An Exercise in Sampling: Rolling Down the River

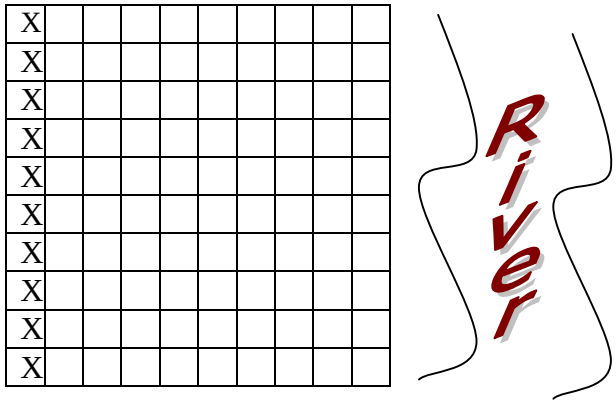
Name: _____ Date: _____

A farmer has just cleared a new field for corn. It is a unique plot of land in that a river runs along one side. The corn looks good in some areas of the field but not others. The farmer is not sure that harvesting the field is worth the expense. He has decided to harvest 10 plots and use this information to estimate the total yield. Based on this estimate, he will decide whether to harvest the remaining plots.

Part I.

A. Method Number One: Convenience Sample

The farmer began by choosing 10 plots that would be easy to harvest. They are marked on the grid below:

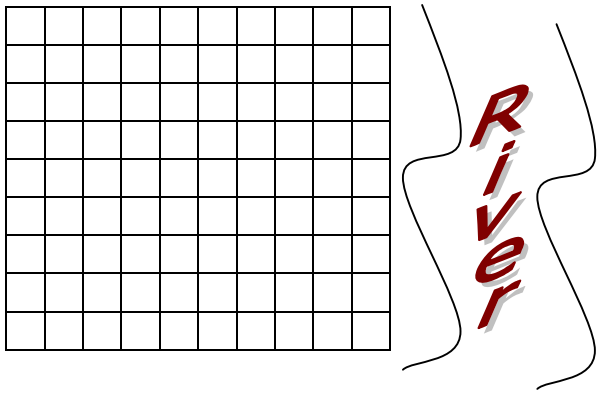


Since then, the farmer has had second thoughts about this selection and has decided to come to you (knowing that you are an AP statistics student, somewhat knowledgeable, but far cheaper than a professional statistician) to determine the approximate yield of the field.

You will still be allowed to pick 10 plots to harvest early. Your job is to determine which of the following methods is the best one to use – and to decide if this is an improvement over the farmer's original plan.

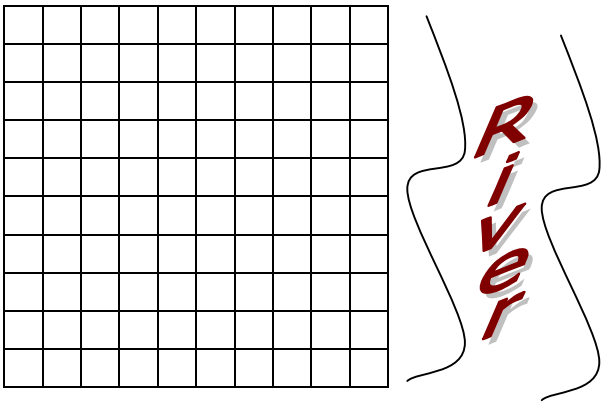
B. Method Number Two: Simple Random Sample

Use your calculator or a random number table to choose 10 plots to harvest. Mark them on the grid below, and describe your method of selection.



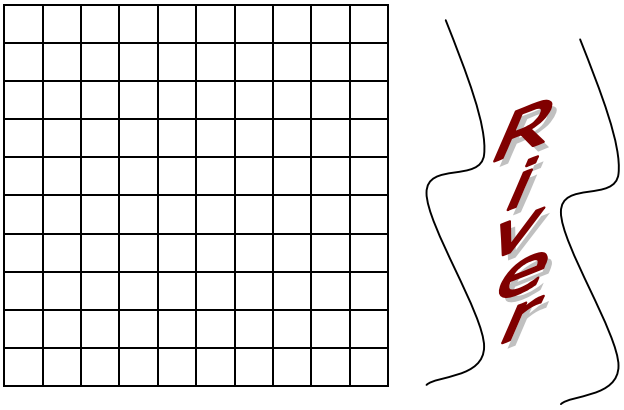
C. Method Number Three: Stratified Sample

Consider the field as grouped in vertical columns (called strata). Using your calculator or a random number table, randomly choose one plot from each vertical column and mark these plots on the grid.



