

4. Populations and Samples

Topics	1. A Feeling for Populations v Samples 3 2. Target Populations, Sampled Populations, Sampling Frames 6 3. On Making Inferences from a Sample 9 4. Simple Random Sampling 11 5. Some Non-Probability Sampling Plans 14 6. Introduction to Probability Sampling 17 a. Sampling WITH v WITHOUT replacement 18 b. How to select a simple random sample 26 7. Some Other Probability Sampling Plans 29 a. Systematic 29 b. Stratified 31 C. Multi-stage 33 8. The Nationwide Inpatient Survey (NIS) 34
---------------	---

Please Quiet Cell Phones and Pagers

Thank you.

1. A Feeling for Populations versus Samples

So far we've seen

- some examples which illustrate that statistics is a tool for evaluating uncertainty and that we must be wary of biases that might enter our thinking;
- how to summarize data in graphs and tables; and
- how to summarize data in numbers

We also remarked that, ultimately, statistics is about

- learning about and describing the population which gave rise to the sample (relevance), or;
- investigating some research question (question);
- assigning probabilities of events (mathematical tool);
- calculating the likelihood of the observed events under some specified postulates (statistical significance);
- *And we're careful - Statistical significance is not the same as biological significance*

Thus, it helps to have a feel for populations versus samples.

Example – The 1948 Gallup Poll

- Before the 1948 presidential election, Gallup polled 50,000
- 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 $n=50,000$
- A feeling for populations versus samples

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

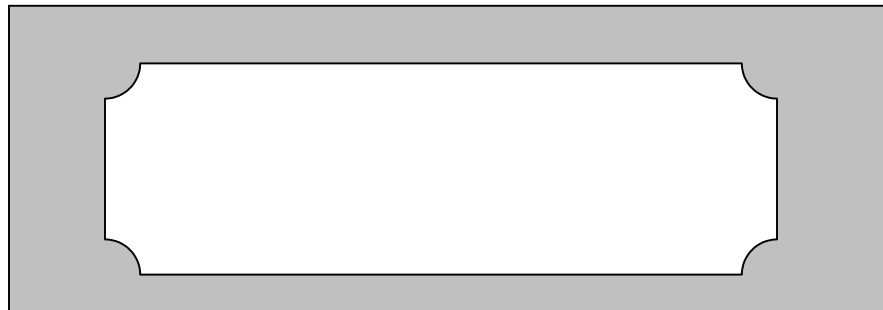
- those with telephones AND ...

SECOND: The over-sampled included a disproportionate number those favoring votes for Dewey. I.e., there was over-sampling in the segment of the population more likely to vote for Dewey

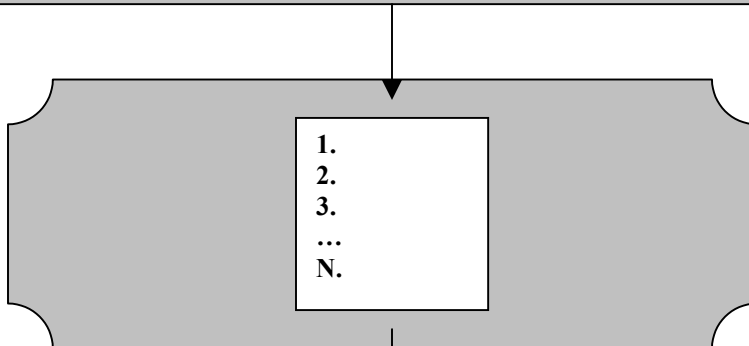
Bias occurred not only because of over-sampling itself but also because the nature of the over-sampling was related to voter preference.

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

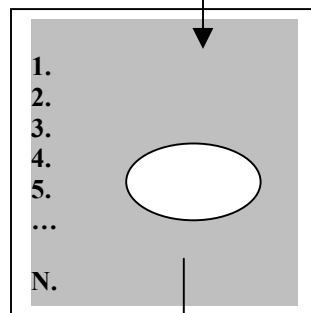
2. Target Population, Sampled Population, Sampling Frame



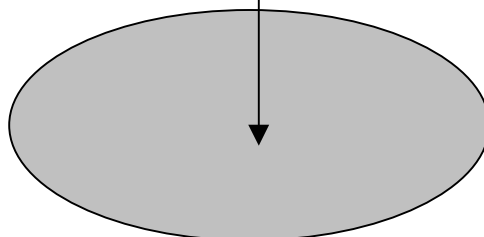
Target Population
The whole group of interest.



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.

