Lecture 1

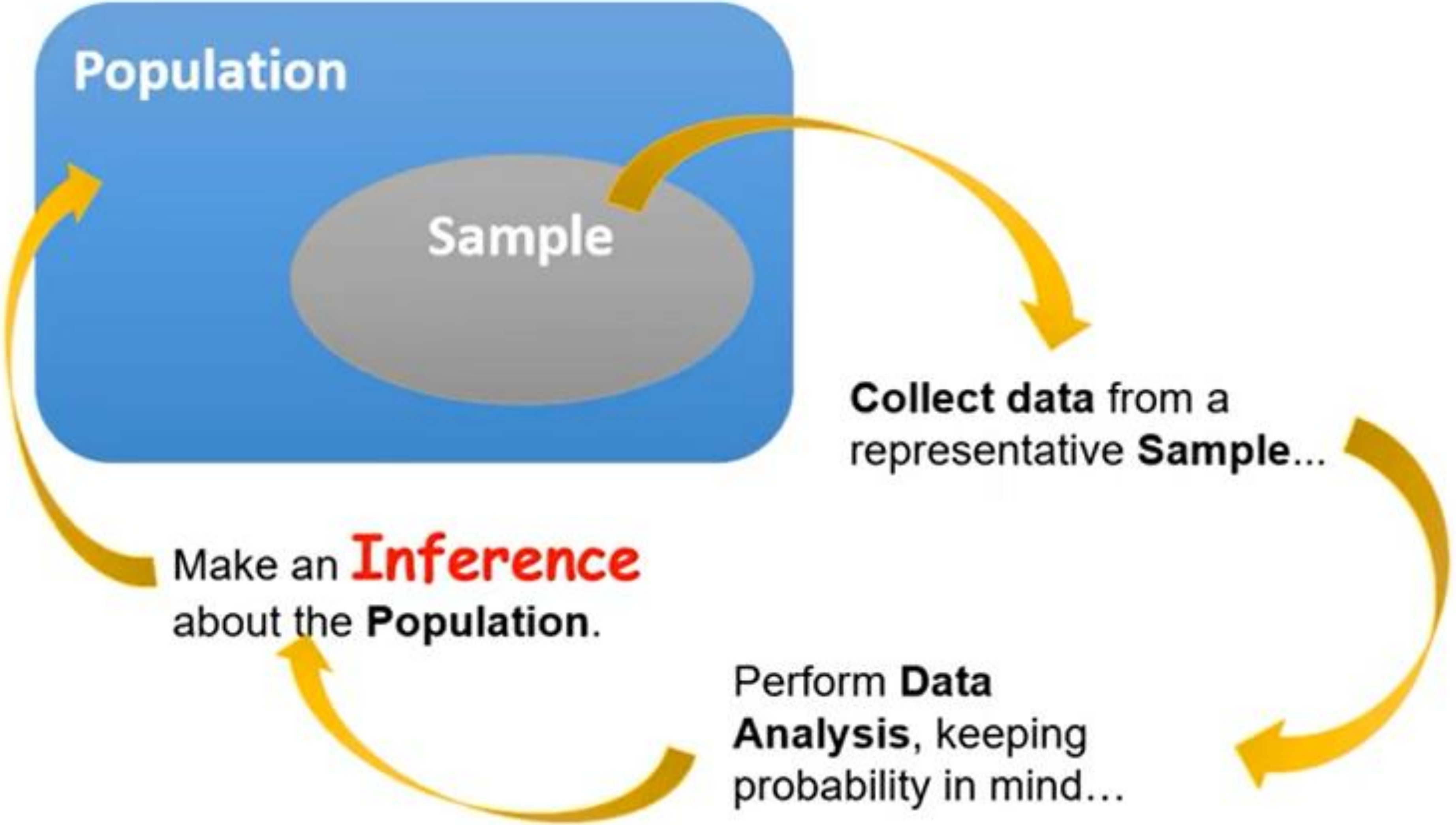
Sampling

You can hardly go a day without hearing the results of a statistical study.

- Mary and James are the most popular names.
- The smallest number of children are born on Saturdays.
- McDonald's sells 75 burgers per second.
- The global rate for washing hands after using the toilet is under 20%

HOW THE DATA WERE PRODUCED

The procedure to realize Statistics:



Collect data from a representative **Sample**...

Perform **Data Analysis**, keeping probability in mind…

Make an **Inference** about the **Population**.

Definition:

Population:

The population in a statistical study is the entire group of individuals we want information about.

How to know the information?

Conduct a census!

A census collects data from every individual in the population.

☹ Census: take too much time and cost too much money ☹

Sampling!!!!!!

Definition:

Population:

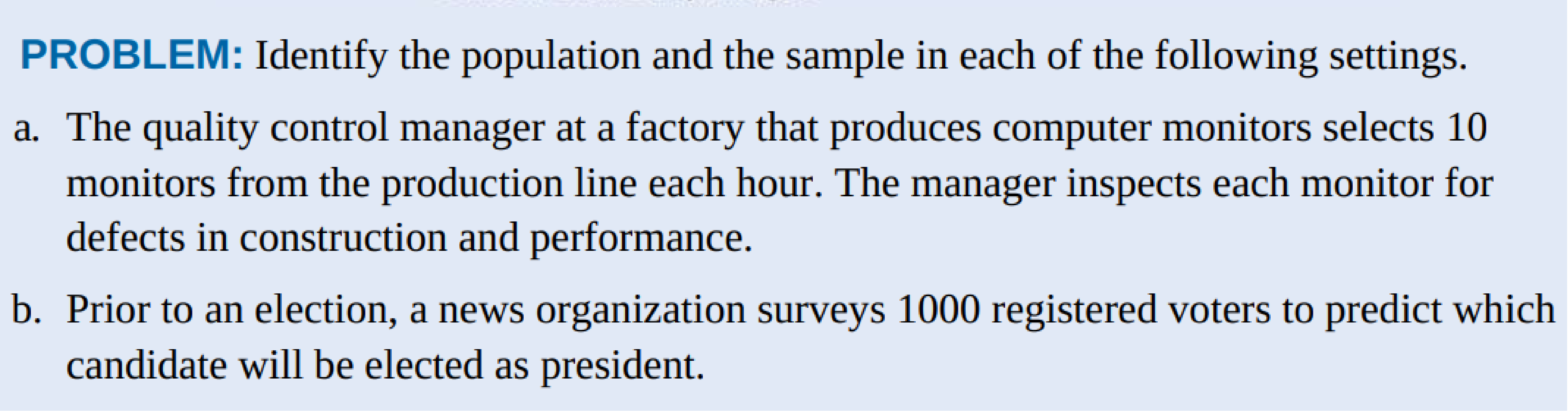
The population in a statistical study is the entire group of individuals we want information about.

How to know the information?

Sampling: Choosing a subset of individuals in the population from which we collect data.

Sample: a subset of individuals in the population from which we collect data

Sampling!!!!!!



PROBLEM: Identify the population and the sample in each of the following settings.

a. The quality control manager at a factory that produces computer monitors selects 10 monitors from the production line each hour. The manager inspects each monitor for defects in construction and performance.

