

Unit 3

Populations and Samples

“To all the ladies present and some of those absent”

- Jerzy Neyman

The collection of all individuals with HIV infection and the collection of all individuals with exposure to mercury are examples of *populations* about which we wish to make inferences. A *census* is one way to obtain information about a population but is usually impractical because it requires the collection of data for every individual in the population.

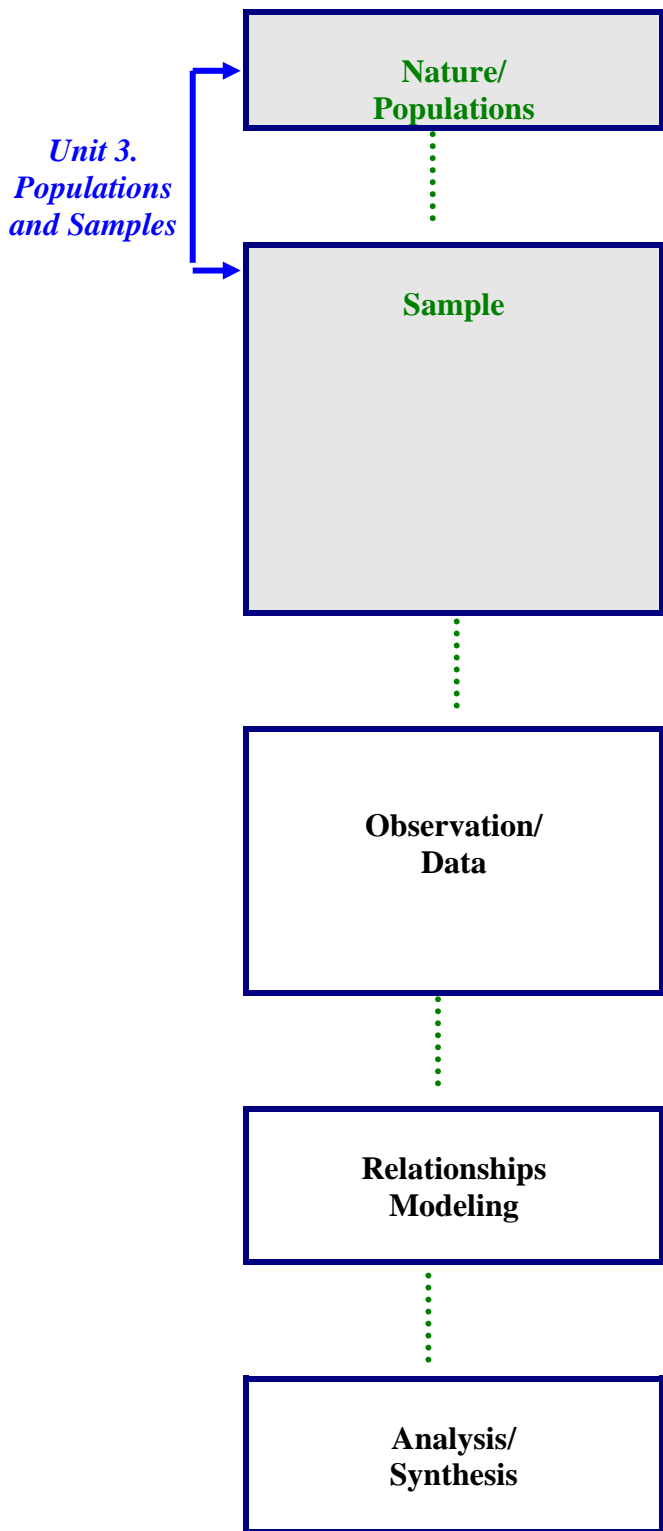
Instead, typically, we study a fraction of the population called a *sample*. If the sample is representative of the population, then quantities computed from available data in the sample are reasonable estimates of the corresponding, but unavailable, quantities in the population.

A sample is not a replica of the population of interest. Thus, the goal of sampling is to obtain a representative sample such that the inferences drawn from the sample are in error by as little as possible.

Table of Contents

Topics	1. Unit Roadmap	3
	2. Learning Objectives	4
	3. A Feeling for Populations v Samples	5
	4. Target Populations, Sampled Populations, Sampling Frames	8
	5. On Making Inferences from a Sample	11
	6. Simple Random Sampling	13
	7. Some Non-Probability Sampling Plans	16
	8. More on Simple Random Sampling	19
	a. Sampling WITH v WITHOUT replacement	20
	b. How to select a simple random sample	28
9. Some Other Probability Sampling Plans	31	
a. Systematic	31	
b. Stratified	33	
c. Multi-stage	35	
10. The Nationwide Inpatient Survey (NIS)	36	

1. Unit Roadmap



Representative - Glance again at the roadmap at the footer of this page. We're headed to estimation and hypothesis tests of data in samples. We hope that the conclusions drawn from these reasonably apply to the population as well. For this to be possible at all, the sample must be representative.

Minimum Variance – A conclusion drawn from a sample will differ from the reality of the population. This is *sampling error*. An additional goal of sampling is to obtain a sample for which sampling error is minimized.

Nature
Population/
Sample
Observation/
Data
Relationships/
Modeling
Analysis/
Synthesis

2. Learning Objectives

When you have finished this unit, you should be able to:

- Explain the distinction between target population, sampled population, and sample.
- Explain why it is important that a sample should be representative of the population from which it is taken.
- Explain the rationale for choosing a sampling method that minimizes sampling error.
- Distinguish non-probability versus probability samples.
- Define simple random sampling.
- Distinguish sampling with versus without replacement.
- Explain the rationale for systematic, stratified, and multi-stage sampling methods.
- Define systematic, stratified, and multi-stage sampling.

