement.

b. Calculate your sample mean and compare it to your population mean.

Problem H3. Repeat the process of generating a random sample of size 10 and calculating the sample mean at least nine more times. Computer software may make this process easier and allow you to take more samples.

Problem H4. For the set of sample means from your random samples of size 10, determine the Five-Number Summary.

Problem H5. You will now generate a new set of estimates, this time based on random samples of size five.

a. Do you expect these estimates to have more or less variation than the estimates from samples of size 10?

b. Generate several random samples of size five. Use the same random process and generate the same number of random samples (at least 10) that you did in Problem H3. Determine the mean for each sample.

c. Determine the Five-Number Summary for these estimates.

Problem H6. Compare the estimates from samples of size 10 with the estimates from samples of size five. Draw comparative box plots. Where is the population mean in relation to each box plot? Which sample size produces estimates with less variation? Is this what you predicted?

Take It Further

Problem H7.

a. Generate the same number of random samples that you've been using, but this time of size 20.

b. Use computer software to compute the average and standard deviation of all the sample means of size five and all the sample means of size 20. Compare the results: Which set has the smaller standard deviation? How much smaller is it?

[See Tip H7, page 285]

Suggested Readings

These readings are available as downloadable PDF files on the *Data Analysis, Statistics, and Probability* Web site. Go to:

www.learner.org/learningmath

Perry, Mike and Kader, Gary (February, 1998). Counting Penguins. *Mathematics Teacher, 91* (2), 110-116.

Woolley, Thomas (Autumn, 1998). A Note on Illustrating the Central Limit Theorem. *Teaching Statistics, 20* (3), 89-90.

Tips

Part A: Random Samples

Tip A1. In Samples B and C, you will need to use the sample results to make a "best guess" for the number of penguins in the entire region. What methods have you learned for coming up with such a guess?

Part B: Selecting the Sample

Tip B1. To select 10 sub-regions from the 100 total sub-regions in a "fair" way requires that each of the 100 subregions has the same chance of being selected. You can accomplish this with random selection. How might you select 10 sub-regions in a random fashion?

Tip B2. You may wish to use a random-number-generating device, such as a calculator, a 10-sided die, or computer software, to generate the random numbers.

Part C: Investigating Variation in Estimates

Tip C1(b). The interval should include most of the data in the stem and leaf plot. For example, "between 200 and 400" would be a very poor interval of values.

Tip C2. Since there are 100 samples, the percentage will be the actual number of estimates found.

Part D: The Effect of Sample Size

Tip D1. Which distribution has more estimates "closer" to the actual answer of 500?

Tip D4. Since the number of estimates is the same as Problem D3's, the quartiles and median will be in the same positions. Count the values in increasing order to find them.

Homework

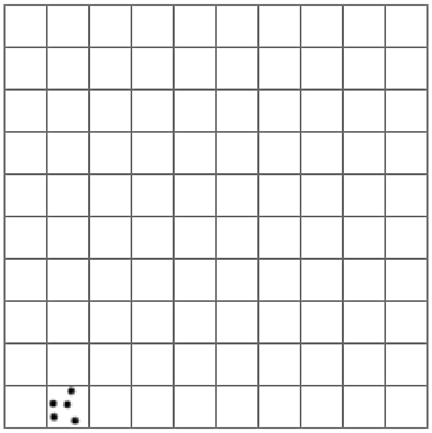
Tip H7. You may need a large number of samples to see a pattern here.

Solutions

Part A: Random Samples

Problem A1. Sample A:

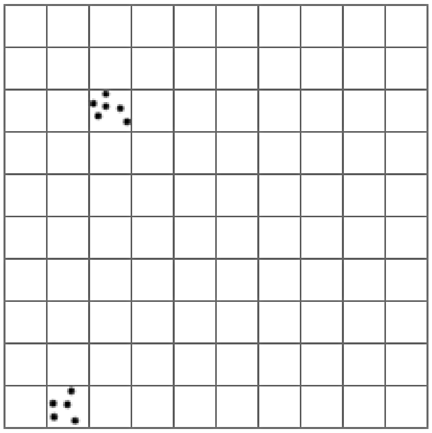
This sample of one sub-region shows five penguins:



Based on this limited information, you might guess that each and every sub-region contains five penguins. Since there are 100 sub-regions, your estimate of the total number of penguins would be 100 x 5 = 500.