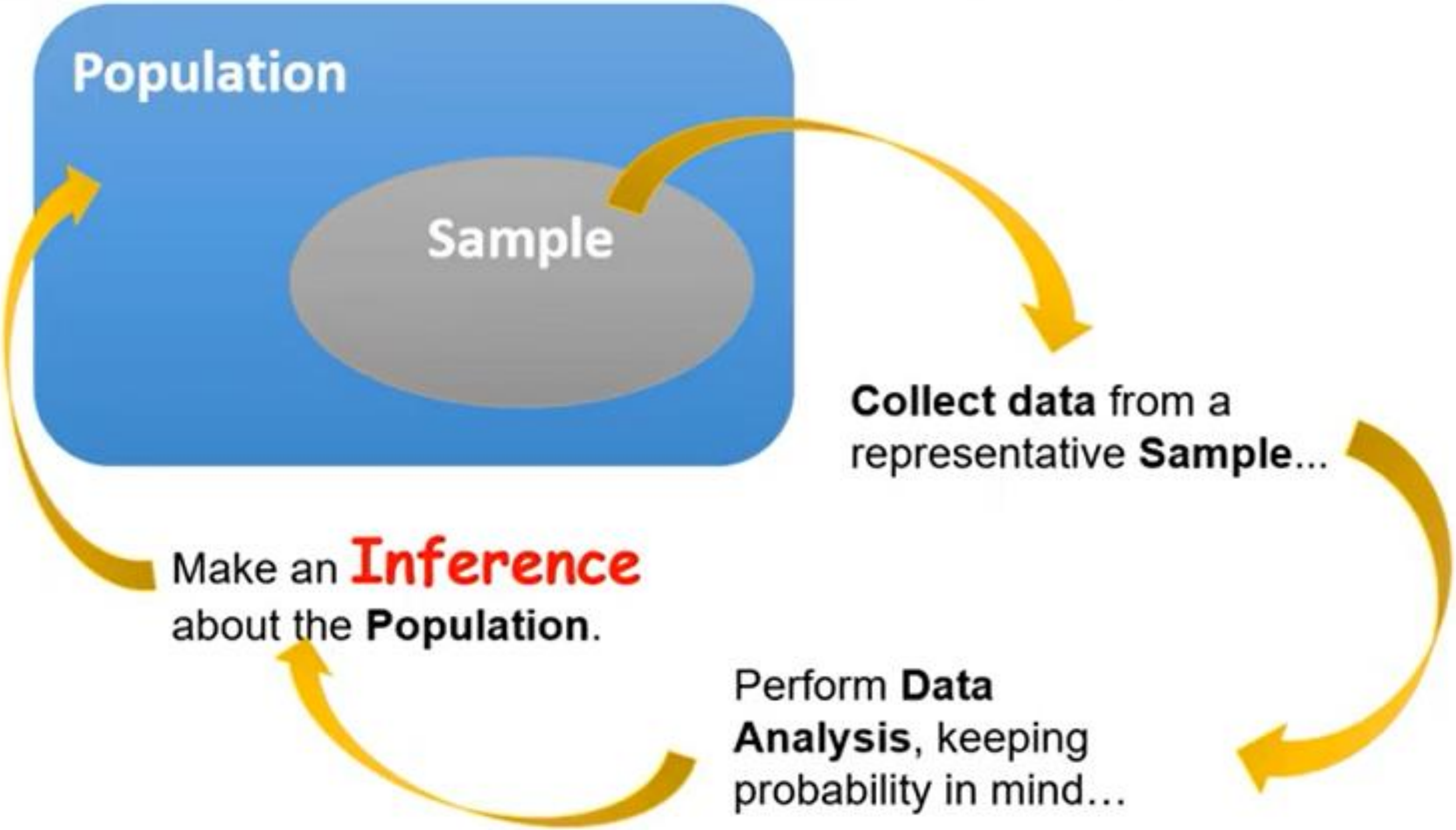
b. Prior to an election, a news organization surveys 1000 registered voters to predict which candidate will be elected as president.

a. The population is all the monitors produced in this factory that hour.

The sample is the 10 monitors selected from the production line.

b. The population is all registered voters.

The sample is the 1000 registered voters surveyed.


Population
Sample
Collect data from a representative **Sample**...
Perform **Data Analysis**, keeping probability in mind…
Make an Inference about the **Population**.

What we want first is the **representative sample**

To get the representative sample, should control: **various kinds of bias!**

How

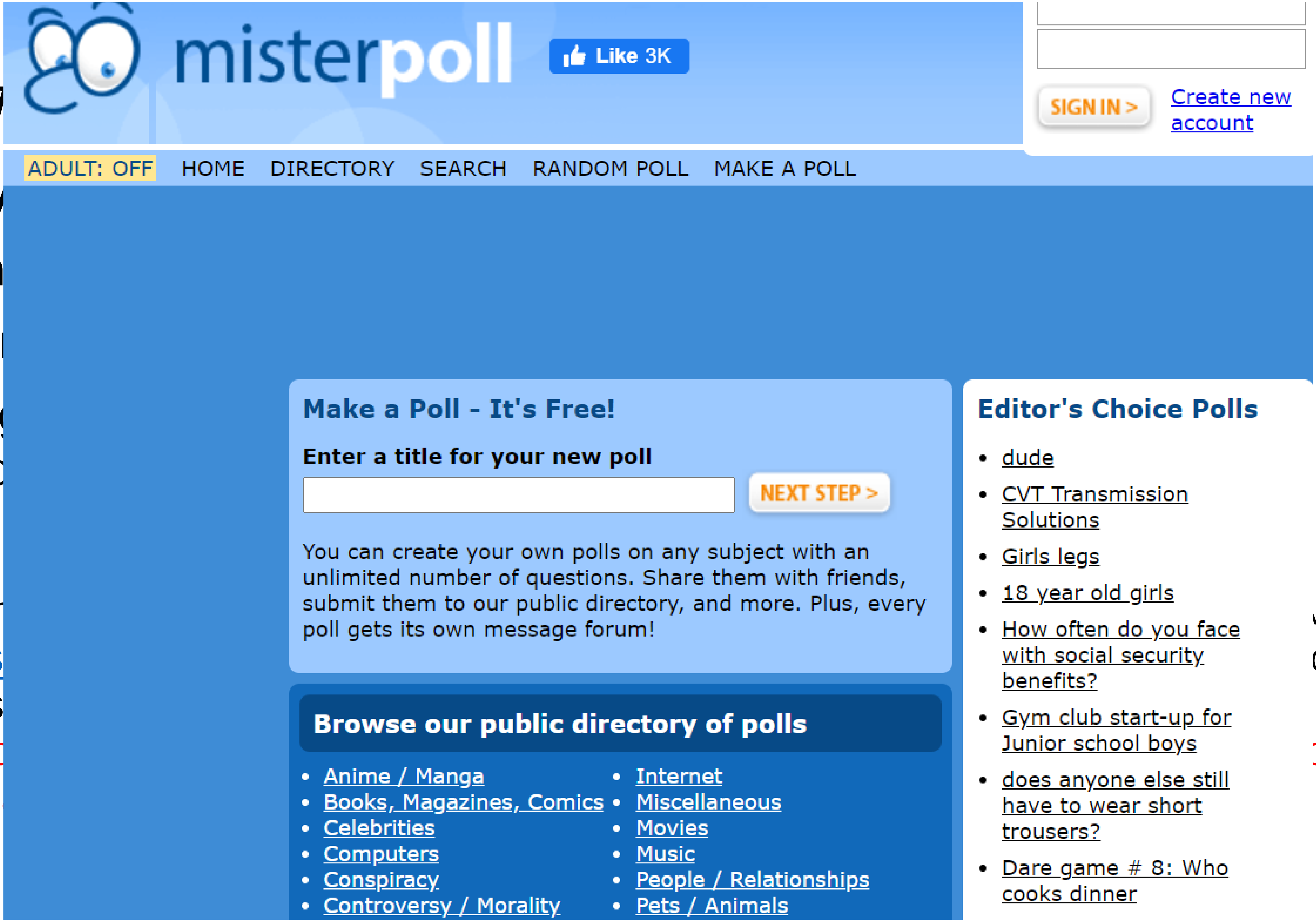
- ➢ **Conv**

Selectin

- ➢ **Volu**

Allowin eneral
invitatio

The Inter /isit
www.mis olls. As
the site s opinion
of everyo out the
views of



misterpoll

Make a Poll - It's Free!

Enter a title for your new poll

NEXT STEP >

You can create your own polls on any subject with an unlimited number of questions. Share them with friends, submit them to our public directory, and more. Plus, every poll gets its own message forum!

Browse our public directory of polls

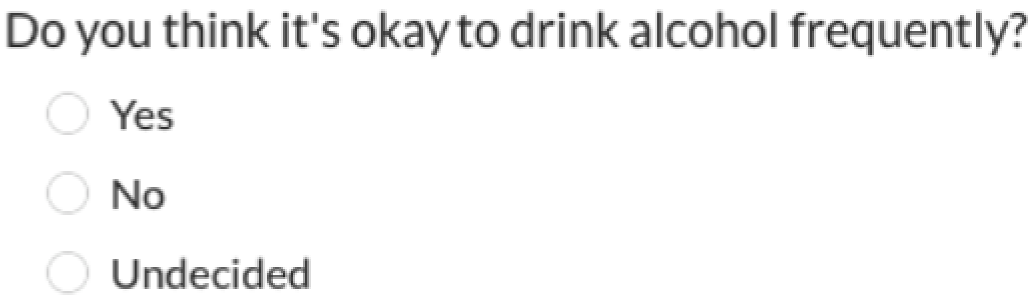
- Anime / Manga
- Books, Magazines, Comics
- Celebrities
- Computers
- Conspiracy
- Controversy / Morality
- Internet
- Miscellaneous
- Movies
- Music
- People / Relationships
- Pets / Animals

Editor's Choice Polls

- dude
- CVT Transmission Solutions
- Girls legs
- 18 year old girls
- How often do you face with social security benefits?
- Gym club start-up for Junior school boys
- does anyone else still have to wear short trousers?
- Dare game # 8: Who cooks dinner

How to Sample Badly (Bias)

- **Convenience bias**
- **Voluntary bias**
- **Selection / Undercoverage bias:** Some groups in the population are left out.
- **Measurement / Response Bias**
- **Nonresponse Bias:** Data are not obtained from all individuals in the sample.
- **Bias caused by the wording of questions**
- **…………**



Do you think it's okay to drink alcohol frequently?