Target Population	<ul style="list-style-type: none">• The aggregate of individuals who are of interest.• <u>Example</u> – The population of actual 1948 presidential election voters.
Sampled Population	<ul style="list-style-type: none">• The aggregate of individuals that was actually sampled.• A listing of the entire sampled population comprises the <u>sampling frame</u>.• 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• <u>Example</u> – 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♣ Consenting♣ 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)**

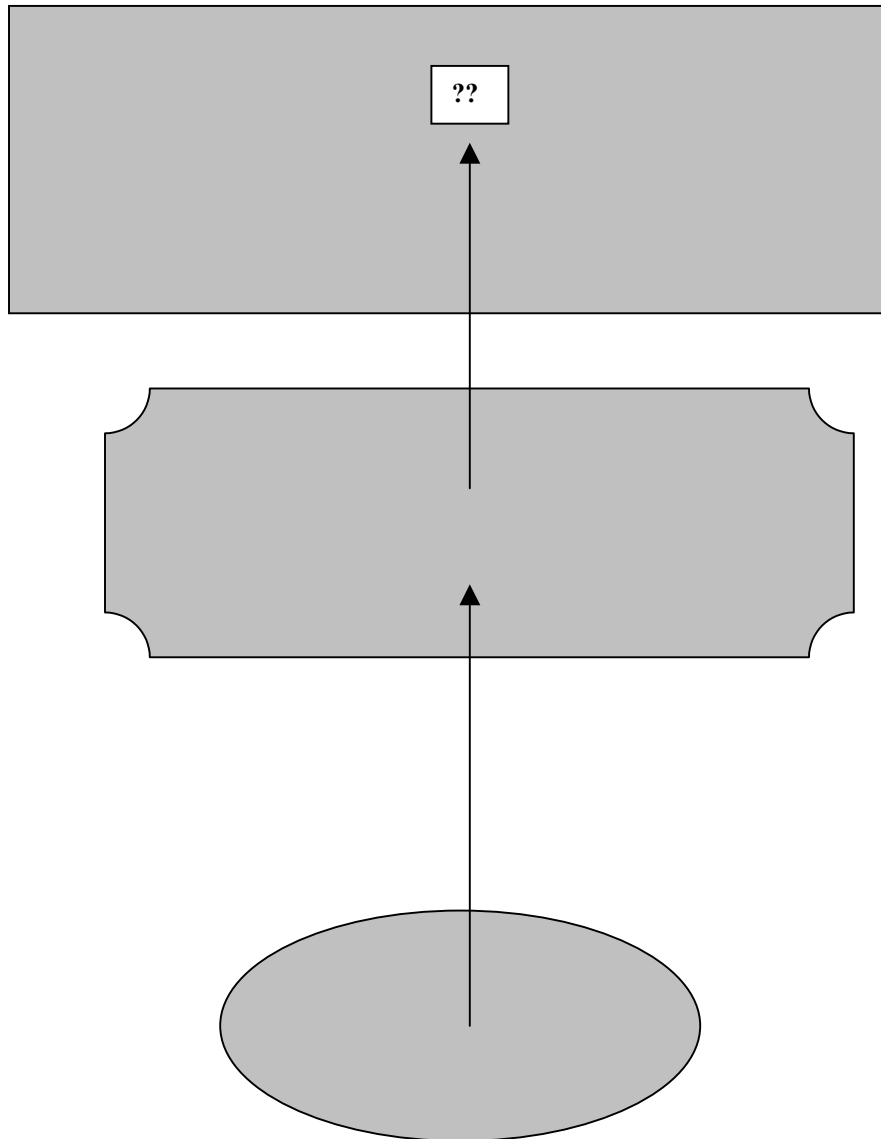
Example –

- ♣ **The League of Women's Voters Registration List might be the sampling frame for the target population who vote in the 2002 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?**

3. On Making Inferences From a Sample (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 applies 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 defined the sampled population

Example -

- ♦ The sampling frame, by definition, is allowed to contain only consenters
- ♦ Therefore, refusers, by definition, are not in the sampling frame.
- ♦ Thus, the sampled population differs from the target population in at least one respect, ALWAYS.
- ♦ 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.

4. Simple Random Sampling

We would like our sampling plan to:

- ◆ permit generalization to the target population “in the long run” (unbiased);
- ◆ permit inferences that are not in error by too much, “in the short run” (minimum variance);

The virtue of simple random sampling:

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

“In the long run? What about here and now?”

An individual sample will likely produce an estimate value that is NOT equal to the population value. It is for this reason that we seek out summary statistics that aren't too noisy (more on this in Unit 7 – Estimation).