3. A Feeling for Populations versus Samples

In unit 1, our task was to summarize the information in a sample. There was no consideration of the source of the data. We didn't ask "*what population did these data come from?*", "*how was the data obtained?*", nor "*was random sampling used?*"? In unit 1, *Summarizing Data*:

- The techniques and advantages of graphical and tabular summaries of data were emphasized; and
- We familiarized ourselves with the basics of summarizing data using numerical approaches (means, variances, etc)

Now, let's step back.

Consider the source of our "sample." It's a population, but what population? Is the sample "representative?" Can we use the conclusions drawn from our sample to say anything about the source population from which the sample came?

- A population – What *can* we say about the population?
- A research question – Is the information in our sample enough to support or refute a specific hypothesis about the population?
- Probabilities – What were the chances of observing a specific event in the sample? High and plausible? Or low and implausible?

Statistical inference requires that the sample studied be a probability sample.

Nature — Population/
Sample — Observation/
Data — Relationships/
Modeling — Analysis/
Synthesis

Example – The 1948 Gallup Poll

- Before the 1948 presidential election, Gallup polled 50,000 registered voters.
- Each was asked who they were going to vote for – Dewey or Truman

	Predicted by Gallup Poll	True Election Result
Dewey	50%	45%
Truman	44%	50%

- How could the prediction have been so wrong? Especially when n=50,000
- How could the sample have been so dissimilar to the population?

The 1948 Gallup Poll

- The actual sample turned out not to be representative of the population of actual voters
- It is thought that the error was the result of two things:

FIRST: The interviewers over-sampled

1. wealthy
2. safe neighborhoods, those with telephones AND . . .

SECOND: The over-sampled included a disproportionate number of voters favoring Dewey; ie - there was an over-sampling of the segment of the population more likely to vote for Dewey

Bias occurred because two things occurred:

- (1) there was over-sampling; *and*
- (2) the nature of the over-sampling was related to voter preference.

Note – Oversampling, per se, does not produce bias in study findings necessarily.

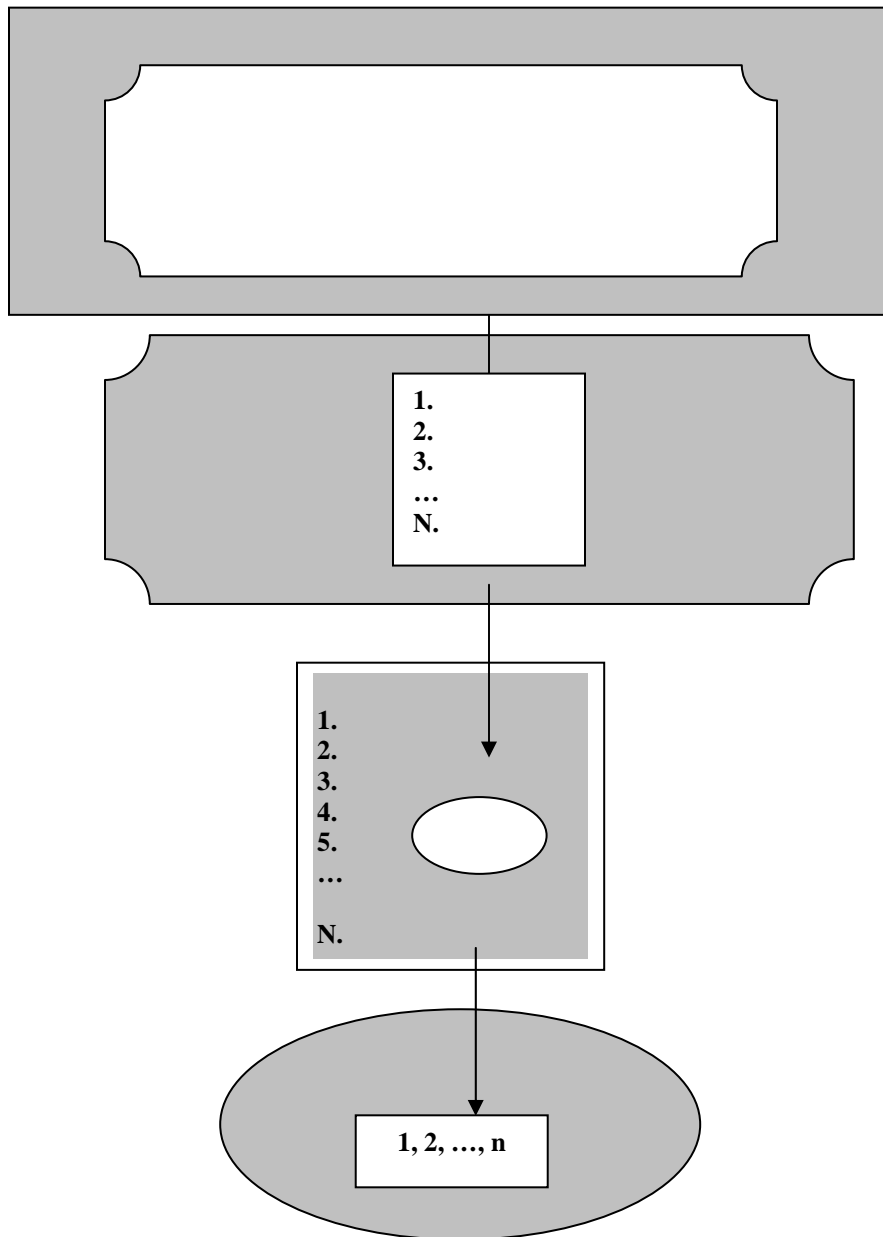
Note – We will say much more about “bias” later. For now, think of bias as the extent to which a finding is incorrect.