D. Method Number Four: Stratified Sample

Consider the field as grouped in horizontal rows (also called strata). Using your calculator or a random number table, randomly choose one plot from each horizontal row and mark these plots on the grid.



OK, the crop is ready. Below is a grid with the yield per plot. Estimate the average yield per plot based on each of the four sampling techniques.

6	17	20	38	47	55	69	76	82	97
7	14	23	34	43	56	63	75	81	92
2	14	28	30	50	50	62	80	85	96
9	15	27	34	43	51	65	72	88	91
4	15	28	32	44	50	64	76	82	97
5	16	27	31	48	59	69	72	86	99
5	18	28	34	50	60	62	75	90	90
8	15	20	38	40	54	62	77	88	93
7	17	29	39	44	53	61	77	80	90
7	19	22	33	49	53	67	76	86	97

Sampling Method	Mean yield per plot	Estimate of total yield
Convenience Sample (farmer's)		
Simple Random Sample		
Vertical Strata		
Horizontal Strata		

Observations:

1) You have looked at four different methods of choosing plots. Is there a reason, other than convenience, to choose one method over another?

2) How did your estimates vary according to the different sampling methods you used?

3) Compare your results to someone else in the class. Were your results similar?

4) Pool the results of all students for the mean yields from the simple random samples and make a class boxplot. Repeat for means from vertical strata and from horizontal strata. Compare the class boxplots for each sampling method. What do you see?

5) Which sampling method should you use? Why do you think this method is best?

6) What was the actual yield of the farmer's field? How did the boxplots relate to this actual value?

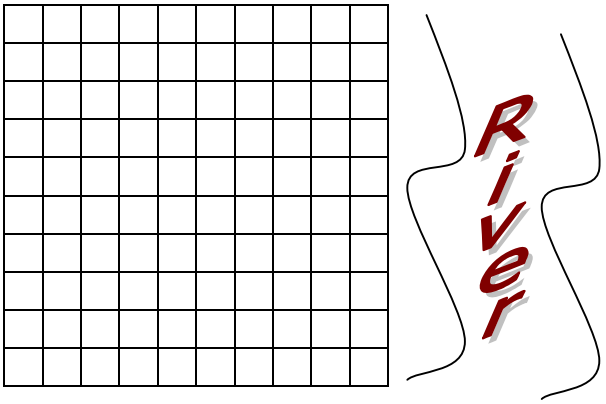
Part II:

The farmer was very impressed with the results of your study and seeks to improve the yield of that part of the field the following year. Believing that irrigation is the answer, a new system was installed. The following year's yield was:

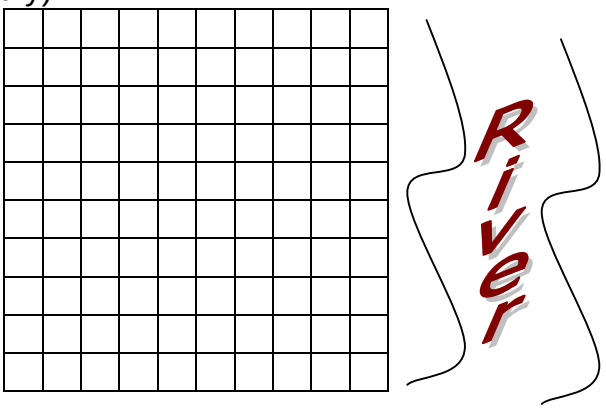
79	81	95	69	65	59	88	65	66	91
80	75	88	80	82	66	76	99	62	61
97	50	92	92	91	84	75	85	63	89
99	71	55	75	65	66	66	86	96	50
57	95	51	79	98	71	70	86	89	76
57	53	90	71	50	76	56	91	85	64
69	95	98	90	93	97	79	95	73	90
58	99	75	51	67	81	55	63	89	74
98	62	73	54	50	76	91	50	90	55
91	59	69	59	71	72	85	85	86	97

Redo your sampling using a SRS, vertical stratification, and horizontal stratification. Be certain to mark on the grids the plots you choose.

