Unit 3
Populations and Samples

"To all the ladies present and some of those absent"

- Jerzy Neyman

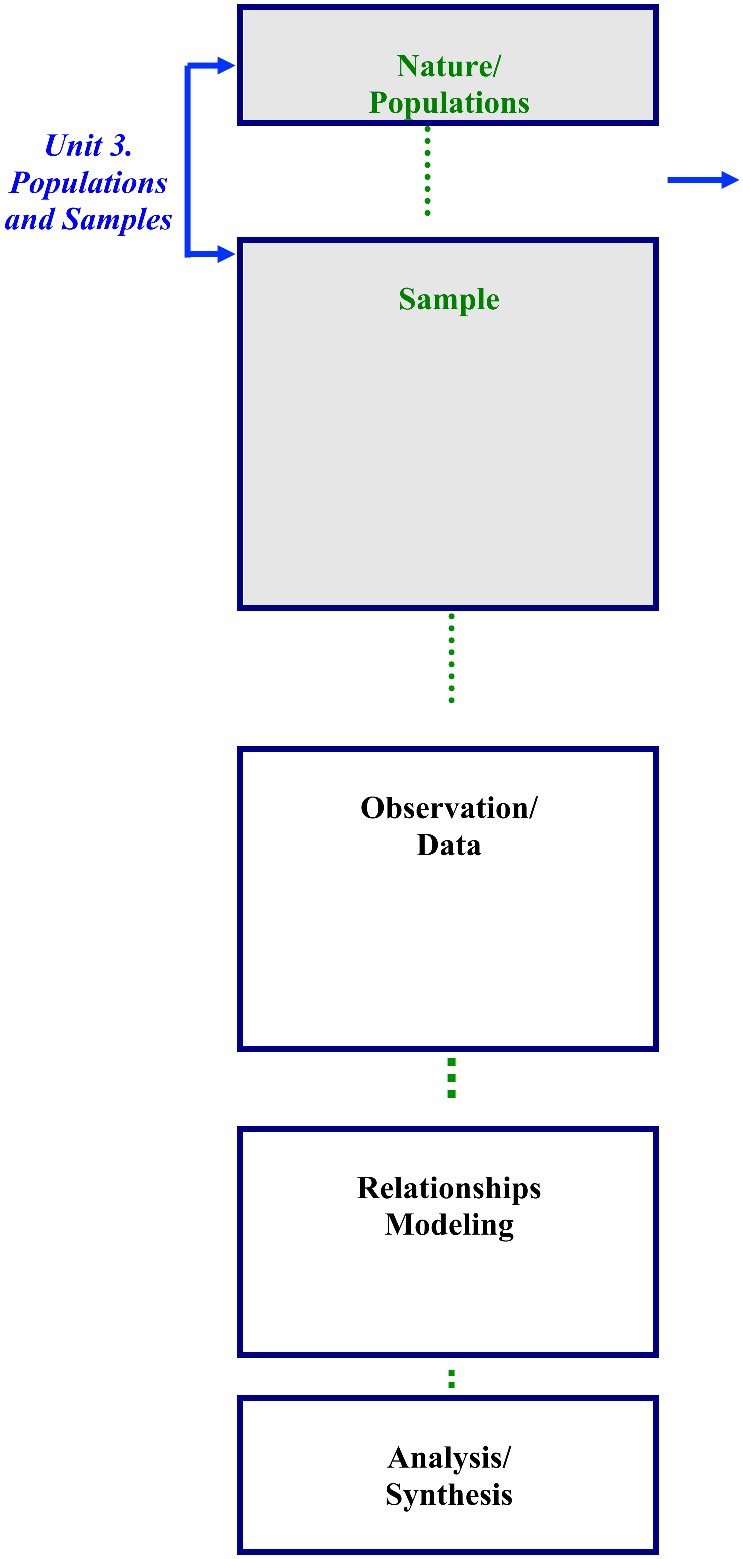
The collection of all individuals with HIV infection and the collection of all individuals with exposure to mercury are examples of ***populations*** about which we wish to make inferences. A ***census*** involves the collection of information on ***every individual*** in the population and is one way to obtain information about a population. How nice! No statistical analyses are required! Unfortunately, though ideal, censuses are usually impractical because we lack the necessary resources for their conduct.