Example, continued - The 1948 Gallup Poll

The population actually sampled (the sampled population) was *not the same* as the population of interest (the target population)



4. Target Population, Sampled Population, Sampling Frame



Target Population

The whole group of interest.

Note – A convention is to use capital “N” to represent the size of a finite population.

Sampled Population

The subset of the target population that has at least some chance of being sampled.

Sampling Frame

An enumeration (roster) of the sampled population.

Sample

The individuals who were actually measured and comprise the available data.

Note – A convention is to use small “n” to represent the size of a sample.

<p>Target Population</p>	<ul style="list-style-type: none"> • The entire collection of individuals who are of interest. • Example – The population of 1948 presidential election voters who actually voted.
<p>Sampled Population</p>	<ul style="list-style-type: none"> • The aggregate of individuals that was actually sampled. • A listing of the entire sampled population comprises the sampling frame. • GOAL: sampled = target • The sampled population is often difficult to identify. We need to ask: Who did we miss? • Constructing the sampling frame can be difficult • Example – The 1948 Gallup poll sample is believed to have been drawn from the subset of the target population who were <ul style="list-style-type: none"> ♣ Easy to contact, and ♣ Consenting, and ♣ Living in safe neighborhoods

Sampling Frames Why They Are Difficult

To Construct a Sampling Frame Requires

- ♣ **Enumeration** of every individual in the sampled population
- ♣ Attaching an **identifier** to each individual
- ♣ (Often, this identifier is simply the individual's **position** on the list)

Example –

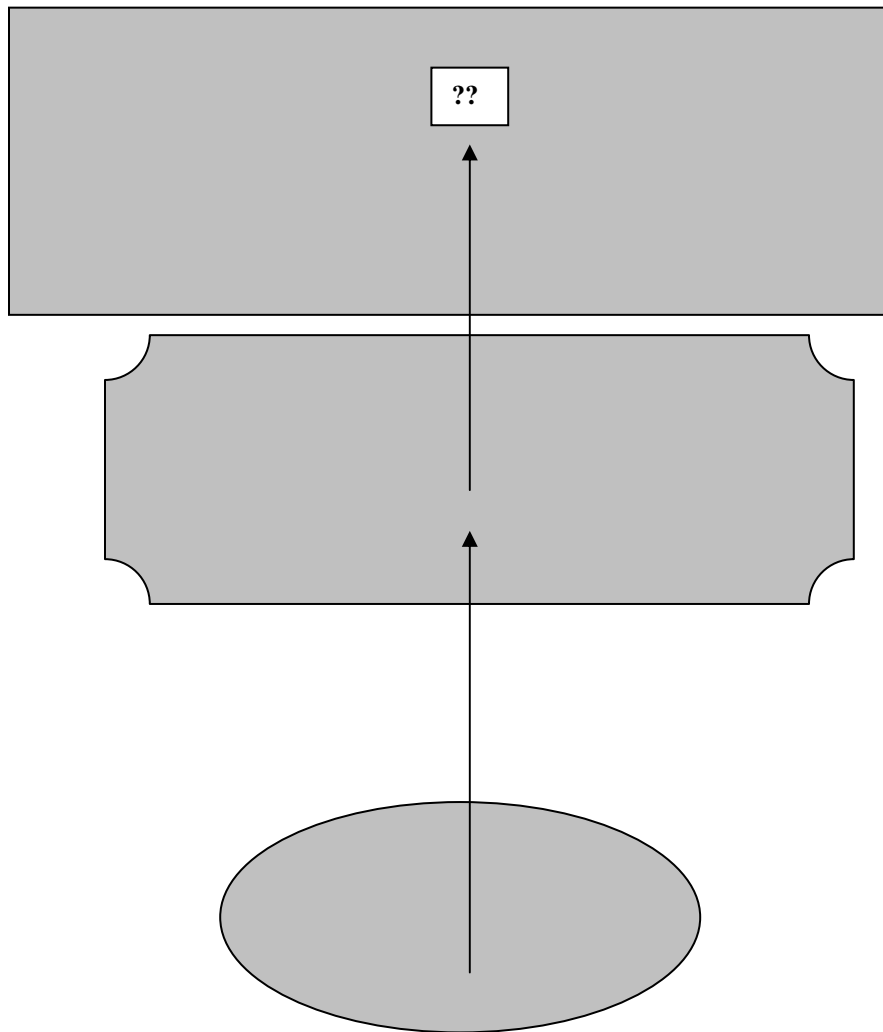
- ♣ The League of Women's Voters Registration List might be the sampling frame for the target population who vote in the 2008 election.
- ♣ Individual identification might be the position on this list.

Now You Try –

- ♣ The target population is joggers aged 40-65 years.
- ♣ How might you define a sampling frame?

Nature
**Population/
Sample**
Observation/
Data
Relationships/
Modeling
Analysis/
Synthesis

5. On Making Inferences From a Sample
(this time - read from the bottom up)



Target Population

The conclusion may or may not generalize to the target population.

- Refusals
- Selection biases

Sampled Population

If sampling is representative, then the conclusion generalizes to the sampled population.

