2.3 – Sampling Principles
MDM4U
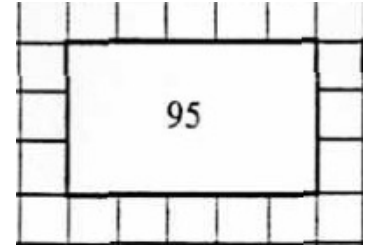*Jensen*

Part 1: Random Rectangles Activity

1. a. Guess the average area of all rectangles on the page: **(guess)** _____

    b. Choose six rectangles (before you calculate any areas) that you think represent the entire
        population of rectangles well.

        6 rectangles – subjective – "rectangle **expert**":

rectangle number	area
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

                                    average: _____

2. a. After setting a new seed value on your calculator, use the randint function to choose six   random
        rectangles for you.

   6 rectangles – **random**:

rectangle number	b. area
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

                                    average: _____

3.   a.   mean of sample averages:

              guesses   _____

              subjective (expert)   _____

              random   _____

       c.   actual area of 100 rectangles (population):   _____

Wrap-up (what have you learned?):

The design of a study is <u>biased</u> if it systematically favors certain outcomes. The design of a study shows bias if it consistently over or under estimates the value you want to know. Random sampling is necessary to get a representative sample.
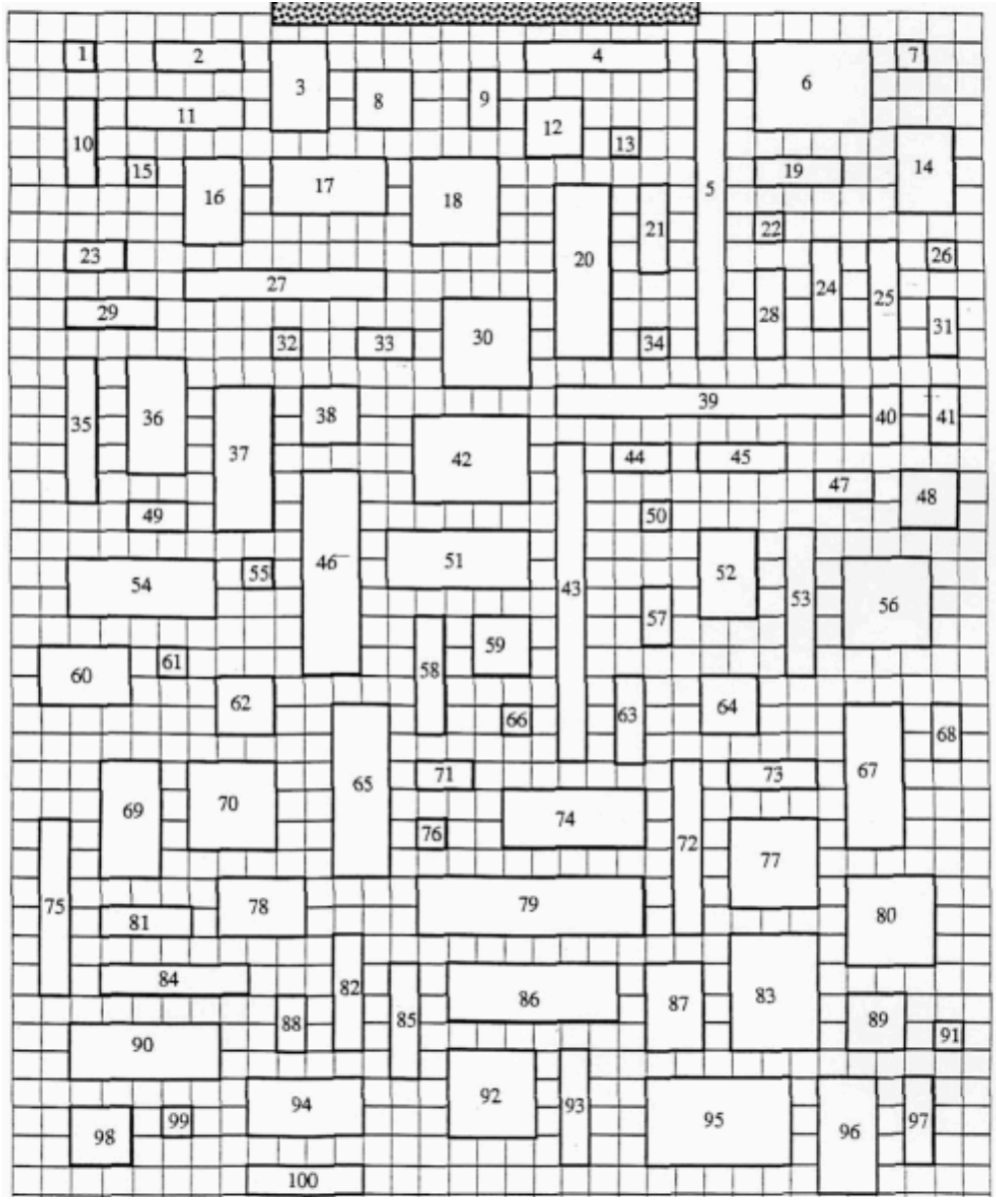
Part 2: Random Sampling Methods