○ Yes

○ No

○ Undecided

Exam Tip: describe how the design of a sample survey leads to bias

(1) describe how the members of the sample might respond differently from the rest of the population

(2) explain how this difference would lead to an underestimate or overestimate.

Example: Explain how using your statistics class as a sample to estimate the proportion of all high school students who own a graphing calculator could result in bias.

Answer: This is a convenience sample. It would probably include a much higher proportion of students with a graphing calculator than in the population at large because a graphing calculator is required for the statistics class. So this method would probably lead to an overestimate of the actual population proportion.

Bias & Error

Bias is not just bad luck in one sample!!

Bias is introduced by **the way in which a sample is selected or by the way in which the data are collected from the sample**.

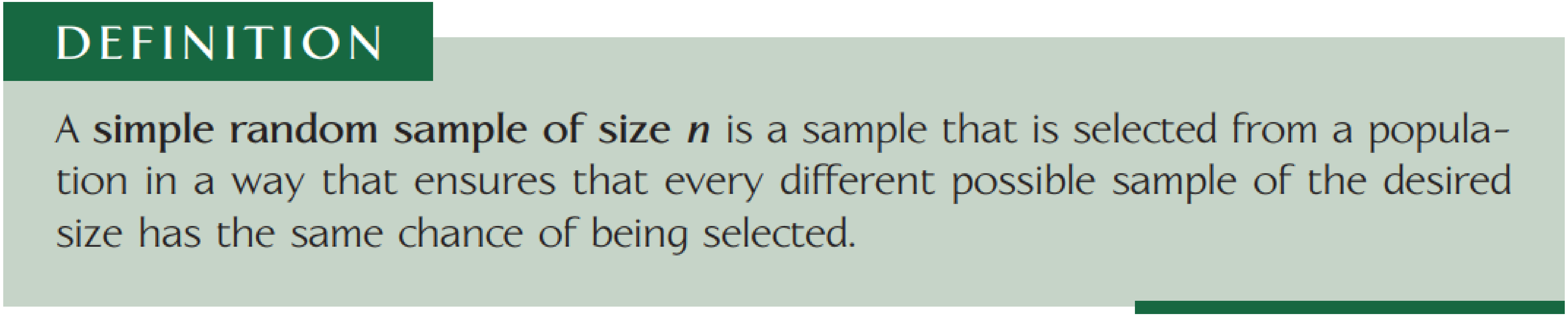
Increasing the size of the sample does nothing to reduce bias if the method of selecting the sample is flawed or if the nonresponse rate remains high.

DEFINITION

Sampling without replacement: Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes *n* distinct individuals from the population.

Sampling with replacement: After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

SRS （Simple Random Sampling）



DEFINITION

A **simple random sample of size *n*** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

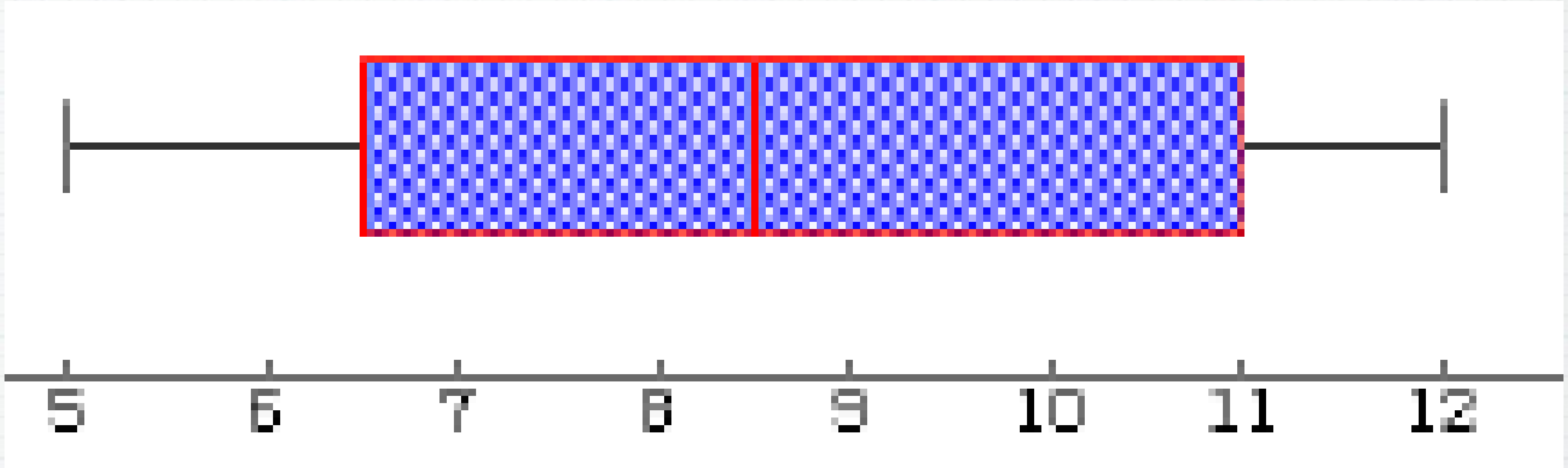
- objective
- free of selection bias

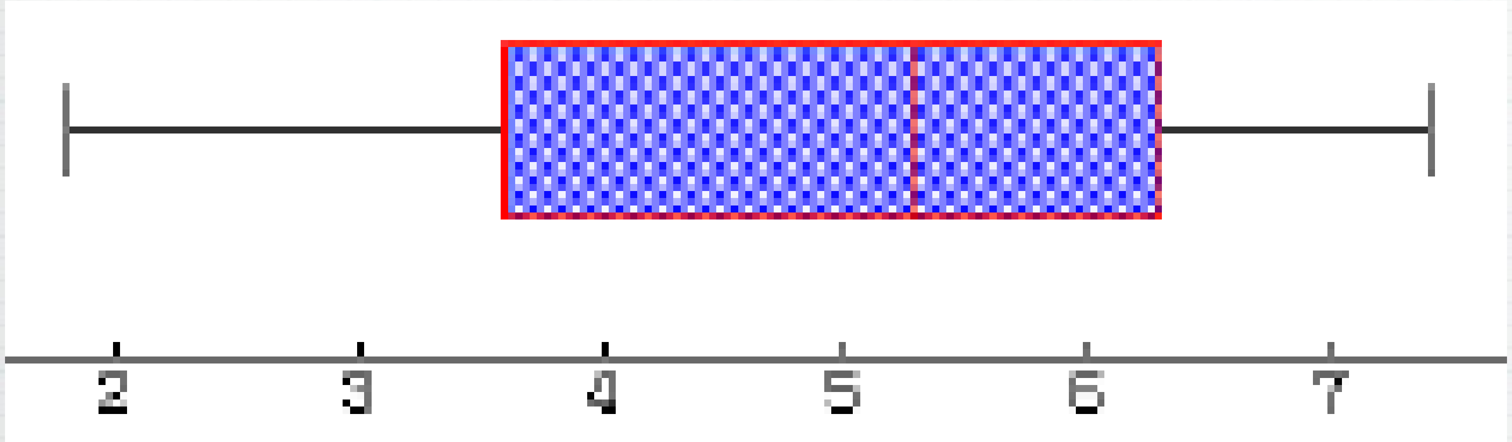
How to choose a SRS using Table of Random Digits

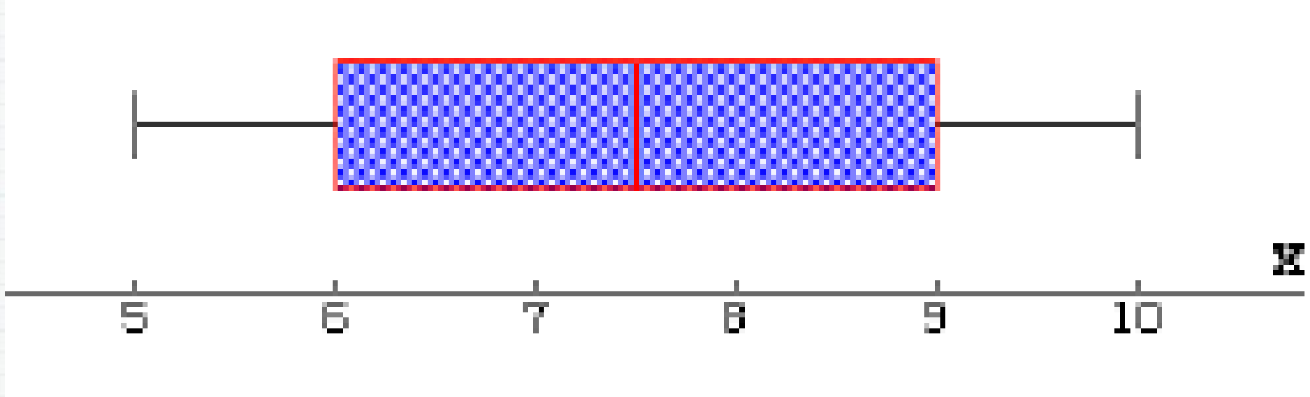
- **Label.**
  Give each member of the population a distinct numerical label with the same number of digits. Use as few digits as possible.