Sample

The conclusion is drawn from the sample.

- ♣ The conclusion is initially drawn from the sample.
- ♣ The question is then: How far back does the generalization go?
- ♣ The conclusion usually applies to the sampled population
- ♣ It may or may not apply to the target population
- ♣ The problem is: It is not always easy to define the sampled population

Example – an NIH Funded Randomized Trial

- ◆ The sampling frame, by definition, is allowed to contain only consenters
- ◆ Thus, refusers, by definition, are not in the sampling frame.
- ◆ Thus, in randomized trial protocol that includes consent, the sampled population differs from the target population because it is restricted to consenters only.
- ◆ This suggests that in any study, the preliminary analyses should always include a comparison of the consenters versus the refusers.

Now You Try ...

- ◆ Suppose the target population is current smokers.
- ◆ How might you construct a sampling frame?
- ◆ What do you end up with for a sampled population?
- ◆ Comment on the nature of generalization, to the extent possible.



6. Simple Random Sampling

We would like our sampling plan to be :

- ◆ **Representative** – If sampling is repeated over and over and over, the “long run” average conclusion about the population should be correct. (note – This is **“unbiasedness”**);
- ◆ **Minimum variance** – The discrepancy between the conclusion drawn from the population and the fact of the population should be as small as possible;

Definition simple random sampling:

Simple random sampling is the method of sampling in which every individual in the sampling frame has the same chance of being included in the sample.

The virtue of **simple random sampling** is that it is **unbiased**, meaning:

- ◆ IF we draw sample after sample after sample after sample
AND IF, for each sample, we compute a sample \bar{X} as our guess of μ ,
so as to compile a collection of sample estimates \bar{X} ,
- ◆ THEN “in the long run”...
the average of all the sample estimates, average of (\bar{X} after \bar{X} after \bar{X} ...)
will be equal to the population parameter value (the true value of μ)

Example of Simple Random Sampling
“Simple Random Sampling Without Replacement is unbiased”

Suppose the Sampling Frame == Target Population

Subject ID	1	2	3	4	5	6
Age, years	21	22	24	26	27	36

The following is true for the population of size N=6

- ◆ Population mean age $\mu = \frac{21 + 22 + 24 + 26 + 27 + 36}{6} = 26$ years
- ◆ The investigator doesn't know this value. That's why s/he is taking a sample!

Sampling Procedure

- ◆ Draw a random sample of n=3 subjects from the population. Do each successive draw “without replacement”. **Note – “Without replacement” means that each selected person, once selected, is NOT returned to the population for future sampling. More on this later.**

Calculation for Each Sample

- ◆ Sample mean $\bar{X} = \frac{\text{1st value} + \text{2nd value} + \text{3rd value}}{n = 3}$
- ◆ In this illustration (but not in real life) we can calculate the error of each \bar{X} by computing

$$\text{error} = 26 - \bar{X}$$

Consider all possible (without replacement) samples of size n=3 from a population of size N=6. There are 20 such samples.

Sample, n=3	Sample \bar{X}	Error = 26 - \bar{X}
{ 21, 22, 24 }	22.333	+3.667
{ 21, 22, 26 }	23	+3
{ 21, 22, 27 }	23.333	+2.667
{ 21, 22, 36 }	26.333	-0.333
{ 21, 24, 26 }	23.667	+2.333
{ 21, 24, 27 }	24	+2
{ 21, 24, 36 }	27	-1
{ 21, 26, 27 }	24.667	+1.333
{ 21, 26, 36 }	27.667	-1.667
{ 21, 27, 36 }	28	-2
{ 22, 24, 26 }	24	+2
{ 22, 24, 27 }	24.333	+1.667
{ 22, 24, 36 }	27.333	-1.333
{ 22, 26, 27 }	25	+1
{ 22, 26, 36 }	28	-2
{ 22, 27, 36 }	28.333	-2.333
{ 24, 26, 27 }	25.667	+0.333
{ 24, 26, 36 }	28.667	-2.667
{ 24, 27, 36 }	29	-3
{ 26, 27, 36 }	29.667	-3.667

$$\frac{\sum \text{sample } \bar{X}}{\text{all 20 possible samples}} = \mu = 26 \qquad \sum_{\text{sample}\#1}^{\text{sample}\#20} [\text{error}] = 0$$

This sampling plan is **unbiased** because:

- * The average of the \bar{X} , taken over all possible samples, is equal to $\mu = 26$.
- * Equivalently, the sum of all the errors, (26 - \bar{X}), is equal to 0.



7. Some Non-Probability Sampling Plans

Non-probability samples are haphazard and, for this reason, results of analyses of non-probability samples cannot be assumed to be representative of the population of interest. Nevertheless, non-probability sampling methods are sometimes used. Three examples of non-probability sampling plans:

- (1) **Quota**
- (2) **Judgment**
- (3) **Volunteer/Convenience**

(1) Quota Sampling Plan

Example –

Population is 10% African American
 Sample size of 100 must include 10 African Americans

How to Construct a Quota Sample

1. Determine relative frequencies of each characteristic (e.g. gender, race/ethnicity, etc) that is hypothesized to influence the outcome of interest.
2. Select a fixed number of subjects of each characteristic (e.g. males or African Americans) so that

Relative frequency of characteristic in **sample** (e.g. 10%)

MATCHES

Relative frequency of characteristic in **population** (e.g. 10%)



(2) Judgment Sampling Plan

Decisions regarding inclusion or non-inclusion are left entirely to the investigator. Judgment sampling is sometimes used in conjunction with quota sampling.