Thus, we typically study instead a subset of the population, called a ***sample.*** Statistical analyses are required if we want to make meaningful inferences about the population. There are lots of ways to obtain a sample; these are called ***sampling designs***. Perhaps the most familiar is the method of ***simple random sampling.*** Loosely, simple random sampling is sampling at random without replacement from the population

The goal of a ***sampling design*** and the statistical analyses that follow are: (1) to obtain a sample with a known probability of selection and for which the conclusions drawn are (2) in the long run correct (unbiased) and (3) in the short run in error by as little as possible (minimum variance).

Table of Contents

1. Unit Roadmap



Take another look at the roadmap at the footer of this page. A study begins with a sample from a population (highlighted here in bold red). → From our sample, we make some observations and record some data. → From there, with the use of an assumed model, we estimate some things (for example – average response to treatment) and test some hypotheses (for example – the new treatment is no better than the control treatment).

In the real world, we have just the ***one sample*** and no luxury to repeat the study over and over. So we rely on the properties of the sampling procedures (for example – all theoretically possible samples were equally likely to have been selected) as the justification for the conclusions we draw.

Unbiased - If we were to repeat our study over and over again, the average of our sample (for example – the average response to treatment) will eventually settle down to a long range average of the average that is equal to the true population average.

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.

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.
- Explain the rationale for systematic, stratified, and multi-stage sampling methods.
- Define systematic, stratified, and multi-stage sampling.

Interested readers, reading the appendix, will also have a feel for:

- The distinction between sampling with versus without replacement.

3. A Feeling for Populations versus Samples

In **unit 1**, our goal was to summarize (and communicate effectively!) the information in a given sample. We didn't concern ourselves with the source of the sample meaning, specifically, the population from which the sample was obtained. We learned about various kinds of summaries (graphical and numerical).

In **unit 2**, we extended our perspective to the population from which the sample came. This unit was an introduction to population probability distributions. Just one, however. We limited our attention to the case of a discrete uniform probability distribution. In such populations, every individual (the "elementary outcome") has the same probability of selection. We learned the names for various kinds of sampling results (elementary outcomes and events). We also learned some tools of basic probabilities that are useful when dealing with various kinds of outcomes: mutually exclusive, dependent/conditional, independent, etc.

In **unit 3**, we will put the two together: **population** and **sample**. A sample is obtained as the result of following a sampling procedure. The nature and specifics of the sampling procedure is called the **sampling design**.

Meaningful statistical inference requires that the sample studied be a probability sample.

- Population – The collection of all the individuals of interest.
- Probability Sampling Design – The rules of probability that govern the likelihood of each sample being selected
- Sample – The subset of the population that is selected as the result of sampling.

Non-representative sampling tends to produce study conclusions that are incorrect.

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

- Not every sample of size 50,000 had an equal chance of being selected by Gallop.

- The actual sample of 50,000 was an under-sampling of some types of voters and an over-sampling of other types of voters.

- The conclusion drawn from the actual sample was an incorrect prediction of the behavior of the population of actual voters. The error in the prediction was the result of two things:

FIRST: The interviewers over-sampled
1. wealthy
2. safe neighborhoods, those with telephones WHICH IS RELEVANT BECAUSE . . .