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

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 Protocol

- ♦ **Include in study sample 3 subjects drawn from the population at random and without replacement.**

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 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

$$\frac{\sum \text{sample } \bar{X}}{\text{all 20 possible samples}} = \mu = 26$$

$$\sum_{\text{sample}\#1}^{\text{sample}\#20} [\text{error}] = 0$$

5. Some Non-Probability Sampling Plans

- ◆ Quota
- ◆ Judgment
- ◆ Volunteer

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%)**

Judgment Sampling Plan

Decision regarding inclusion or non-inclusion is 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

Volunteer 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, since it is based on some sort of probability model, does not apply.**

- 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**

6. Introduction to Probability Sampling Plans

The way to go.

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

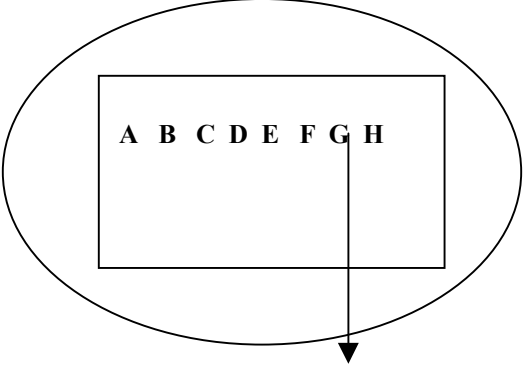
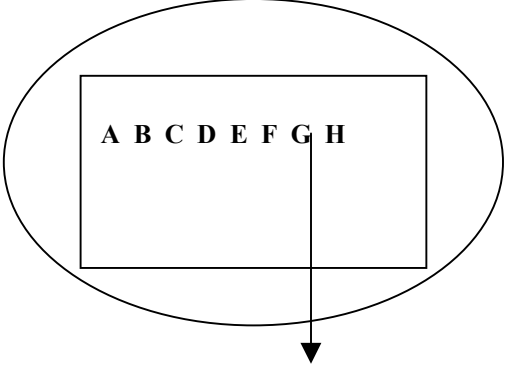
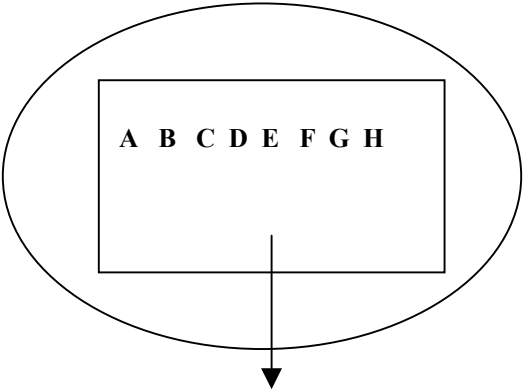
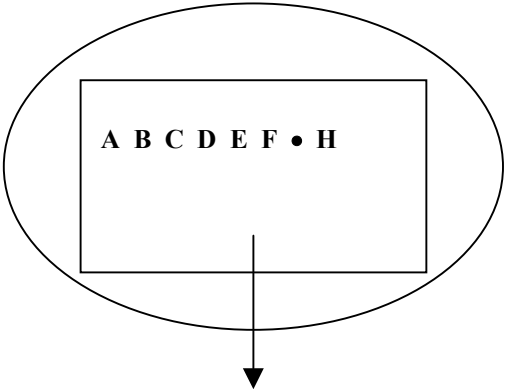
- ◆ **Each individual has a known probability of inclusion in the sample, prior to sampling.**
- ◆ **The investigator has no discretion regarding the inclusion or exclusion of an individual**
- ◆ **This eliminates one source of potential bias – that on the part of the investigator.**

How do we know if a sample is REPRESENTATIVE?