Example –

“Interview 10 persons aged 20-29, 10 persons aged 30-39, etc”
 Sample size of 100 must include 10 African Americans.

Example –

Market research at shopping centers

(3) Volunteer/Convenience Sampling Plan

Volunteers are recruited for inclusion in the study by word of mouth, sometimes with an incentive of some sort (eg. gift certificate at a local supermarket)

Example –

For a study of a new diet/exercise regime, volunteers are recruited through advertising at local clinics, health clubs, media, etc.

Problem?



Limitations of a Non-Probability Sampling Plan

They're serious!

1. We have no idea if the sampling plan produces unbiased estimators. It probably doesn't.
2. Any particular sample, by using fixed selection, may be highly unrepresentative of the target population.
3. Statistical inference, by definition based on some sort of probability model, is not possible.
4. Regarding quota sampling
 - We have no real knowledge of how subjects were selected (recall the Gallup poll example)
5. Regarding judgment sampling
 - Likely, there is bias in the selection – at least for the reasons of comfort and convenience
6. Regarding volunteer sampling
 - Volunteers are likely to be a “select” group for being motivated

Nature Population/
Sample Observation/
Data Relationships/
Modeling Analysis/
Synthesis

8. More on Simple Random Sampling

A **probability sampling plan** is “out of the hands” of the investigator.

- ◆ Each individual has a **known probability** of inclusion in the sample, prior to sampling.
- ◆ The investigator has **no discretion** regarding the inclusion or exclusion of an individual
- ◆ This eliminates one source of potential bias – that on the part of the investigator.

How do we know if a sample is REPRESENTATIVE?

- ◆ Ultimately, we don’t know!.
- ◆ So, instead, we use an unbiased sampling plan and hope for the best.
- ◆ In the meantime, we can generate some descriptive statistics and compare these to what we know about the population.

Simple Random Sampling is the basic probability sampling method.

Recall its definition from page 13:

Simple random sampling is the method of sampling in which every individual in the sampling frame has the same chance of being included in the sample.

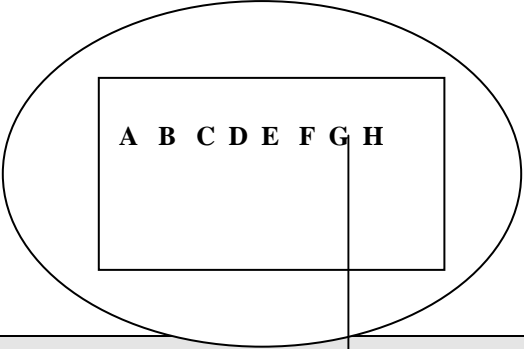
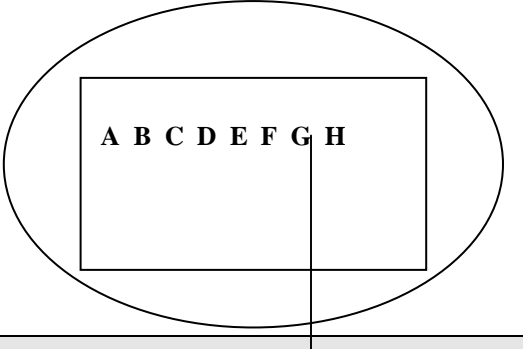
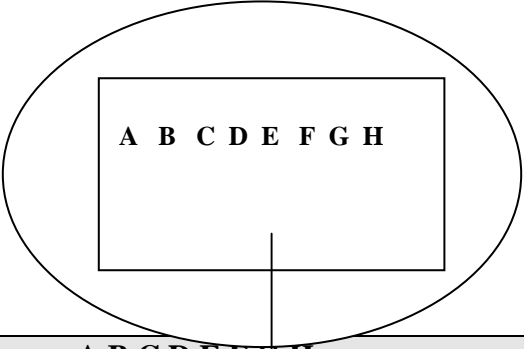
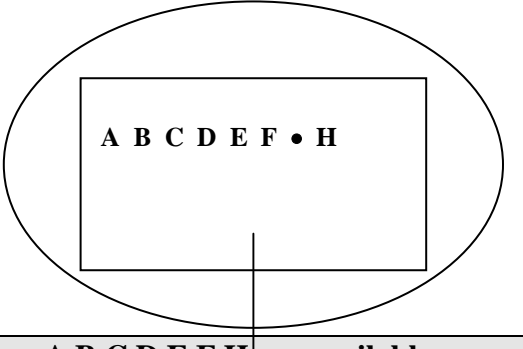
Under simple random sampling

$$\text{Probability \{each equally likely sample\}} = \frac{1}{\text{number of equally likely samples}}$$

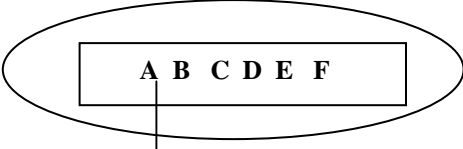

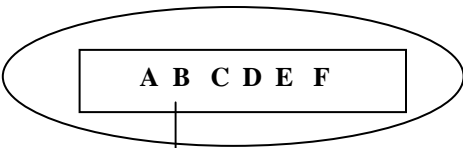
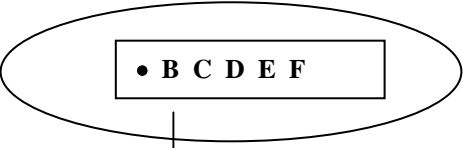
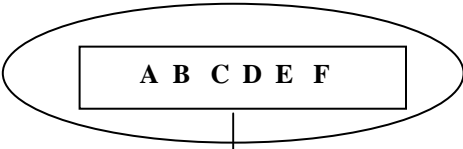