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

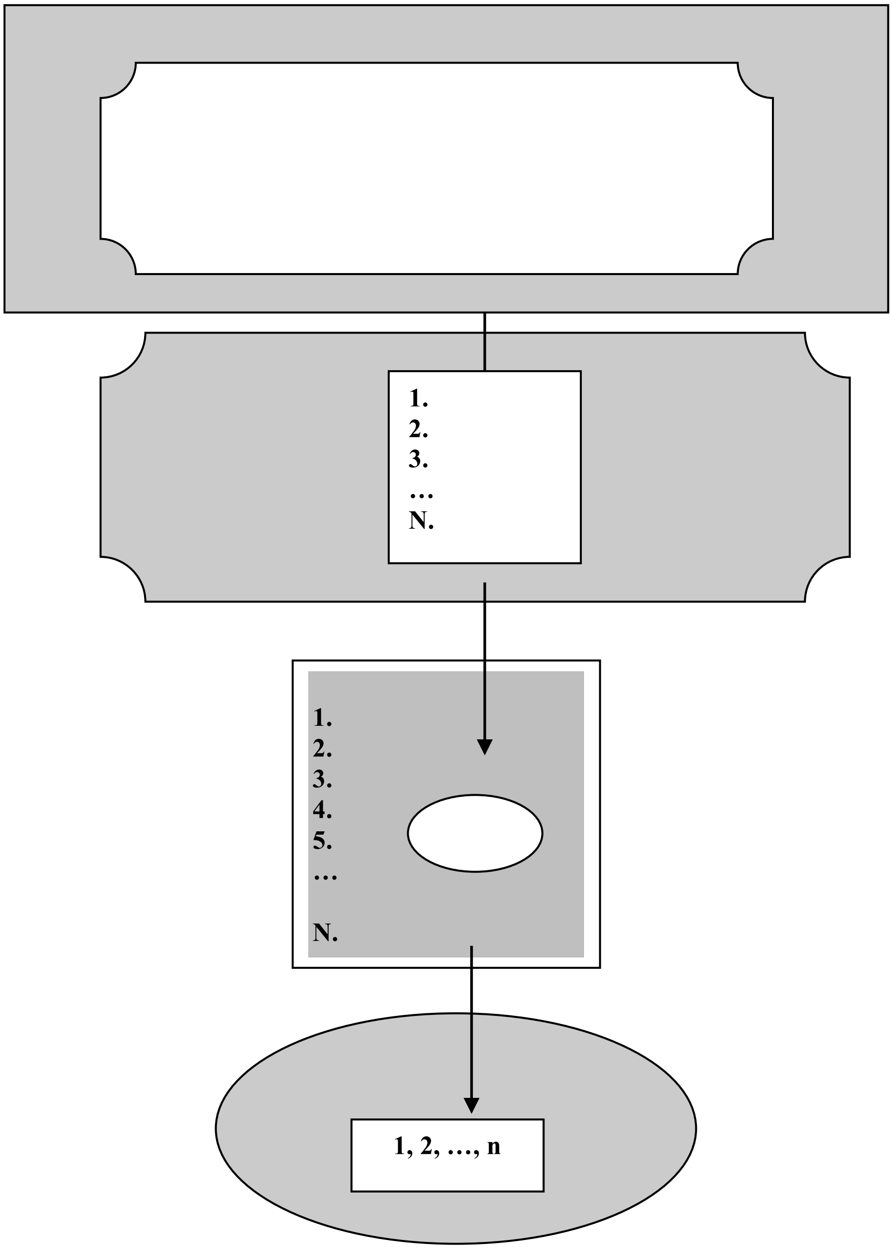
Bias occurred because of two issues:

(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. **So, yes. The sampling frame and the sampled population are the same thing. The one that has been put into a list is the sampling frame.**

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.

Target Population	• The entire collection of individuals who are **of interest**. • **Example** – The population of 1948 presidential election voters who actually voted.
Sampled Population	• 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: Whom 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 ♣ 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 will vote in the 2016 election.

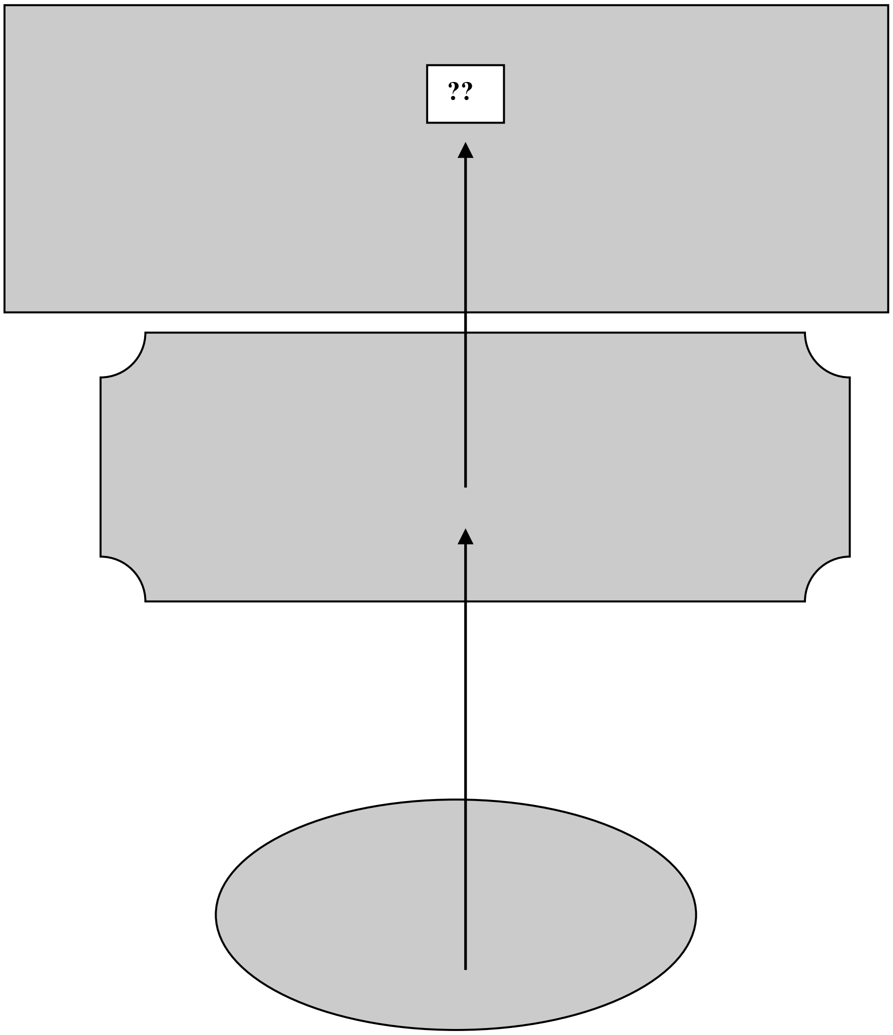
- Individual identification might be the position on this list.