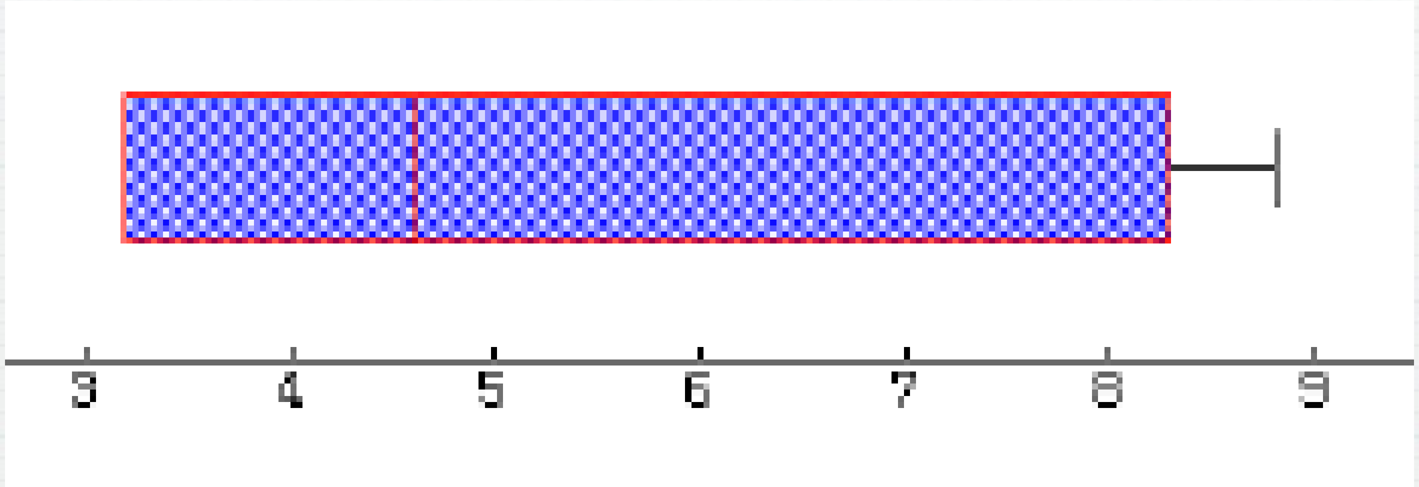
- **Randomize.**
  Read consecutive groups of digits of the appropriate length from left to right across a line in the table. Ignore any group of digits that wasn't used as a label or that duplicates a label already in the sample. Stop when you have chosen n different labels.

- **Select.**
  Choose the individuals that correspond to the randomly selected integers.


5
6
7
8
9
10
11
12


2
3
4
5
6
7


x
5
6
7
8
9
10


3
4
5
6
7
8
9

HOW to choose a SRS using Table of Random Digits

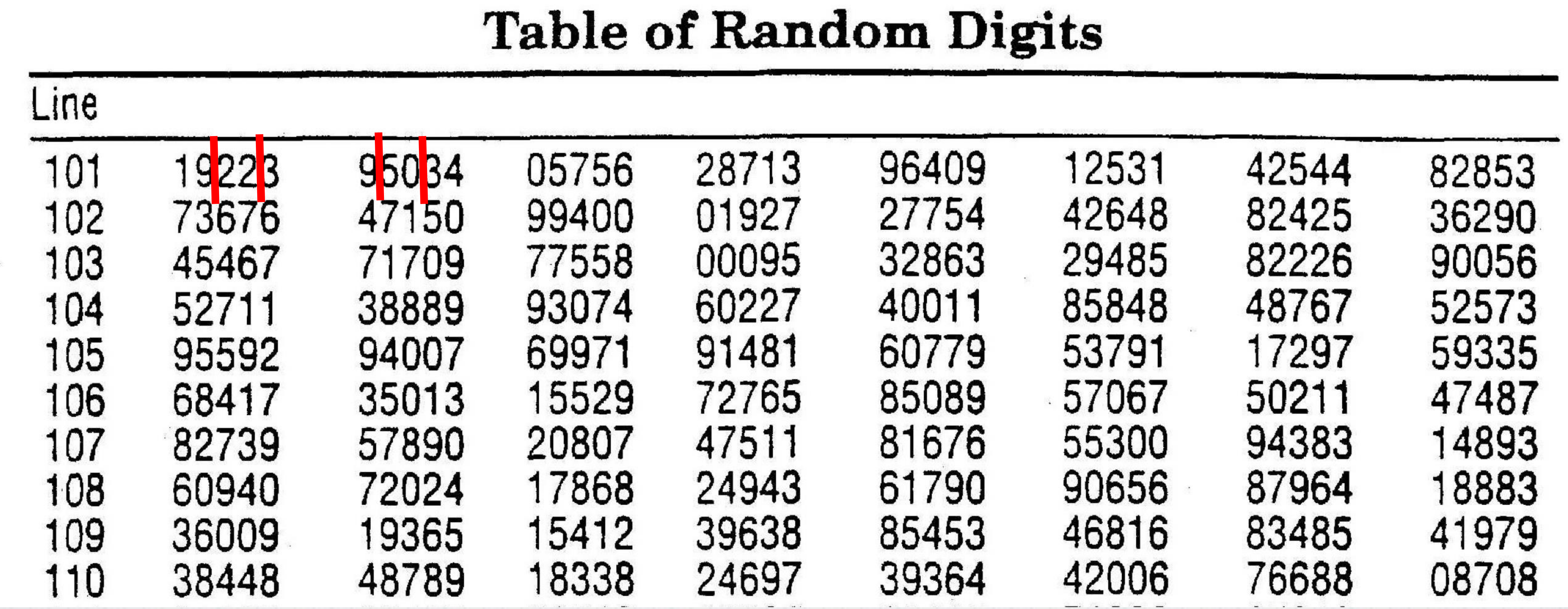


Table of Random Digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708

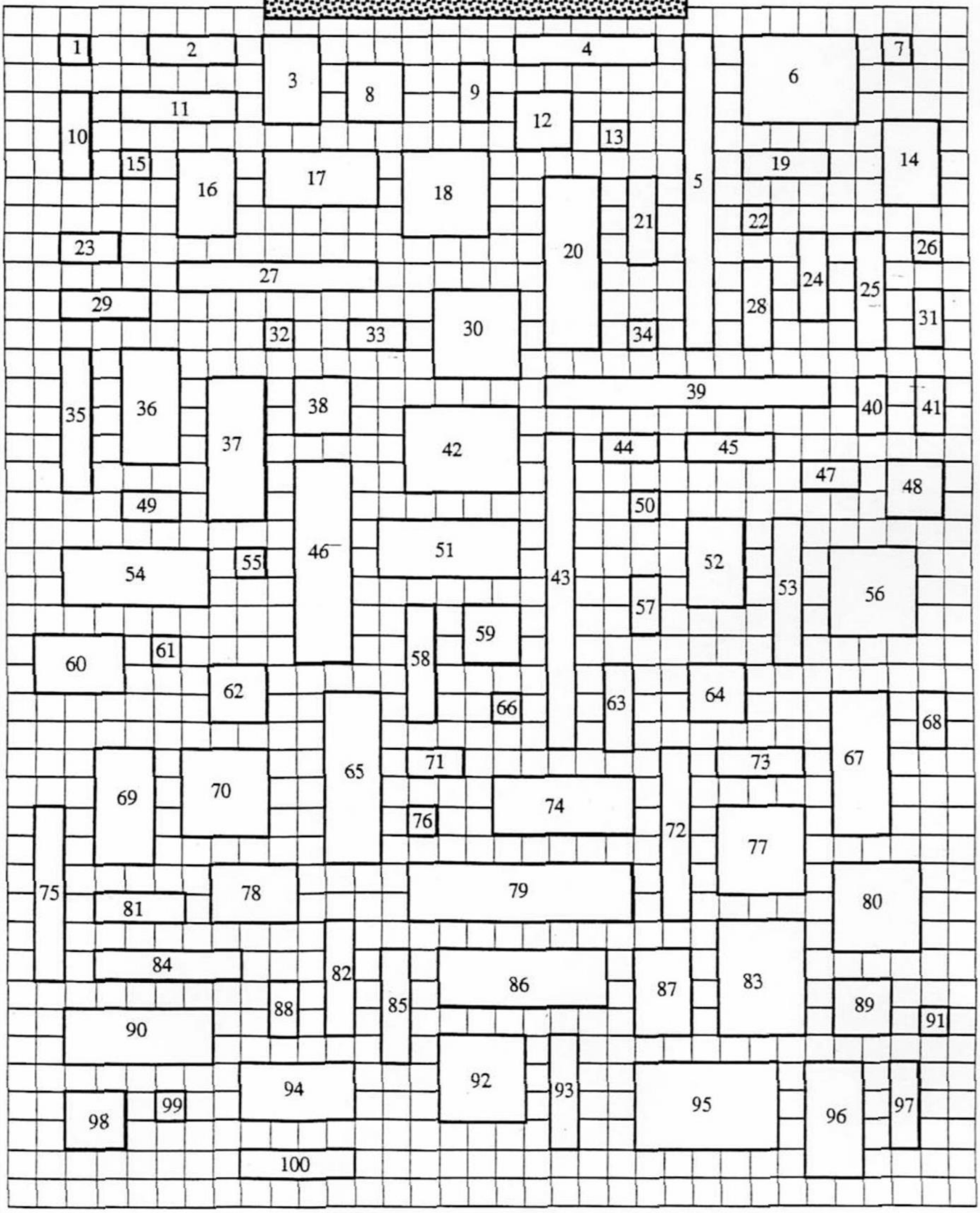
Guess - ten seconds to guess the average area of all the rectangles.