Thus, we need to solve for **the number of equally likely samples!**

Nature ——— Population/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

How many Equally Likely Samples Are there?
a. Simple Random Sampling *With* Replacement
versus
Simple Random Sampling *Without* Replacement

With Replacement	Without Replacement
	
<p>1st Draw: A B C D E F G H are available</p> <p>- Suppose “G” is selected for inclusion</p>	<p>1st Draw: A B C D E F G H are available</p> <p>- Suppose “G” is selected for inclusion</p>
	
<p>2nd Draw: A B C D E F G H are ALL available</p> <p>- Thus “G” is available for inclusion a 2nd time.</p>	<p>2nd Draw: A B C D E F H are available but “G” is not available anymore</p> <p>- Thus, “G” can only be included once.</p>
Etc	etc

Example – What is the Probability of Each Equally Likely Sample?
What is the Probability of {A 1st B 2nd C 3rd}?

With Replacement	Without Replacement
	
Probability {A 1st} = 1/6	Probability {A 1st} = 1/6
	
Probability {B 2nd given A} = 1/6	Probability {B 2nd given A} = 1/5
	
Probability {C 3rd given A, B} = 1/6	Probability {C 3rd given A,B} = 1/4
Probability{sample=A,B,C}=(1/6)(1/6)(1/6)	Probability{sample=A,B,C}=(1/6)(1/5)(1/4)

**Example – How many equally likely samples are there?
Simple Random Sampling WITH Replacement**

Population

Four Queens in a deck of cards

Sampling Plan

- Draw one card at random
- Note its suit
- *Return the selected card*
- Draw one card at random
- Note its suit

Population size, N=4

Sample size, n=2

Total # samples possible = $(4)(4) = 4^2 = N^n = 16$

Probability of each sample = $\frac{1}{N^n} = \frac{1}{16}$

Here are the 16 possible samples:

(spade, spade)	(spade, club)	(spade, heart)	(spade, diamond)
(club, spade)	(club, club)	(club, heart)	(club, diamond)
(heart, spade)	(heart, club)	(heart, heart)	(heart, diamond)
(diamond, spade)	(diamond, club)	(diamond, heart)	(diamond, diamond)

What if the Order of the Sample Doesn't Matter?

*Tip! – We will see this again when we learn the **Binomial Distribution**...*

Ordered Samples - In the previous examples, the sample obtained was defined by BOTH its membership (eg – A B C) and the ORDER of its members (eg – A first, B second, C third).

Unordered Samples – In an Unordered sample, the sample is defined ONLY by its membership.

Example of an Unordered Sample –

{ A B C } is the same as
 { A C B } is the same as
 { B A C } is the same as
 { B C A } is the same as
 { C A B } is the same as
 { C B A }

How many ordered samples that include {A, B, and C} are included in the Unordered sample {A, B, C}?

The answer is obtained by answering the question: *How many **different orderings** are there of “A”, “B” and “C”?* This is the same as asking: *How many **different permutations** are there of “A”, “B” and “C”?*

Solution:

- **First position:** How many choices are possible?
Answer = 3
- **Second position, given 1st is filled:** How many choices are possible?
Answer = $(3 - 1) = 2$.
- **Third position, given 1st and 2nd filled:** How many choices are possible?
Answer = $(3-2) = 1$.
- Thus, the total number of **different orderings (permutations)** of 3 items
= $(3)(3-1)(3-2) = 6$

How many different orderings (permutations) are there of “n” items, such as “A”, “B”, …, “n”?

- **First position:** # choices = n
- **Second position, given 1st is filled:** # choices = $(n - 1)$.
- **Third position, given 1st and 2nd are filled:** # choices = $(n - 2)$,

Etc for 4th position, 5th position and so on to the nth position

- **Answer:** The number of **permutations** of n items is = $(n)(n-1)(n-2) \cdots (2)(1)$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Example – How many equally likely samples are there?
Simple Random Sampling WITHOUT Replacement
Order Does NOT Matter

Population

Four Queens in a deck of cards

Sampling Plan

- Draw one card at random
- Note its suit
- *Set the selected card aside (Do NOT return it to the pile)*
- Draw another card at random
- Note its suit

Population size, N=4

Sample size, n=2

Total # samples possible = $(4)(4-1) = (4)(3) = 12$

Probability of each sample, ordered = $\frac{1}{N(N-1)} = \frac{1}{(4)(3)} = \frac{1}{12}$

If order matters, there are 12 possible equally likely samples, each with probability = 1/(12):

(spade, club) (club, spade)	(heart, club) (club, heart)	(diamond, heart) (heart, diamond)
(heart, spade) (spade, heart)	(diamond, club) (club, diamond)	(spade, diamond) (diamond, spade)

But, if order DOES NOT matter, then there are 6 equally likely samples, each with probability = 1/6:

(spade, club)	(heart, club)	(diamond, heart)
(heart, spade)	(diamond, club)	(spade, diamond)

Putting it all together: How to Calculate the Probability of an Equally Likely Unordered sample

The solution involves two calculations of # orderings (permutations)