Now You Try –

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

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 a randomized trial protocol that includes consent, the sampled population differs from the target population because the sampled population 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 produce estimates that are:

- **Unbiased** – If sampling is repeated over and over and over, the "long run" average conclusion about the population should be correct**.**

- **Minimum variance** – The discrepancies **(and their variance)** between the conclusions drawn from the separate samples versus what is true in 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 (equal) 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 $\overline{X}$ as our guess of μ,
so as to compile a collection of sample estimates $\overline{X}$,

- THEN "in the long run"…
the average of all the sample estimates, average of ($\overline{X}$ after $\overline{X}$ after $\overline{X}$ …)
will be equal to the population parameter value (the true value of μ)

Example of Simple Random Sampling, without replacement

"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

In this (admittedly tiny) illustration, we can actually calculate the value of the population mean parameter. This is handy, as we'll refer back to it later for the purposes of demonstrating unbiased:

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

The Investigator Executes the Following Sampling Procedure

- Draw a random sample of n=3 subjects from the population. Moreover, 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.**

Calculate the sample mean, over and over again, once for each possible sample:

- For each sample, Sample mean $\overline{X} = \dfrac{\text{1st value } + \text{ 2nd value } + \text{3rd value}}{n=3}$
- In this illustration **(but not in real life because, remember, in real life the investigator does not know μ)** we can calculate the error of each $\overline{X}$ as an estimate of μ by computing

$$\text{error} = (\ \mu\ \text{-}\ \overline{X}\ ) = (26 - \overline{X})$$

How many different samples of size n=3 are possible from a population of size N=6, when the sampling design calls for simple random sampling without replacement? To understand what it meant by a sampling design being unbiased, we want to consider all these samples. The answer is that there are 20 possible samples.

The table below shows all 20 samples. For each, it shows the three sampled values of age (2nd column), the sample average age (3rd column) and the discrepancy ("error") between the sample mean and the population mean (3rd column):

Sample #	Sample (each of n=3)	Sample mean, $\bar{X}$	Error = μ- $\bar{X}$ = (26 - $\bar{X}$)
1	**{ 21, 22, 24}**	**22.333**	**+3.667**
2	**{ 21, 22, 26}**	**23**	**+3**
3	**{ 21, 22, 27}**	**23.333**	**+2.667**
4	**{ 21, 22, 36}**	**26.333**	**-0.333**
5	**{ 21, 24, 26}**	**23.667**	**+2.333**
6	**{ 21, 24, 27}**	**24**	**+2**
7	**{ 21, 24, 36}**	**27**	**-1**
8	**{ 21, 26, 27}**	**24.667**	**+1.333**
9	**{ 21, 26, 36}**	**27.667**	**-1.667**
10	**{ 21, 27, 36}**	**28**	**-2**
11	**{ 22, 24, 26}**	**24**	**+2**
12	**{ 22, 24, 27}**	**24.333**	**+1.667**
13	**{ 22, 24, 36}**	**27.333**	**-1.333**
14	**{ 22, 26, 27}**	**25**	**+1**
15	**{ 22, 26, 36}**	**28**	**-2**
16	**{ 22, 27, 36}**	**28.333**	**-2.333**
17	**{ 24, 26, 27}**	**25.667**	**+0.333**
18	**{ 24, 26, 36}**	**28.667**	**-2.667**
19	**{ 24, 27, 36}**	**29**	**-3**
20	**{ 26, 27, 36}**	**29.667**	**-3.667**

We have two ways of seeing that this sampling design is **unbiased**: 1) the average of the sample means is equal to the population mean; and 2) the errors "balance out" to zero.