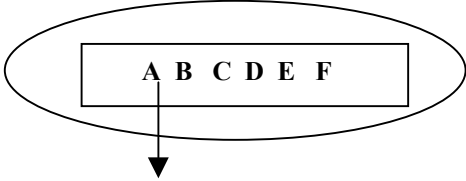
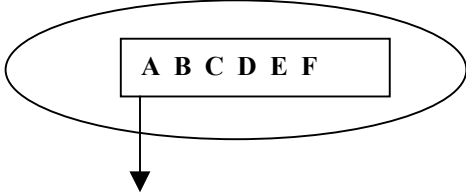
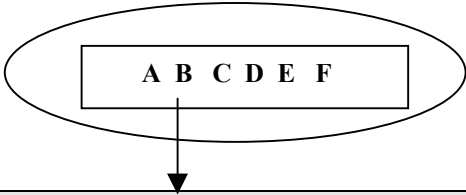
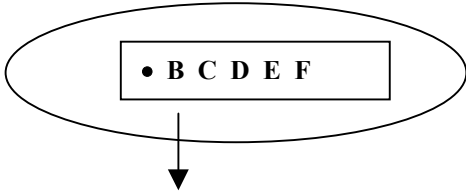
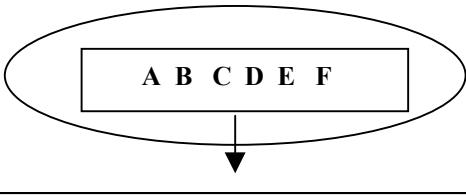
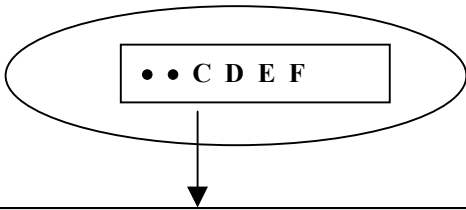
- ◆ **Ultimately, we don’t know.**
- ◆ **So, instead, we use an unbiased sampling plan and hope for the best.**
- ◆ **In the meantime, we can generate some descriptive statistics and compare these to what we know about the population.**

Simple Random Sampling is the basic probability sampling method.

Sampling With Replacement
V
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>
<p style="text-align: center;">Etc</p>	<p style="text-align: center;">etc</p>

**Example –
What is the Probability of {A B C}?**

With Replacement	Without Replacement
	
Probability {A} = 1/6	Probability {A} = 1/6
	
Probability {B given A} = 1/6	Probability {B given A} = 1/5
	
Probability {C given A, B} = 1/6	Probability {C 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)

Sampling With versus Without Replacement General

With Replacement	Without Replacement
Population size = N Sample size = n	Population size = N Sample size = n
Each individual has a 1/N chance of inclusion in the sample.	Each individual has a 1/N chance of inclusion in the sample, overall.
<p>Total # possible samples of size = n is</p> $\begin{array}{ccccccc} (N) & (N) & (N) & \dots & (N) & = & N^n \\ \uparrow & \uparrow & \uparrow & & \uparrow & & \\ 1^{\text{st}} & 2^{\text{nd}} & 3^{\text{rd}} & & \text{nth} & & \\ \text{draw} & \text{draw} & \text{draw} & & \text{draw} & & \end{array}$	<p>Total # possible samples of size = n is</p> $\begin{array}{ccccccc} (N) & (N-1) & (N-2) & \dots & (N-n+1) \\ \uparrow & \uparrow & \uparrow & & \uparrow & & \\ 1^{\text{st}} & 2^{\text{nd}} & 3^{\text{rd}} & & \text{nth} & & \\ \text{draw} & \text{draw} & \text{draw} & & \text{draw} & & \end{array}$
<p>Probability {each sample of size n} =</p> $\frac{1}{N^n}$	<p>Probability {each sample of size n} =</p> $\frac{1}{N(N-1)(N-2) \dots (N-n+1)}$

Example Sampling WITH Replacement

Population

Four Queens in a deck of cards

Sampling Plan

5. Draw one card at random
6. Note its suit
7. *Return the selected card*
8. Draw one card at random
9. 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
This applies to the Binomial Distribution to Come...

Example –

$\{ A \ B \ C \}$ 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 ways are there to order “A”, “B”, and “C”

- # letter choices for the first position = 3
- Having specified a selection for first position,
letter choices for the second position = $(3 - 1) = 2$.
- Having specified selections for the first and second positions,
letter choices for the third position = $(3 - 1 - 1) = 1$.
- Therefore the # ways to order “A”, “B”, “C” = $(3)(3-1)(3-2) = 6$

How many ways are there to order “A”, “B”, ..., “n”?

- # letter choices for the first position = n
- Having specified a selection for first position,
letter choices for the second position = $(n - 1)$.
- Having specified selections for the first and second positions,
letter choices for the third position = $(n - 2)$, etc
- Therefore the # ways to order “A”, ..., “n” = $(n)(n-1)(n-2) \cdots (2)(1)$

This is called the number of permutations of 1, 2, ..., n

Example
Sampling WITHOUT Replacement
Order Does NOT Matter

Population

Four Queens in a deck of cards

Sampling Plan

10. Draw one card at random
11. Note its suit
12. *Set the selected card aside (Do NOT return it to the pile)*
13. Draw another card at random
14. 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 = $\frac{1}{N(N-1)} = \frac{1}{(4)(3)} = \frac{1}{12}$

Here are the 12 possible ordered samples:

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

Notice that each “net” result of two cards has two possible orderings.