(1) Calculate the total number of samples, *as if you were doing ordered sampling*

$$\begin{aligned} \left(\begin{array}{l} \text{Total \#} \\ \text{ordered samples} \end{array} \right) &= \left(\begin{array}{l} \text{\# ways to} \\ \text{draw 1st} \end{array} \right) \left(\begin{array}{l} \text{\# ways to} \\ \text{draw 2nd} \end{array} \right) \\ &= (N)(N - 1) \\ &= (4)(3) \\ &= 12 \end{aligned}$$

(2) Next, calculate the *number of rearrangements of the sample obtained.*

$$\begin{aligned} \left(\begin{array}{l} \text{\# orderings of} \\ \text{given sample} \end{array} \right) &= \left(\begin{array}{l} \text{\# choices for} \\ \text{position 1} \end{array} \right) \left(\begin{array}{l} \text{\# choices for} \\ \text{position 2} \end{array} \right) \\ &= (n)(n - 1) \\ &= (2)(1) \\ &= 2 \end{aligned}$$

Pr[each equally likely UNordered sample]

$$= \frac{\text{\# rearrangements of sample obtained}}{\text{\# of ordered samples possible}}$$

$$= \frac{(n)(n-1)(n-2) \dots (2)(1)}{(N)(N-1)(N-2) \dots (N-n+1)}$$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

$$\left(\begin{array}{l} \text{Probability of} \\ \text{1 club and 1 heart} \end{array} \right) = \frac{\# \text{ orderings of a "net" result}}{\text{total \# of ordered samples}} = \frac{(n)(n-1)}{(N)(N-1)} = \frac{(2)(1)}{(4)(3)} = \frac{2}{12}$$

Simple Random Sampling **WITHOUT** Replacement Summary

IF Order **DOES** Matter

- ◆ Total # ordered samples = $N(N-1)(N-2) \cdots (N-n+1)$
- ◆ Probability [Each equally likely ordered sample] = $\frac{1}{(N)(N-1)\cdots(N-n+1)}$

IF Order Does **NOT** Matter

- ◆ Total # ordered samples = $N(N-1)(N-2) \cdots (N-n+1)$
- ◆ # rearrangements of the sample obtained = $(n)(n-1)(n-2)\cdots(2)(1)$
- ◆ Probability [Each equally likely UNordered sample] = $\frac{(n)(n-1)(n-2)\cdots(2)(1)}{(N)(N-1)\cdots(N-n+1)}$
- ◆ The total # of UNordered samples is the reciprocal of this.

$$\text{Total \# UNordered samples} = \frac{(N)(N-1)(N-2)\cdots(N-n+1)}{(n)(n-1)(n-2)\cdots(2)(1)}$$

b. How to Select a Simple Random Sample **WITHOUT Replacement**
(Using a random number table)

<p><u>Step 1:</u></p> <p>List the subjects in the sampled population.</p> <ul style="list-style-type: none"> ♣ This is the <u>sampling frame</u>. 	<p><u>Example – Obtain a simple random sample of n=30 from a population of size N=500-</u></p> <ul style="list-style-type: none"> ♣ Make a list of all N=500
<p><u>Step 2:</u></p> <p>Number this listing from “1” to “N”</p> <ul style="list-style-type: none"> ♣ where N = size of sampled population 	<p><u>Example, continued -</u></p> <p style="text-align: center;">N = 500 n = 30</p>
<p><u>Step 3:</u></p> <p>The size of “N” tells you how many digits in a random number to be looking at:</p> <ul style="list-style-type: none"> ♣ For $N \leq 10$ Need only read 1 digit ♣ For N between 10 and 99 Read 2 digits ♣ For N between 100 and 999 Read 3 digits <p style="text-align: center;">etc</p>	<p><u>Example, continued –</u></p> <p>Since N=500 is between 100 and 999 and is 3 digits long.</p> <ul style="list-style-type: none"> ♣ Read 3 digits

Step 4:

Using the random number table, pick a random number as a starting point

79889	75532	28704
48895	11196	34335
89604	41372	10837



Example, continued -

The first 3 digits of this number is “111”. So we will include the 111th subject in our sample

Step 5:

Proceed down your selected column of the random number table, row by row.

With each row, if the required digits are $\leq N$, INCLUDE

With each row, if the required digits are $> N$, PASS BY

With each row, if the required digits are a repeat of a previous selection, PASS BY

79889	75532	28704
48895	11196	34335
89604	41372	10837



Example, continued -

The first 3 digits of the second random number is “413”. So we will include the 413th subject in our sample

Step 6:

Repeat “Step 5” a total of n=30 times, which is your desired sample size.



Remarks on Simple Random Sampling

Advantages:

- Selection is entirely left to chance.
- Selection bias is still possible, but chances are small.
- No chance for discretion on the part of the investigator or on the part of the interviewers.
- We can compute the probability of observing any one sample. This gives a basis for statistical inference to the population, our ultimate goal.

Disadvantages:

- We still don't know if a particular sample is representative
- Depending upon the nature of the population being studied, it may be difficult or time-consuming to select a simple random sample.
- An individual sample might have a disproportionate # of skewed values.

Nature ——— Population/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

9. Some Other Probability Sampling Plans

a. Systematic Sampling

- Population size = N . Desired sample size = n .
- Desired is an $(n/N) = X$, or a $100X\%$ sample
- Pick the first item by simple random sampling. Thereafter, select every $(1/X)$ th item

Example:

Suppose $n=20$ is desired from a population of size $N=100$

→ $(n/N) = 20/100 = .05$ means we want a 5% sample, obtained by systematic sampling.