Sample B:

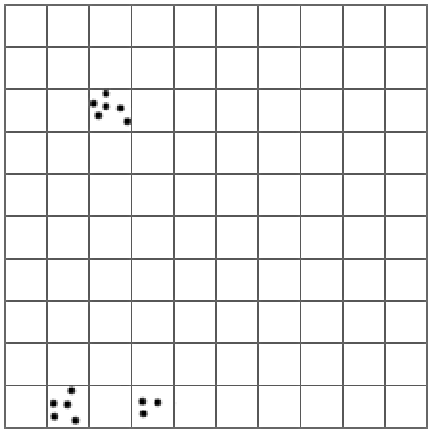
This sample of two sub-regions contains 5 + 6 = 11 penguins, or an average of 11/2 penguins per sub-region:



Based on this limited information, you might guess that the average for all 100 sub-regions is 11/2 penguins. Since there are 100 sub-regions, your estimate of the total number of penguins would be 100 x (11/2) = 550.

Sample C:

This sample of three sub-regions contains 5 + 6 + 3 = 14 penguins, or an average of 14/3 penguins per sub-region:



Based on this limited information, you might guess that the average for all 100 sub-regions is 14/3 penguins. Since there are 100 sub-regions, your estimate of the total number of penguins would be 100 x (14/3) = 1,400/3, or, to the nearest penguin, 467 penguins.

Solutions, cont'd.

Problem A1, cont'd.

Here is the completed table:

Sample	Photo 1	Photo 2	Photo 3	Estimate of Total
A	5	N/A	N/A	500
B	5	6	N/A	550
C	5	6	3	467

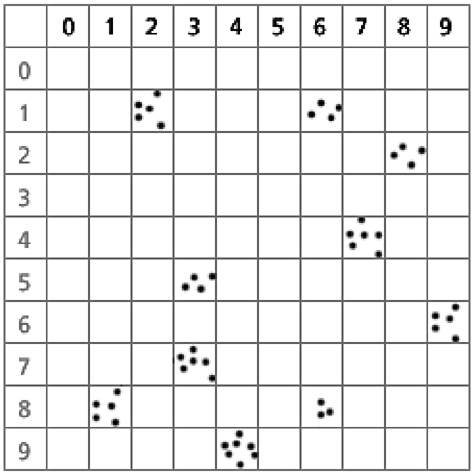
Problem A2. First, find the average number of penguins in each sub-region of the sample. The total number of penguins is 5 + 6 + 6 + 7 + 5 + 2 + 1 + 5 + 5 + 3 = 45. Since there are 10 sub-regions in the sample, the average number of penguins is 45/10. Therefore, a good estimate for the total number of penguins is 100 x 45/10 = 450 penguins.

Part B: Selecting the Sample

Problem B1. Answers will vary, as there are many possible ways to do this. One possibility is to take the 100 pictures of the sub-regions, shuffle them thoroughly, then look at the first 10. Another is to assign each sub-region to a number from 00 to 99, and use the last two digits of the daily lottery number for each of the last 10 days. A commonly used method for assigning regions to numbers is to use a random-number-generating device, such as a calculator, a die, or computer software.

Problem B2. With a calculator, the first two decimal digits of the random number will range from 00 to 99, and each of the 100 values is equally likely. If a number appears more than once, it is rejected so that 10 different sub-regions will be selected. Another idea is to use a 10-sided die or spinner and to generate two random digits by two tosses or spins (and get your 10 random numbers by 20 tosses or spins).

Problem B3. Answers will vary, depending on which region you selected in Problem B2. As an example, the random sequence (96, 74, 61, 21, 49, 37, 82, 35, 18, 68) determines this sample of 10 sub-regions:



The estimate of the total number of penguins is:

100 x [(5 + 4 + 4 + 6 + 4 + 5 + 6 +5 + 3 + 7)/10] = 100 x (49/10) = 490

Problem B4. While it is possible for the two estimates to be equal, it is pretty unlikely, due to the variation in the individual sub-regions. If the number of sub-regions in the sample increases to 20, the variation in the estimates should be reduced. The estimates should be closer to the actual value, but it is no more likely that they will be equal.

Problem B5. Answers will vary. To determine how many penguins there are in the region, you might calculate the mean or median of the set of five estimates.

Part C: Investigating Variation in Estimates

Problem C1.

a. A good estimate might be the median of the 100 estimates, which is in position (100 + 1)/2 = 50.5. This means that the median is the average of the 50th and 51st values in the ordered list. Both the 50th and 51st values are 500, so 500 penguins is a good estimate, based on the median.

b. It seems very likely that the actual number is between 360 and 620, since all 100 estimates fall in this range. A tighter range is 450 to 550, which includes 69 of the 100 estimates.