(1) The **average of the sample averages** $\bar{X}$, taken over all 20 possible samples, is $\mu = 26$.

(2) The **sum of the errors, (μ - $\bar{X}$)** = (26 - $\bar{X}$), is 0.

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

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 the 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.

8. Some Other Probability Sampling Plans

a. Systematic Sampling

- Population size = N. Desired sample size = n.

- The percent of the population we want in our sample is thus p = (n/N).
  Thus, we want a (p)(100) % sample

- Step 1: Pick the first item by simple random sampling.
  Steps 2 onward: Thereafter, select every $(N/n)^{th}$ item

Example:
Suppose n=20 is desired from a population of size N=100.
→ This is a 20% sample, since p = (n/N) = (20/100)= .20 or 20%.
Step 1: Pick the first individual by simple random sampling
Steps 2 onward: Thereafter, select every (N/n)th = (100/20) = 5th individual by systematic sampling.
The first individual is selected by simple random sampling (so chances of inclusion are 1 in 100)
Thereafter, take every 5th 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.
→ This is a 10% sample, since p = (n/N) = (100/1000)= .10 or 10%.
Step 1: Pick the first individual by simple random sampling
Steps 2 onward: Thereafter, select every (N/n)th = (1000/100) = 10th individual by systematic sampling.
That is, 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 (e.g. males and females) that are NOT similarly represented (eg. population is disproportionately male)

Take care:

- Strata MUST be mutually exclusive and exhaustive

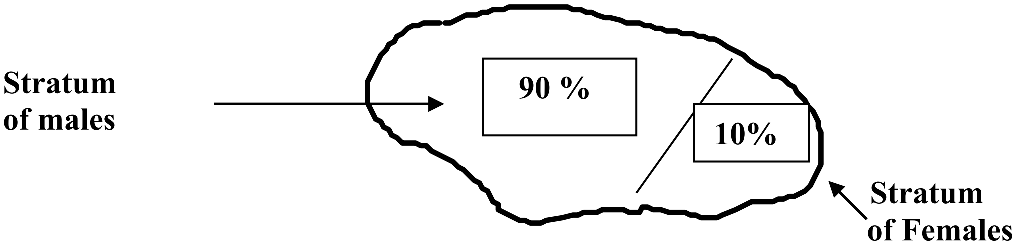
- To compute an overall population estimate requires use of weights that correspond to representation in the population. Following is an example.

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)

- ♣ Since males are 90% of the population, let's weight the average observed for males w = 0.90
- ♣ Since females are 10% of the population, let's weight the average observed for females = 0.10
- ♣ Note that weights total 1.00



$$\begin{bmatrix}\text{Weighted}\\ \text{average, } \bar{X}_w\end{bmatrix} = \begin{pmatrix}\text{weight}\\ \text{males}\end{pmatrix}\left(\bar{X}_{\text{males}}\right) + \begin{pmatrix}\text{weight}\\ \text{females}\end{pmatrix}\left(\bar{X}_{\text{females}}\right)$$