1. Simple Random Sampling

A sample is a simple random sample if it is selected so that:

- each member of the population is equally likely to be chosen and the members of the sample are chosen independently of one other;

OR

- every set of n units has an equal chance to be the sample actually selected.

Example: Put names in hat and draw until have desired sample size; more commonly, number names and use random number generator or other source of random numbers to select sample. Notice that some type of unbiased method must be used; haphazard ≠ random.

2. Systematic Random Sampling

A sample is a systematic random sample if you randomly choose some starting point; then select every n^{th} element in the population, where n is the sampling interval. This guarantees that the sample is taken from throughout the population but it requires an ordered list of everyone in the population.

Example: If we wanted to get a systematic random sample of 10% of the students from King's which has approximately 600 students...

- Calculate number of students required for sample: $600×0.10 = 60$
- Calculate the sampling interval: $sampling\ interval = \frac{population\ size}{sample\ size} = \frac{600}{60} = 10$
- Choose a random starting point using a random number generator
- Include every 10th student from the randomly chosen starting point in your sample

3. Stratified Random Sampling

When using a stratified random sample, the population is divided into groups called strata (e.g. age, geographical areas, grade, etc.)

A simple random sample of the members of each stratum is then taken. The size of the sample for each stratum is proportionate to the stratum's size (you must survey the same percentage of people from each stratum).

Example: If we want a stratified random sample of 10% of the 600 King's students, we can divide the population in to four groups based on grade (9, 10, 11, 12) and then take a simple random sample of 10% of the students in each grade.

4. Cluster Random Sampling

When using a cluster random sampling method, divide the population into groups or clusters; randomly select a few of those groups and then sample all members from the selected groups.

**Example:** Randomly select 5 block C classes—survey all students in each class selected.

5. Multi-Stage Random Sampling

When using multi-stage random sampling, the population is organized in to groups, a simple random sample of groups is chosen, and then a simple random sample of people within the chosen groups is taken.

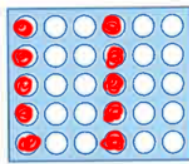Example: Randomly select 5 block C classes—survey a random sample of 10% of the students in each class selected.