Problem C2.

a. The best estimate is 500, which is exactly right. Our sampling found this estimate eight of 100 times.

b. The six worst estimates are 360, 360, 390, 610, 610, and 620. These are the only six estimates that are more than 100 penguins away from the actual value.

c. These are the estimates between 450 and 550 (inclusive); 69% (69/100) of the estimates are within this range.

d. These are the estimates between 400 and 600 (inclusive); since only six estimates are more than 100 penguins away, 94% (94/100) of the estimates are within this range.

Problem C3.

a. These are the estimates between 425 and 575 penguins (inclusive); the proportion is 84/100, since 84 of the 100 estimates are within this range.

b. The proportion is 16/100 (obtained as 1 - 84/100).

Problem C4. Here is the completed table:

Interval (Inclusive)	Proportion of Estimates in Interval	Proportion of Estimates Not in Interval
350-650	100/100	0/100
375-625	98/100	2/100
400-600	94/100	6/100
425-575	84/100	16/100
450-550	69/100	31/100
475-525	37/100	63/100

Solutions, cont'd.

Problem C5.

a. When the proportion of estimates in an interval is high, it is a strong suggestion that the actual population value lies somewhere in that range.

b. A small interval gives greater precision to the estimates. If we can say that the actual value lies between 475 and 525, it is more meaningful than saying that the actual value lies, say, between 400 and 600.

c. As the interval range decreases, the proportion of estimates in that interval decreases. Thus, there is an important tradeoff: A wide interval will contain more estimates but will be less meaningful, whereas a small interval will be more meaningful but will contain fewer estimates.

Problem C6.

a. The expected probability is 0.84, or 84%, since 84 of the 100 estimates fall in this interval.

b. The probability is 37%, since 37 of the estimates fall in the smallest interval (475 to 525).

c. It is very likely—94% of the estimates fall in the interval within 100 penguins of the actual total (400 to 600).

Part D: The Effect of Sample Size

Problem D1. There are more estimates from the distribution for sample size 20 that fall in the 4H and 5L stems (i.e., in the range 450-549). This suggests that the estimates from 20 sub-regions are more accurate.

Problem D2.

a. Here is the completed table:

		Proportion of Estimates in Interval	
Interval	**Interval Length**	**Sample Size 10**	**Sample Size 20**
350-650	300	100/100	100/100
375-625	250	98/100	100/100
400-600	200	94/100	98/100
425-575	150	84/100	94/100
450-550	100	69/100	83/100
475-525	50	37/100	55/100

b. Each interval of the samples of 20 sub-regions contains a higher proportion of estimates. For instance, the interval 450-550 contains 83/100 samples of size 20, compared to 69/100 samples of size 10. A higher proportion of the estimates falls within 50 penguins of the actual population size (500) when samples of size 20 were used. This suggests that the increased sample size has a significant effect on the accuracy of the estimates.

Solutions, cont'd.

Problem D3.

a. The median is in position (100 + 1) / 2 = 50.5, so it is the average of the 50th and 51st values in the ordered list. Each of these values is 500.

b. The quartiles will be at position (50 + 1) / 2 = 25.5, so they are the average of the 25th and 26th values in their respective halves.

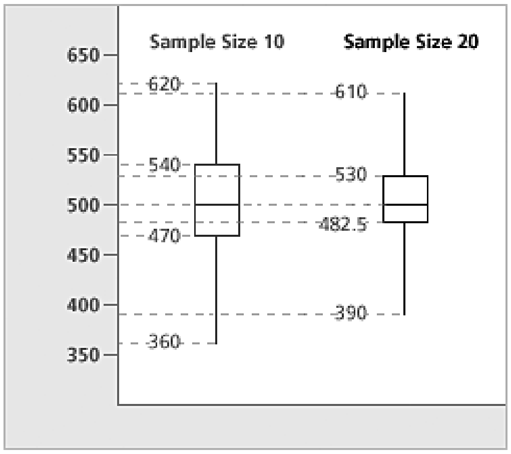
c. Here is the completed table:

	Sample Size 10
Maximum	620
Upper Quartile (Q3)	540
Median	500
Lower Quartile (Q1)	470
Minimum	360

Problem D4. Here is the completed table:

	Sample Size 20
Maximum	610
Upper Quartile (Q3)	530
Median	500
Lower Quartile (Q1)	482.5
Minimum	390

Problem D5.



Problem D6.

a. The sample-to-sample variation goes down as the sample size increases. This is exhibited by the shrinking box portion of the graphs.

b. The estimates are closer to the actual value as the sample size increases. Both the range and the interquartile range decrease significantly from the estimates using sample size 10 and sample size 20.