Expert or Judgment Sample – Select 5 rectangles that, in your judgment, are representative of the rectangles. Record the size of your five rectangles and find their mean.

Simple Random Sample (SRS) – Use the numbers on the rectangles and a random number table to choose five rectangles at random. Locate the corresponding rectangles and record their size and find the mean.

Random Rectangles

by Richard Scheaffer



actual mean = 5.27

Other probability sampling methods

1. systematic sample

- randomly choose some starting point;

- select every kth element in the population.

- easier than random sampling;

- guarantees that sample is taken from throughout the population.

2. cluster sample

- divide the population into sections or clusters (mutually exclusive)
- randomly select a few of those sections and then choose **all** members from the selected sections

- **faster** way to obtain sample

When to use this method:

- clusters are relatively heterogeneous, i.e. the population should contain distinct subpopulations of different types.

3. stratified random sample

- the entire population are divided into a set of nonoverlapping subgroups (strata)
- draw a random sample from each stratum

Purpose:

- the sample is more representative than a SRS might be
- might be interested in results from separate strata.

Sampling

The entire group of individuals that we want information about is called the __________.

A census is an attempt to gather information about____________
_________________________________.

Problems with census:____.

A _____ is a part of the population that we actually examine in order to gather information.

The _________________ refers to the method used to choose the sample from the population.

The design of a study is biased if it
____

Sampling

The entire group of individuals that we want information about is called the **population**.

A census is an attempt to gather information about **every individual member of the population**.

Problems with census: **costs、time、sometimes can destroy items•••**

A **sample** is a part of the population that we actually examine in order to gather information.

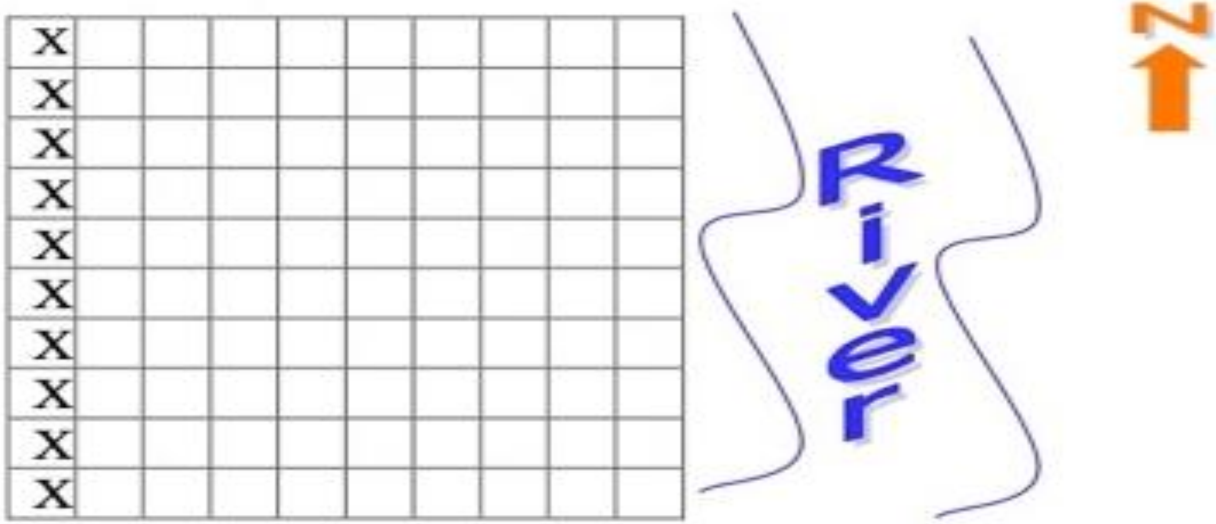
The **design of a sample** refers to the method used to choose the sample from the population.

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:



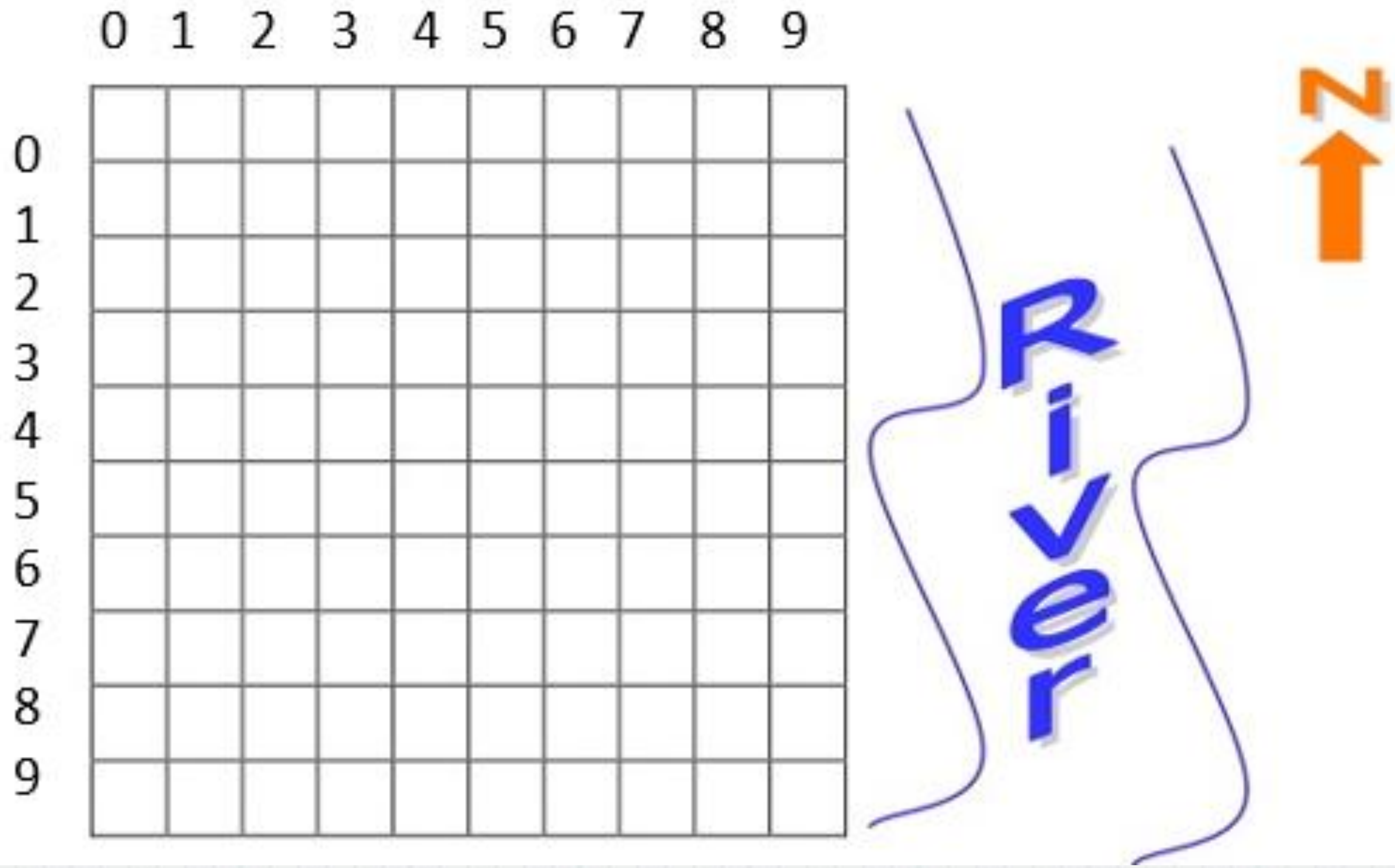
X									
X									
X									
X									
X									
X									
X									
X									
X									
X									

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.