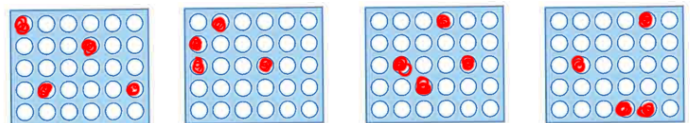## Review of Different Random Sampling Techniques:

Simple Randon
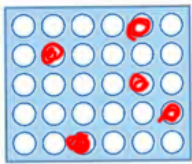
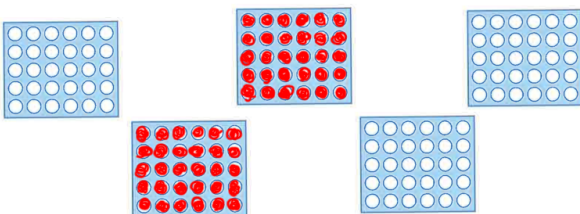- all selections are equally likely

Systematic Random

- random starting poi choose individuals at interval (every *n*th pe
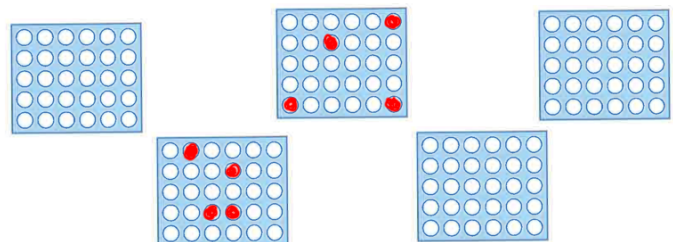
Stratified Random

- Divide population into groups then survey an equal percentage of each group.

Cluster Random

- Divide the population into groups. Choose a random sample of groups and then survey **every member** of the groups chosen.

Multi-Stage Random

- Divide the population into groups. Choose a random sample of groups and then choose a random sample of members of the chosen groups.

1. Convenience sampling

The easiest way to obtain a sample is to choose it without any random mechanism (also called haphazard sampling). Choosing individuals from the population who are easy to reach results in a convenience sample. Convenience sampling often produces unrepresentative data.

Example: Suppose we want to know how long students at a large high school spent doing homework last week. We might go to the school library and ask the first 30 students we see about their homework time.

2. Voluntary Response Sampling

A voluntary response sample consists of people who choose themselves by responding to a general invitation. Voluntary response samples attract people who feel strongly about an issue, and who often share the same opinion. This leads to bias.

Example: A radio host invites listeners to call in to give opinions on a new band.

Part 4: River Activity

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.

0 1 2 3 4 5 6 7 8 9

N
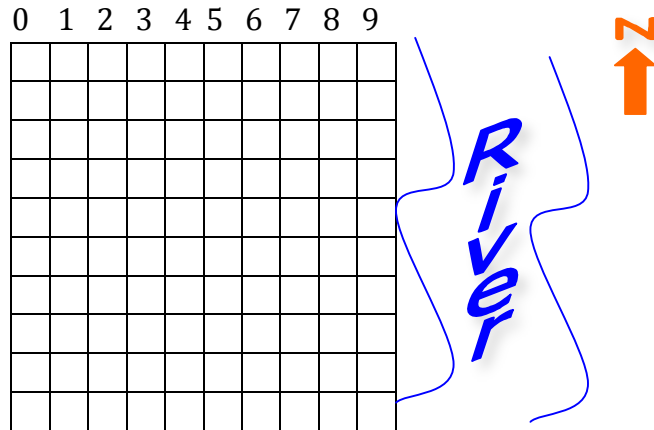
River

C. Method Number Three: Stratified Sample

You and the farmer think the river might have a strong influence on corn production so you decide to consider the field as grouped in vertical columns (called strata—remember you can only stratify data your sample when you think a factor will have a strong influence on the outcome.). Using your random number table, randomly choose one plot from each vertical column and mark on the grid. (Label your columns A through J, rows 0 through 9.)