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

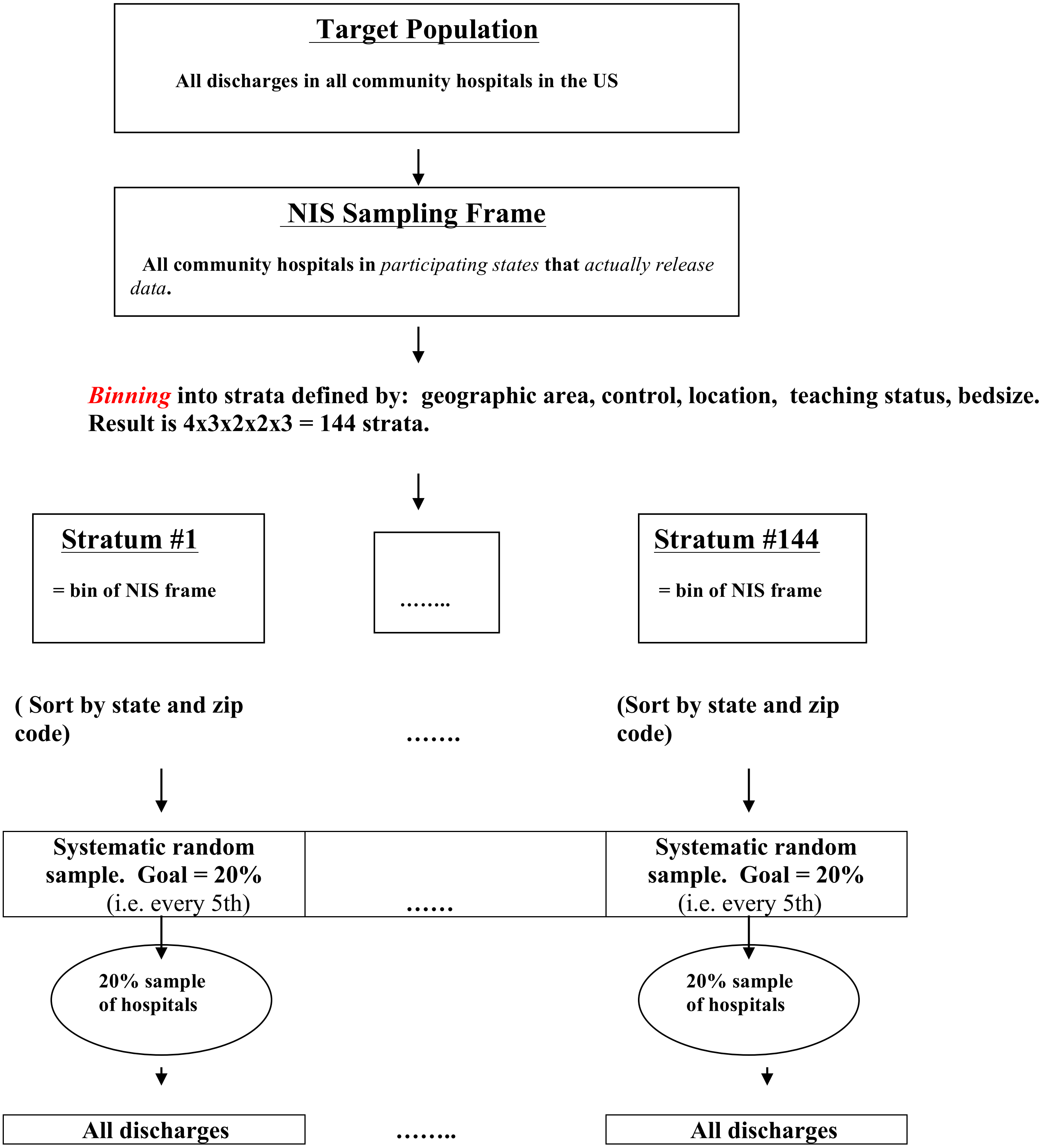
•Inference then applies to the entire population

CAUTION!!!

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

9. 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 4x3x2x2x3 = 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)

20% sample of hospitals

20% sample of hospitals

All discharges

……..

All discharges

Appendix
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 (equal) chance of being included in the sample.**

Under simple random sampling

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

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

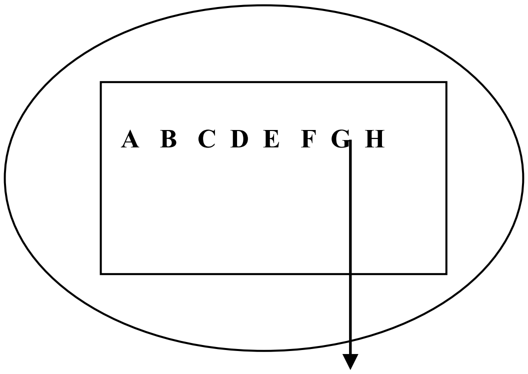
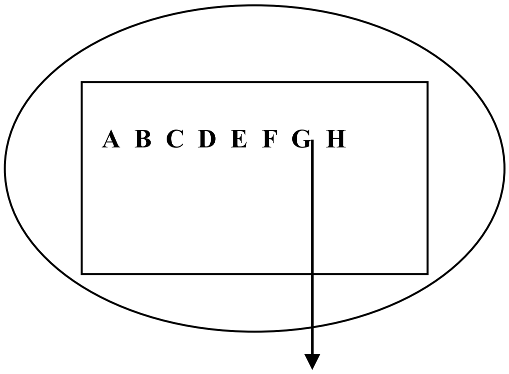
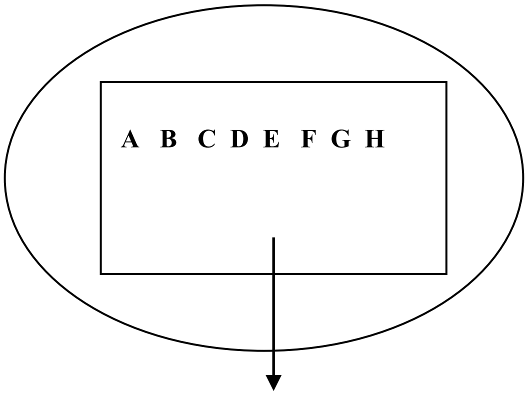
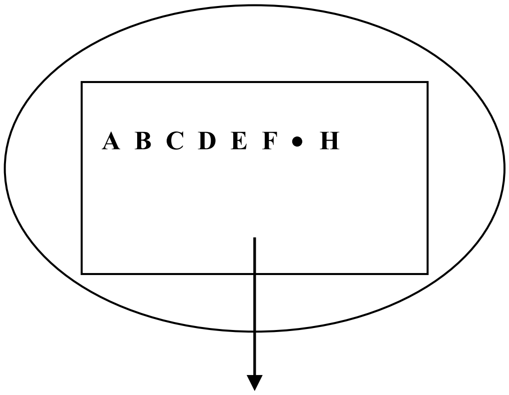
There are two kinds of sampling: WITHOUT replacement and WITH replacement