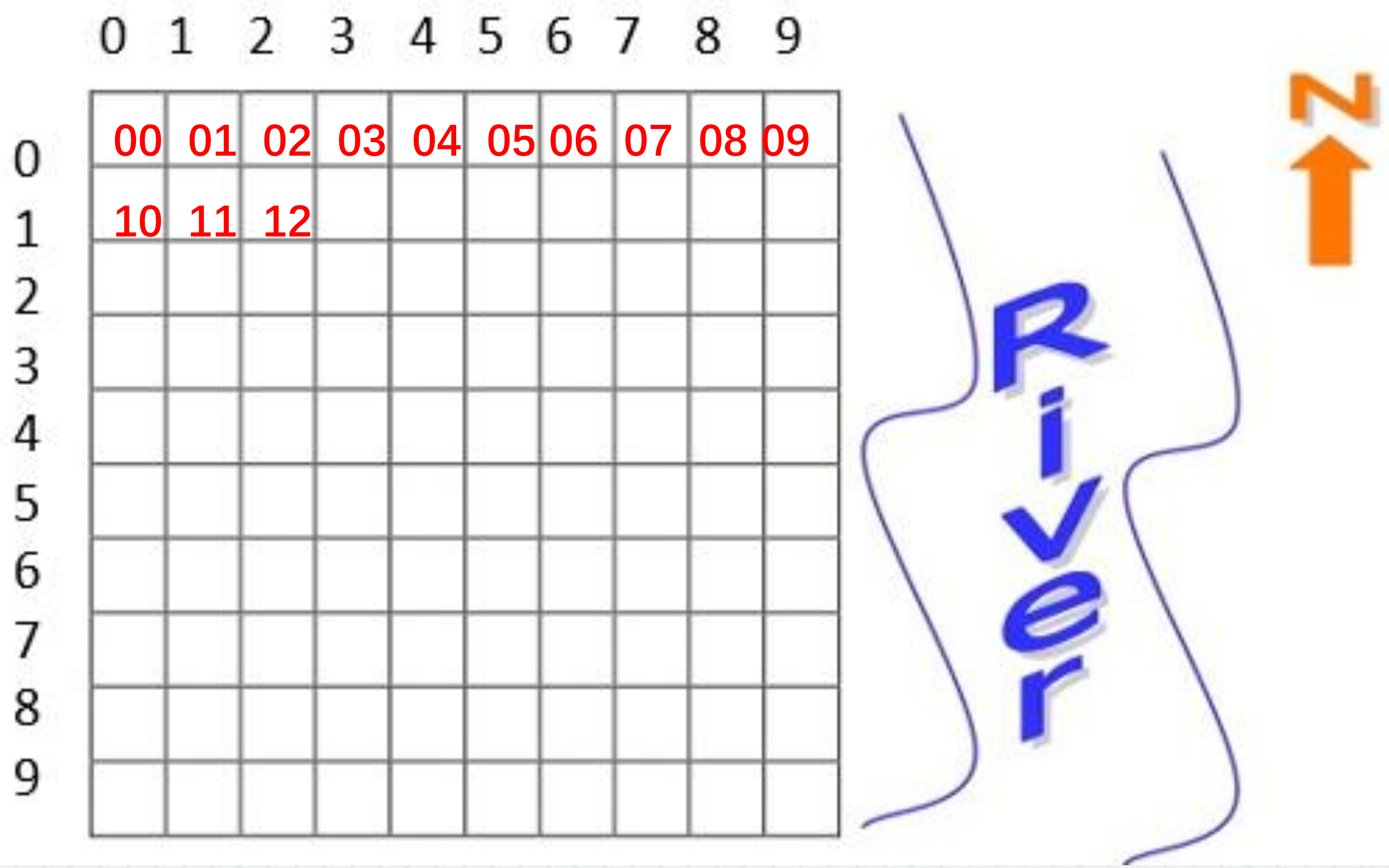
B. Method Number Two: Simple Random Sample

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.



B. Method Number Two: Simple Random Sample

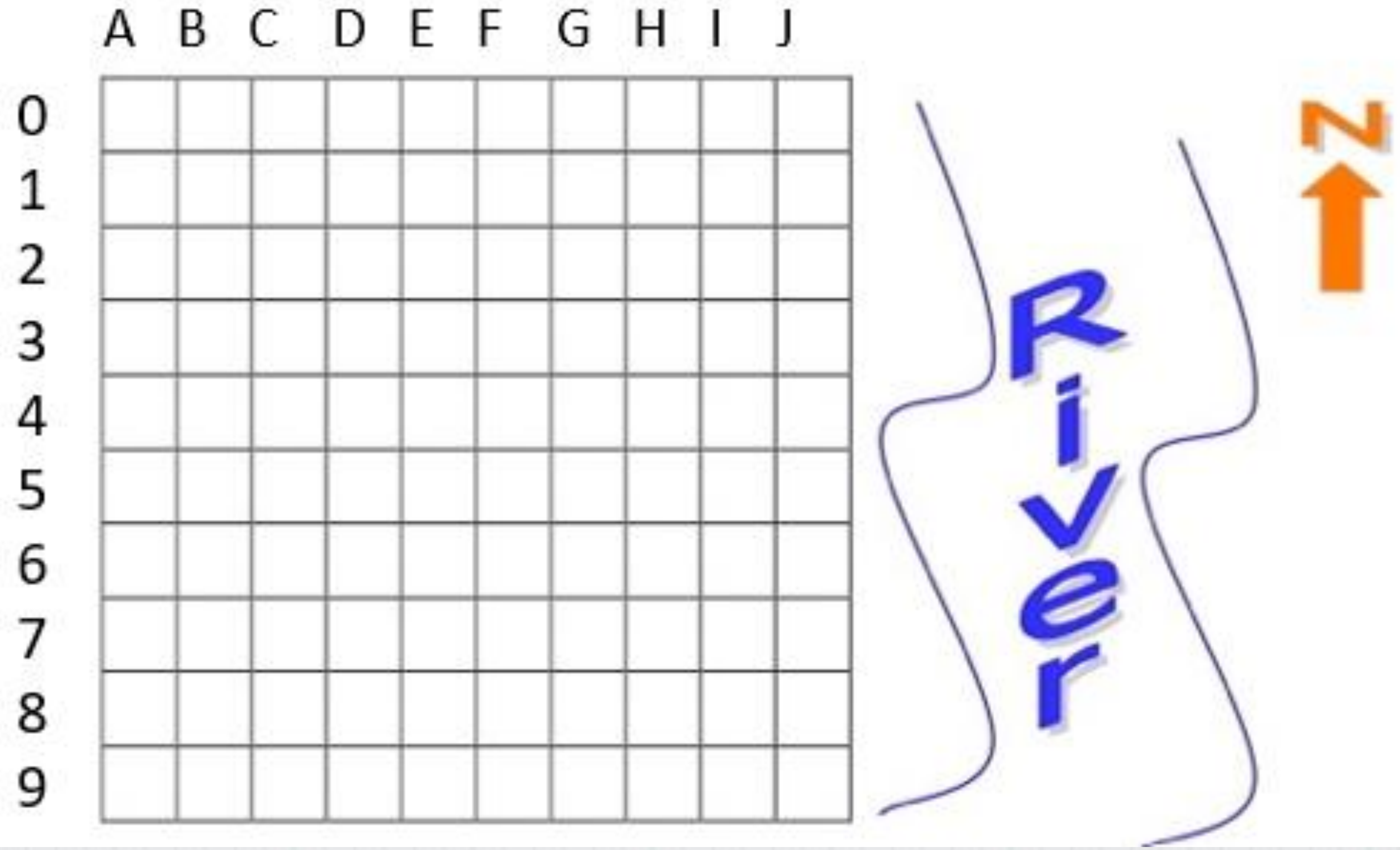
Use your calculator or a random number table to choose 10 plots to harvest. Mark them on the grid below, and describe your method of selection.