A B C D E F G H I J

N

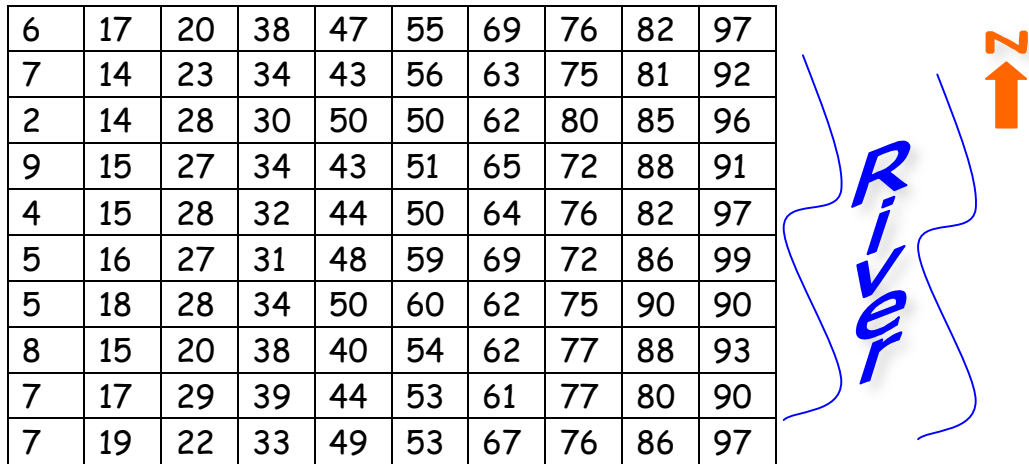*River*
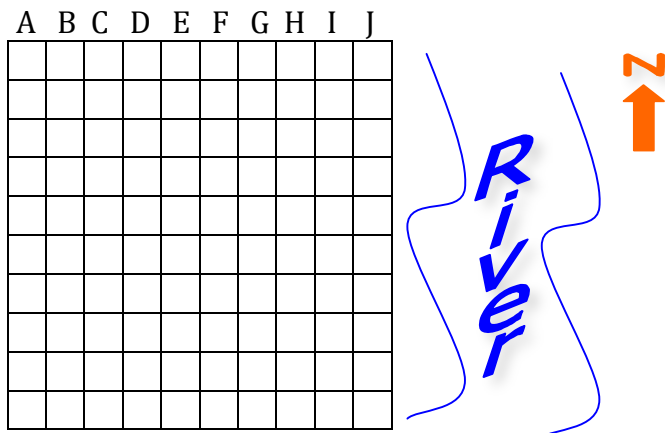
D. Method Number Four: Stratified Sample

You and the farmer rethink the plan and decide that direction (north—south) may have a strong influence on corn production. You decide to consider the field as grouped in horizontal rows (also called strata). Using your random number table, randomly choose one plot from each horizontal row and mark them on the grid. (Label your rows A through J, columns 0 through 9.)

0 1 2 3 4 5 6 7 8 9

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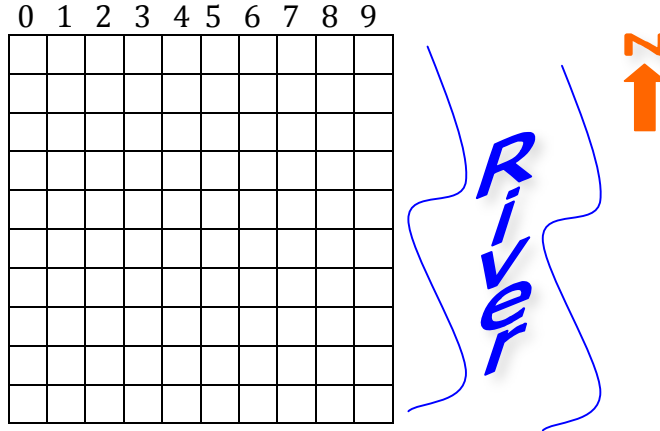
OK, the crop is ready! Below is a grid with the yield per plot. Estimate the average yield per plot based on each of the four sampling techniques.

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

River N ↑

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?

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

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

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

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) When we compare the class boxplots for each sampling method. What do you see?

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

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

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

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.