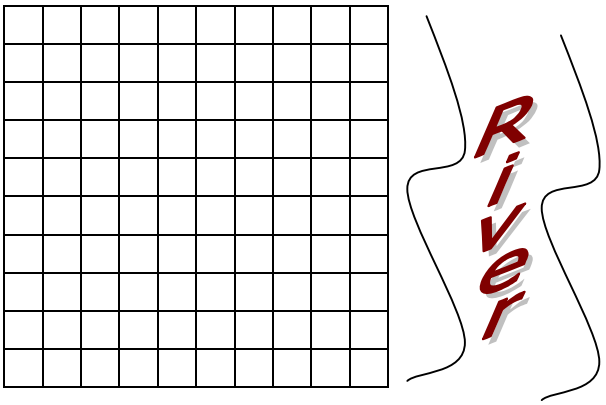
A. Simple Random Sample:



B. Stratified Sample (vertically):



C. Stratified Sample (horizontally):



Sampling Method	Mean yield per plot	Estimate of total yield
Simple Random Sample		
Vertical Strata		
Horizontal Strata		

Observations:

1) Compare the class boxplots of the sample means obtained from the SRS and the two methods of stratified sampling.

2) Based on the results of both activities, under what conditions is it more useful to use stratified sampling?

3) Based on the results of both activities, under what conditions is it more useful to use a simple random sample?

Teacher's Notes for Rolling Down the River

The purpose of this exercise is to allow students to see the effects of different methods of sampling in different situations.

Part I.

A. Convenience Sample

This is rarely a good choice. Although it is attractive to the farmer to harvest the plots as easily as possible, it often leads to large bias in the result.

B. Simple Random Sample

With simple random sampling, all possible sets of 10 plots have an equal chance of being selected. By using this impartial selection method, higher yield plots should be balanced out by lower yield plots. However, it may be the case, since all possible combinations are possible, that all of the selected plots have a high yield or that all of the selected plots have a low yield. Thus there is large variability in the sample statistic.

C/D. Stratification

When there is some factor that can influence or affect the response, (in this case the river has an effect on the yield), then using a stratified sample should reduce some of the variability in the means of repeated samples. However, it is necessary to choose the strata correctly. Strata should be constructed so that within the strata the data are very similar (homogeneous) while the individual strata contain sets of data that are as different as possible (heterogeneous). The farmer should be consulted as to the direction of the strata. His experience would determine the best approach.

Note: The data were purposely set up so that the effects of proper stratification would be startling. This does not mean to suggest that a crop grown near a river would necessarily result in such a large difference in yields.

Observations / Answer Key:

1) One needs to choose a method that will give the best estimate of the yield. This can be affected by factors that cannot be controlled: e.g. the placement of the river. That's why one shouldn't choose the ten plots chosen by the farmer.

2) The student will see that the farmer's sample yields a very low estimate compared to the other methods used.

3) Comparing results with a peer helps the student verify that the sampling was done correctly. This does not mean the students will have the same sample, but each student should use the same process of drawing a sample for a given method. Some methods will produce highly variable results while others are much more consistent.

4) The variability of the means of the sample yields, as shown by the length of the boxplot and the width of the middle 50%, will reduce drastically once the student has stratified appropriately. Thus the strata that are effective are the vertical ones, in which the values in each stratum are similar. This stratification reduces the variation in the sample means since the values chosen for a particular stratum vary little from sample to sample relative to the variability in the population.

5) Vertical stratification should be used since the sample would then include higher yielding plots as well as lower yielding ones.

6) The actual yield is 5004. The class boxplot for the means resulting from the vertical stratification should be centered near 5004/100 or about 50.

Part II.

Observations / Answer Key:

1) Since the river effect has been cancelled out by the irrigation process, there is no discernable pattern in the yield (in effect, the data are randomly distributed). Therefore, there should be no improvement in using a stratified random sample over a SRS. The boxplots should be centered near the total yield divided by 100 (7603/100 or approximately 76).

2) It is more useful to stratify when one suspects that there is some outside factor affecting the response variable.

3) It is more useful to use a SRS when there is no reason to stratify; that is, when there is no reason to expect that an outside factor is affecting the response variable. It certainly is easier and is often less expensive to use a SRS.