	0	1	2	3	4	5	6	7	8	9
0	00	01	02	03	04	05	06	07	08	09
1	10	11	12							
2										
3										
4										
5										
6										
7										
8										
9										

C. Method Number Three: Stratified Sample

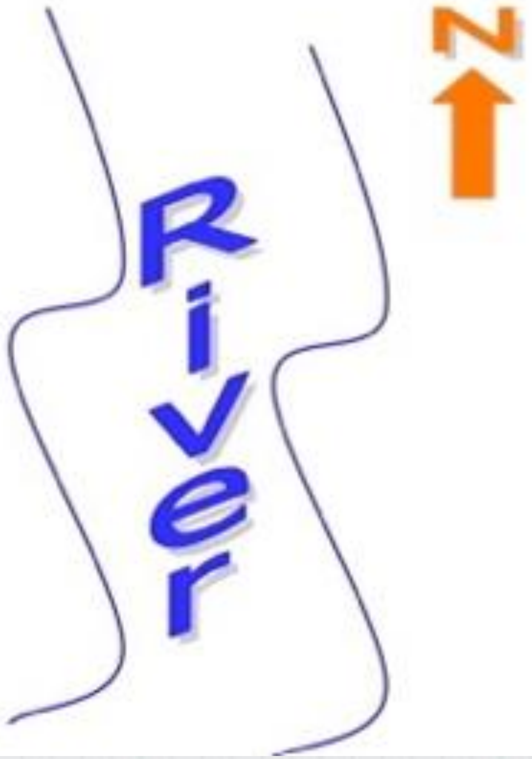
You and the farmer think the river might have a strong influence on corn production so you decide to consider the field as grouped in vertical columns (called strata—remember you can only stratify data your sample when you think it will have a strong influence on the outcome.). Using your random number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



C. Method Number Three: Stratified Sample

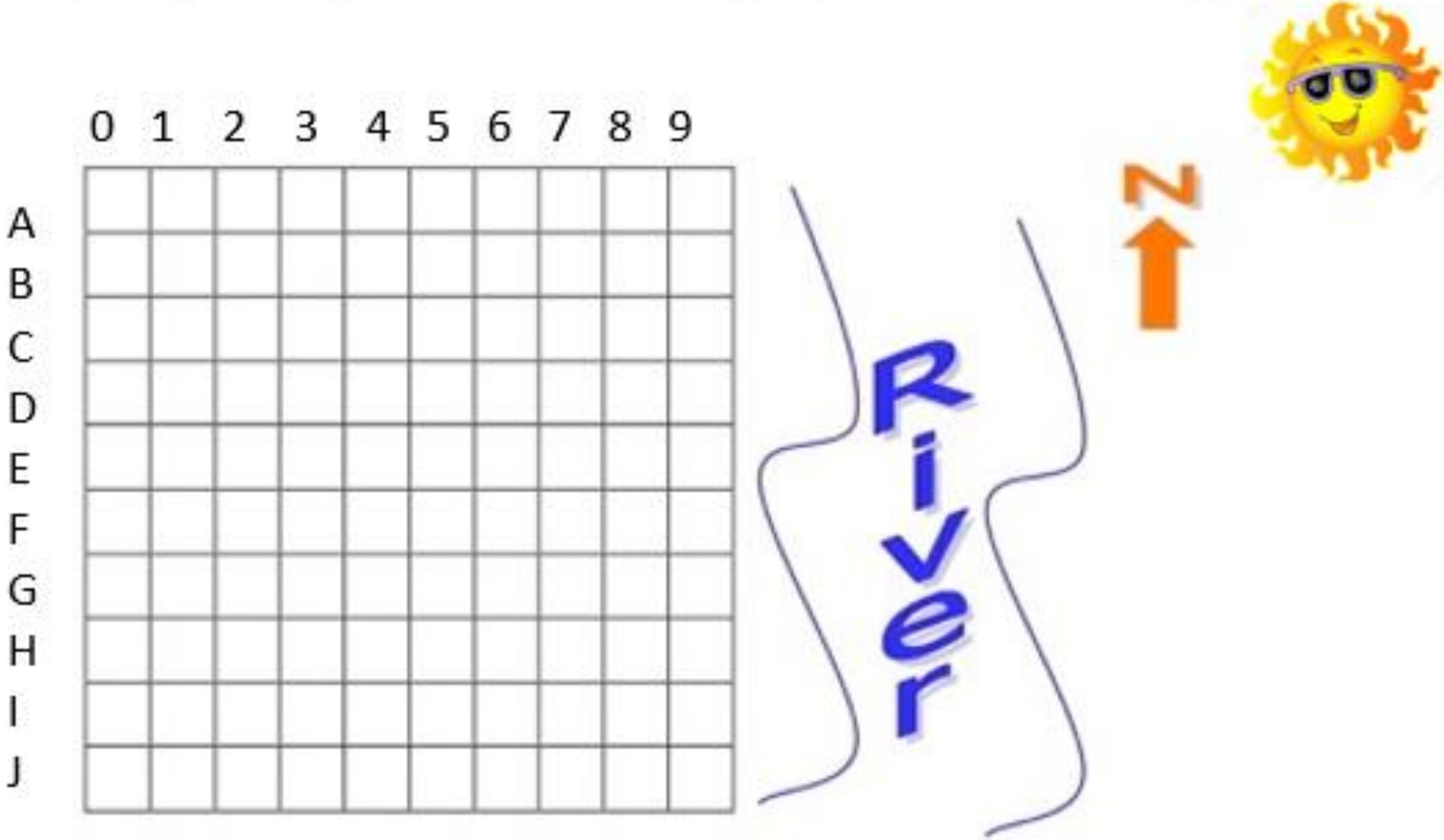
You and the farmer think the river might have a strong influence on corn production so you decide to consider the field as grouped in vertical columns (called strata—remember you can only stratify data your sample when you think it will have a strong influence on the outcome.). Using your random number 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
0	0									
1	1									
2	2									
3										
4										
5										
6										
7										
8										
9										



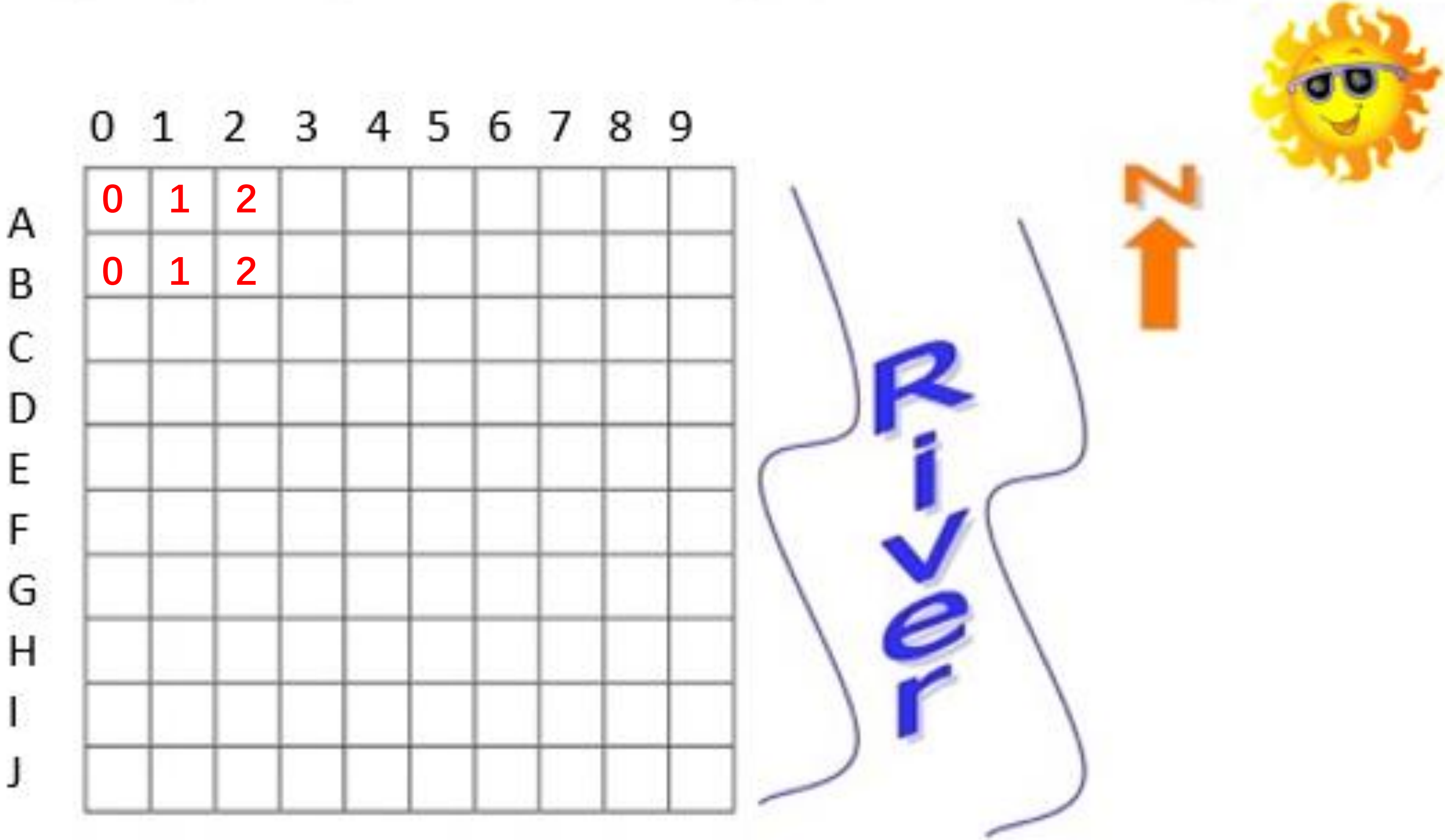
D. Method Number Four: Stratified Sample

You and the farmer think direction (north—south) might have a strong influence on corn production so 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.)