- **Example of selection without replacement** – You are selected to participate in a survey. You can only be surveyed once .

- **Example of selection with replacement** – You play the lottery multiple times and so you are available for selection multiple times (not sure this is a great example …).

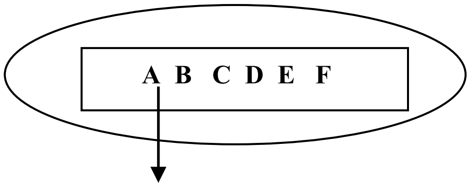
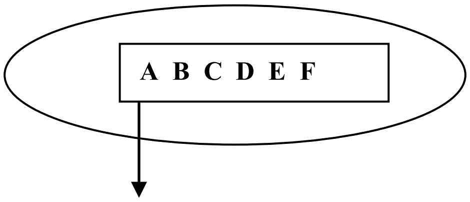
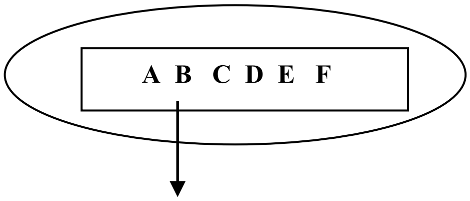
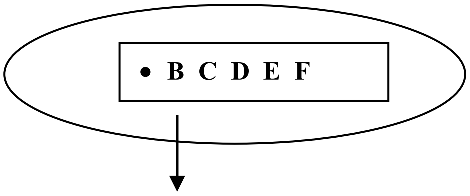
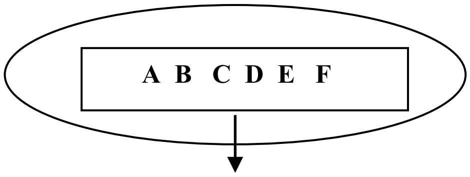
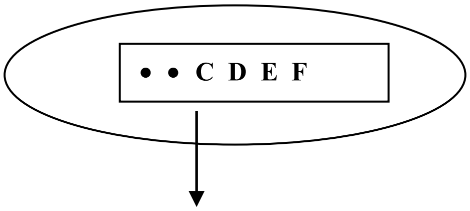
How many Equally Likely Samples Are there?

a. Simple Random Sampling With Replacement
versus
Simple Random Sampling Without Replacement

With Replacement **Answer: N^n**	**Without Replacement** **Answer: N (N-1) (N-2) … (N-n+1)**
	
1st Draw: A B C D E F G H are available **- Suppose "G" is selected for inclusion**	**1st Draw: A B C D E F G H are available** **- Suppose "G" is selected for inclusion**
	
2nd Draw: A B C D E F G H are ALL available **- Thus "G" is available for inclusion a 2nd time.**	**2nd Draw: A B C D E F H are available but "G" is not available anymore** **- Thus, "G" can only be included once.**
Etc	**etc**

Example – What is the Probability of Each Equally Likely Sample Sequence?

Eg- What is the Probability of {1st draw is A, 2nd draw is B, 3rd draw is C}?