By inspection, we can see easily that
 Probability {1 club and 1 spade, regardless of order} = 2/12

More generally –

Total # of ordered samples = $\frac{N!}{(N-2)!}$ = # ways to draw 1st \times # ways to draw 2nd

$$= n(n-1)$$

$$= (4)(3)$$

$$= 12$$

orderings of given sample = $\frac{2!}{(2-2)!}$ = # choices for position 1 \times # choices for position 2

$$= 2(1)$$

$$= (2)(1)$$

$$= 2$$

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

Sampling WITHOUT Replacement Summary

IF Order DOES Matter

- ◆ **Total # ordered samples = $N(N-1)(N-2) \cdots (N-n+1)$**
- ◆ **Probability of each ordered sample = $\frac{1}{(N)(N-1)\cdots(N-n+1)}$**

IF Order Does NOT Matter

- ◆ **Total # ordered samples = $N(N-1)(N-2) \cdots (N-n+1)$**
- ◆ **# ways to order (“permute”) each “net” result of size $n = (n)(n-1)(n-2)\cdots(2)(1)$**
- ◆ **Total # of Unordered samples =**

$$\frac{\text{Total \# ordered samples}}{\text{\# orderings of 'net' result}} = \frac{(N)(N-1)(N-2)\cdots(N-n+1)}{(n)(n-1)\cdots(2)(1)}$$

- ◆ **Probability of each “net” unordered sample = $1/(\text{total \# unordered samples}) =$**

$$\frac{\text{\#orderings of the "net" result}}{\text{Total \# ordered samples}} = \frac{(n)(n-1)(n-2)\cdots(2)(1)}{(N)(N-1)\cdots(N-n+1)}$$


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</u> -</p> <p>Select a simple random sample of 30 subjects from a population of 500</p>
<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 <li style="text-align: center;">etc 	<p><u>Example, continued</u> –</p> <p>N=500 is between 100 and 999</p> <ul style="list-style-type: none"> ♣ For N between 100 and 999 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” until you have included the required number ($n=30$) in your sample

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 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.

7. Some Other Probability Sampling Plans

- ♦ **Systematic**
- ♦ **Stratified**
- ♦ **Multi-stage**

Systematic Sampling

- **Population 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: $n=20, N=100 \rightarrow (n/N) = 20/100 = .05$
or $X=.05$, for a $100(.05)=5\%$ sample \rightarrow select $1/.05=20$, or every 20th item
- **Probability an item is included on the 1ST draw = $1/N = .01$**
- **Given the 1ST item has been drawn, probability an item is included on any other draw is 0 or 1.**

Example –

Suppose we want a sample of size $n=100$ from the 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.

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.

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 is likely to be comprised predominately of males.

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:

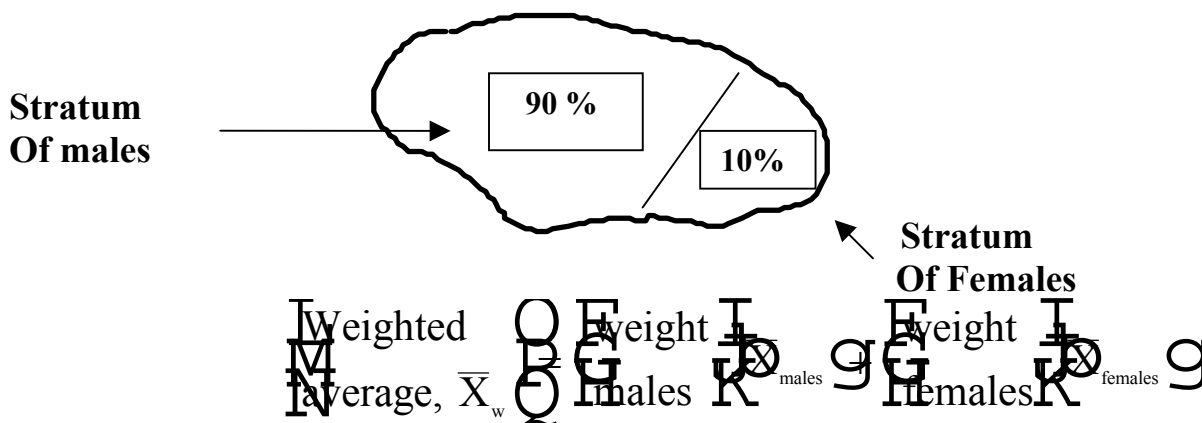
- Strata MUST be mutually exclusive and exhaustive
- To compute an overall population estimate requires use of weights that correspond to representation in the population. Following is an example.

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



Multi-Stage Sampling
Good, Sometimes essential, for Difficult Populations

Example Suppose we want to study the 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.

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