The first individual is selected by simple random sampling (so chances of inclusion are 1 in 100)

Thereafter, take every 20th individual (so chances are then 0 or 1 depending on position in list!!)

Example –

Suppose we want a sample of size $n=100$ from the $N=1000$ medical charts in a clinic office.

Pick the 1ST chart by simple random sampling.

$$n/N = 100/1000 = .10 \rightarrow 1/.10 = 10$$

Thereafter, select every 10th chart.

Remarks on Systematic Sampling

Advantages:

- It's easy.
- Depending on the listing, the sampled items are more evenly distributed.
- As long as there is no association with the order of the listing and the characteristic under study, this should yield a representative sample. **Note – This was not the case in the 1948 Gallup Poll!**



Disadvantage:

- If the sampling frame has periodicities (a regular pattern) and the rule for systematic sampling happens to coincide, the resulting sample may not be representative.

Example of a Periodicity that Results in a Biased Sample:

- Clinic scheduling sets up 15 minute appointments with physicians
- Leaves time for an emergency, or walk-in visit at 15 minutes before the hour, every hour.
- Doing a chart-audit, you sample every 4th visit and get only the emergency visits selected into the sample, or else none of them.

b. Stratified Sampling
Simple Random Sampling within Strata

Example-

Do construction workers experience major health problems?

Do health problems differ among males and females?

Construction workers, as a group, are likely to be comprised predominately of males.

Thus, if we take a simple random sample we may get very few women in the sample.

Procedure:

- 1. Define mutually exclusive strata such that the outcome of interest is likely to be**

similar within a stratum; and very different between strata.

outcome:	health problems
strata:	Males / Females

- 2. Obtain a simple random sample from each stratum**

We want to be sure to get a good overall sample.
Sampling each stratum separately ensures this.



Remarks on Stratified Sampling

Advantage:

- Good when population has high variability, especially when the population includes a mix of people (eg. males and females) that are NOT similarly represented (eg. population is disproportionately male)

Take care:

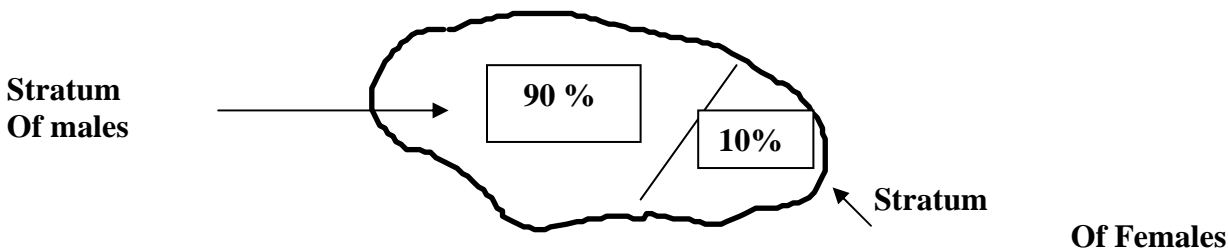
- Strata MUST be mutually exclusive and exhaustive
- To compute an overall population estimate requires use of weights that correspond to representation in the population. Following is an example.

Example of Calculation of Weighted Mean from a Stratified Sample -

Goal – To estimate the average # cigarettes smoked per day among all construction workers.

Population is disproportionately male (90% male, 10% female)

- ♣ Weight given to average observed for males = 0.90
- ♣ Weight given to average observed for females = 0.10
- ♣ Note that weights total 1.00



$$\left[\begin{array}{l} \text{Weighted} \\ \text{average, } \bar{X}_w \end{array} \right] = \left(\begin{array}{l} \text{weight} \\ \text{males} \end{array} \right) (\bar{X}_{\text{males}}) + \left(\begin{array}{l} \text{weight} \\ \text{females} \end{array} \right) (\bar{X}_{\text{females}})$$

c. Multi-Stage Sampling

Good, Sometimes essential, for Difficult Populations

Example Suppose we want to study a gypsy moth infestation.

A **multistage** sample plan calls for

- 1ST - Select individual trees
(Primary sampling units - PSU's)
- 2nd - Select leaves from only the selected trees
(Secondary sampling units)

Multistage Sampling

- The selection of the primary units may be by simple random sampling
- The selection of the secondary units may also be by simple random sampling
- Inference then applies to the entire population

CAUTION!!!

- Take care that the selection of primary sampling units is NOT on the basis of study outcome. Bias would result.



10. The Nationwide Inpatient Survey (NIS)
Sampling Designs Can Be Quite Complex

Target Population
 All discharges in all community hospitals in the US



NIS Sampling Frame
 All community hospitals in *participating states* that actually release data.



Binning into strata defined by: geographic area, control, location teaching status, bedsize. Result is $4 \times 3 \times 2 \times 2 \times 3 = 144$ strata.



Stratum #1
 = bin of NIS frame



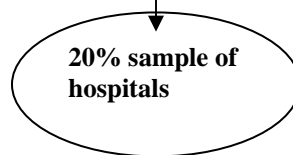
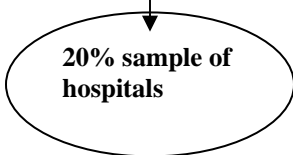
Stratum #144
 = bin of NIS frame

(Sort by state and zip code)

(Sort by state and zip code)



Systematic random sample. Goal = 20% (i.e. every 5th)	Systematic random sample. Goal = 20% (i.e. every 5th)
--	-------	--



All discharges



All discharges