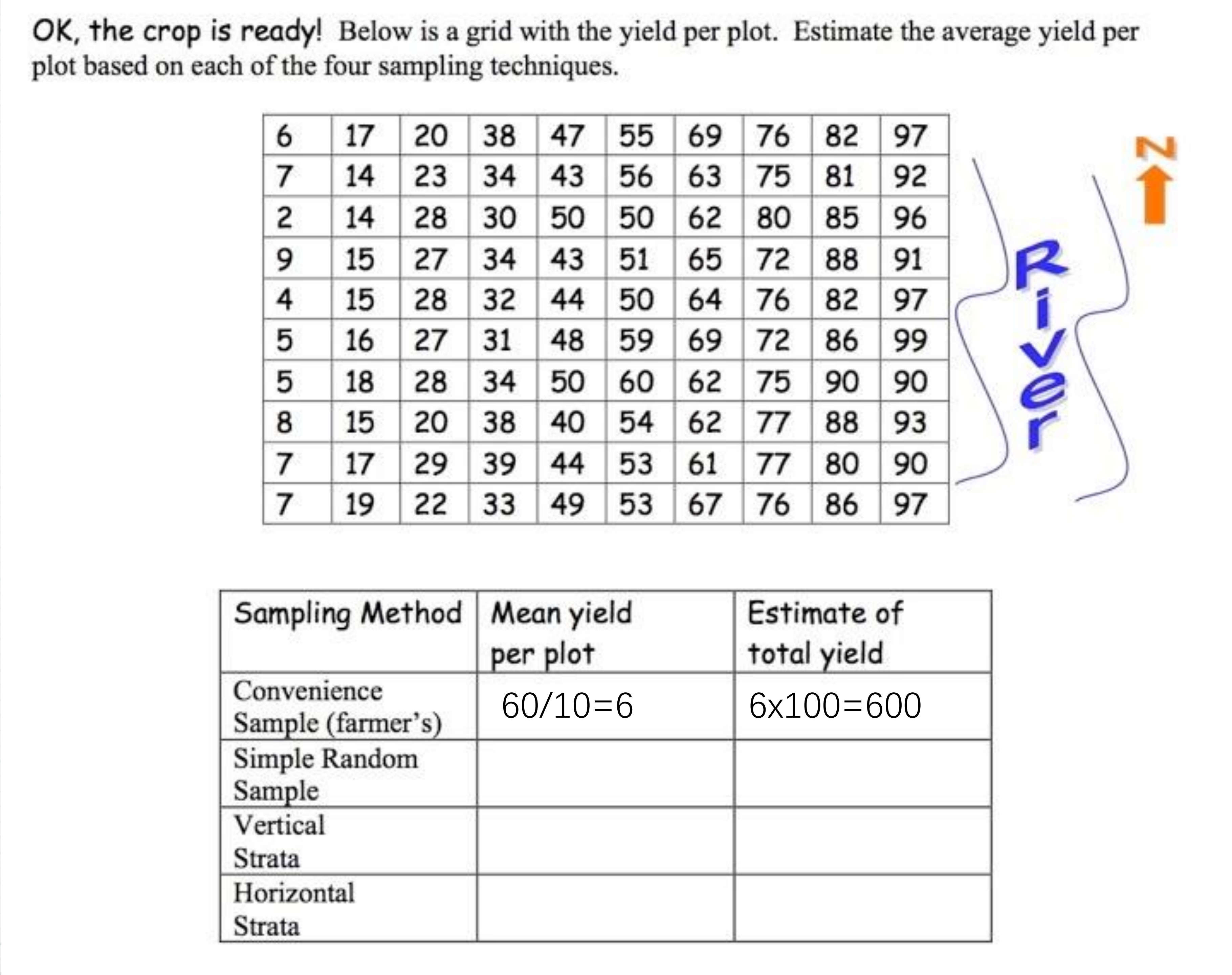


D. Method Number Four: Stratified Sample

You and the farmer think direction (north—south) might have a strong influence on corn production so 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
A	0	1	2							
B	0	1	2							
C										
D										
E										
F										
G										
H										
I										
J										

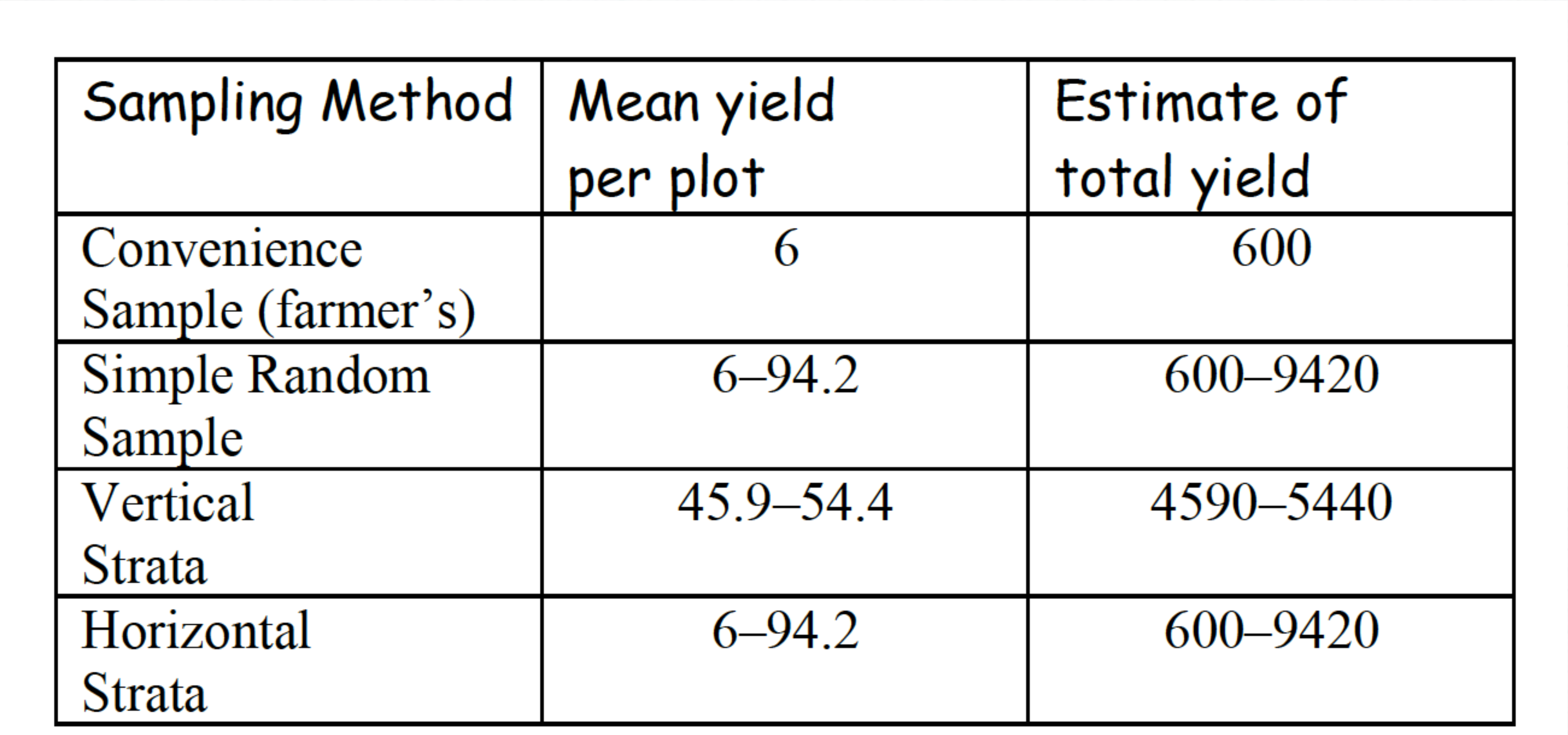


OK, the crop is ready! Below is a grid with the yield per plot. Estimate the average yield per plot based on each of the four sampling techniques.

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)	60/10=6	6x100=600
Simple Random Sample		
Vertical Strata		
Horizontal Strata		

Possible yield values



Sampling Method	Mean yield per plot	Estimate of total yield
Convenience Sample (farmer's)	6	600
Simple Random Sample	6–94.2	600–9420
Vertical Strata	45.9–54.4	4590–5440
Horizontal Strata	6–94.2	600–9420

actual total yield = 5004

Stratification may produce a **smaller error** of estimation than would be produced by a simple random sample of the same size. This result is particularly true if measurements within strata are very homogeneous – “stratification reduces variation”.