

Chapter 5 is all about producing data:

- 5.1 Designing Samples
- 5.2 Designing Experiments
- 5.3 Simulating Experiments

**Sample Design:** The point of sampling from a population is to gather good information about a specific group of interest so that a generalization can be made about it.

- reliable statistics come from good sampling techniques involving the use of chance
- worthless data comes from poor sampling techniques
- the population is the entire group about which we want information
- the sample is the part of the population we actually examine.
  - it should represent the population well
- the frame is a list of all of the sampling units (ideally the whole pop.)
  - an incomplete frame can lead to bias in sampling
- a census is an attempt to gather information from every individual in a population

We sample a population when we want to do one of two things:

- Observe
- Experiment

Observational Study vs. Experiment

Observational Study	Experiment
<ul style="list-style-type: none"> <li>• no treatment imposed</li> <li>• no influencing the response</li> <li>• watching for trends that exist</li> </ul>	<ul style="list-style-type: none"> <li>• implement a treatment</li> <li>• trying to cause and measure change</li> <li>• must be well designed to be effective</li> </ul>

Observational Studies and Experiments BOTH require good

**Sample Design:**

Probability Sampling	Non-Probability Sampling
<ul style="list-style-type: none"> <li>SRS (Simple Random Sample) names in a hat/lottery/raffle</li> <li>Stratified Random Sample group, then SRS</li> <li>Multistage (cluster) layers, then SRS</li> </ul>	<ul style="list-style-type: none"> <li>Voluntary Response respondents choose selves</li> <li>Convenience Sampling respondents are chosen by location</li> </ul>

When non-probability based methods are used, they often lead to biased design (systematically favoring certain outcomes)

There are 5 types of bias that we will refer to:

- voluntary response bias (self selected subjects)
- non-response bias (subjects given opportunity, but do not participate)
- response bias (lying)
- undercoverage bias (subjects not given opportunity to participate)
- poor wording bias (extra information to sway choice)

When probability based methods are used correctly, they lead to good data that can often be analyzed and used to make inferences about a population.

How do we choose a good sample with probability based methods?...

Start with SRS (Simple Random Sample)

We will use our random table of digits (Table B in the back of your book) to generate random samples from a population.

The **Label & Table** method:

1. Know how many individuals are in the population, assign numbers
 

for $n < 10$ :	0 - 9
$11 < n < 99$	01 - 99
$11 < n < 100$	00 - 99
$101 < n < 999$	001 - 999
$101 < n < 1000$	000 - 999
  
2. Use the table of random digits to select individuals for sample
  - choose a random line to start from
  - using every digit, check in groups of the same size as the assignments (0-9: check one digit at a time, 01-99: check in pairs)

Example: I want to choose three people to go to Islands of Adventure and there are 5 people to choose from:

	Assign:		Use an efficiency method:
Amy	0		0,1
Bill	1		2,3
Chuck	2	OR	4,5
Dave	3		6,7
Edward	4		8,9

Choose a starting line in the table:

101 19223 95034 05756

General Rules for using a table of random digits:

- Use every number - don't worry about "wasting" the digits
- When assigning numbers, use the smallest number of digits possible
- Use the table in the same number of digits as the size of your labels
- ignore numbers that are not assigned as labels