With Replacement **Answer: = 1 / [N^n]**	**Without Replacement** **Answer: = 1/ [N (N-1) (N-2) … (N-n+1)]**
	
Probability {1st draw is A} = 1/6	**Probability {1st draw is A} = 1/6**
	
Probability {2nd draw is B, given 1st is A} = 1/6	**Probability {2nd draw is B, given 1st is A}= 1/5**
	
Probability {C 3rd given A, B} = 1/6	**Probability {C 3rd given A,B} = ¼**
Probability{sample sequence A,B,C} **= (1/6) (1/6) (1/6)** **= 1 / [6^3]** **= 1 / [N^n]**	**Probability{sample sequence A,B,C}** **= (1/6) (1/5) (1/4)** **= 1 / [6 * 5 * 4]** **= 1 / [N (N-1) (N-2) … (N-n+1)]**

a. Simple Random 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*.**
How many equally likely samples are there? **Answer: N^n** $\underset{\substack{\uparrow \\ \text{1st draw}}}{(N)} \quad \underset{\substack{\uparrow \\ \text{2nd draw}}}{(N)} \quad \underset{\substack{\uparrow \\ \text{3rd draw}}}{(N)} \quad \ldots \quad \underset{\substack{\uparrow \\ \text{nth draw}}}{(N)} \quad = \quad N^n$	**How many equally likely samples are there?** **Answer: (N)(N-1) …(N-n+1)** $\underset{\substack{\uparrow \\ \text{1st draw}}}{(N)} \quad \underset{\substack{\uparrow \\ \text{2nd draw}}}{(N\text{-}1)} \quad \underset{\substack{\uparrow \\ \text{3rd draw}}}{(N\text{-}2)} \quad \ldots. \quad \underset{\substack{\uparrow \\ \text{nth draw}}}{(N\text{-}n\text{+}1)} = \underset{\uparrow}{(N)}\underset{\uparrow}{(N\text{-}1)} \ldots \underset{\uparrow}{(N\text{-}n\text{+}1)}$
Probability {each equally likely sample of size n} = $$\frac{1}{N^n}$$	**Probability {each equally likely sample of size n} =** $$\frac{1}{N(N-1)(N-2)\ldots(N-n+1)}$$

Again, under simple random sampling:

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

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* ***("with replacement")***
- 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 = $\dfrac{1}{N^n} = \dfrac{1}{16}$

Here are the 16 possible (ordered) samples:

(spade, spade) **(club, spade)** **(heart, spade)** **(diamond, spade)**	**(spade, club)** **(club, club)** **(heart, club)** **(diamond, club)**	**(spade, heart)** **(club, heart)** **(heart, heart)** **(diamond, heart)**	**(spade, diamond)** **(club, diamond)** **(heart, diamond)** **(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 and B and C". There are 6 "qualifying" ordered sequences:

{ 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 }

What if we just want to know the number of "qualifying" ordered sequences of "A and B and C" that satisfy the event of "A and B and C"? **We don't really want to have to list them all out every time; how tedious.**
The answer is obtained by answering the question: *How many **rearrangements of the 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:**
  From among my selected, what letter should I put into the 1st position?
  How many choices are possible?
  Answer = 3

- **Second position, given 1st is filled:**
  Having "filled" position 1, from the remainder of my selected, what letter should I put into the 2nd position? How many choices are possible?
  Answer = (3 – 1) = 2.

- **Third position, given 1st and 2nd filled:** And so on … How many choices are possible?
  Answer = (3-2) = 1.

- **Thus, the total number of ordered rearrangements (permutations) of 3 items**
  = (3) (3-1) (3-2) = 6

How many different ordered rearrangements (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) …. (2)(1)**

The number of *ordered* rearrangements (permutations) of n items is

$$n! = (n)(n-1)(n-2) \ldots . (2)(1)$$

Note – n! (called the factorial) is just an abbreviation, so that we don't have to write out long hand expressions like (n)(n-1)(n-2)…(2)(1). More on this in Unit 4.

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

Example – You're playing cards on a Friday night. The dealer has just the 4 queens. He/she gives you 2 of them. What are the chances that your hand contains the queen of clubs and the queen of hearts, regardless of which you got first and which you got second?

Population

Four Queens (N=4)

Sampling Plan

- Draw one card at random
- Note its suit
- *Set the selected card aside (Do NOT return it to the pile)*
- Draw a 2nd card at random from the 3 that are remaining.
- Note its suit

Population size, N=4
Sample size, n=2
Total # (ordered sequence) samples possible = (4) (4-1) = (4)(3) = 12

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

Here is the sample space comprised of all 12 ordered sequence sample points.
Under simple random sampling each occurs with equal probability = 1/(12):

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

But, if order DOES NOT matter, then the events of interest are 6 equally likely "hands"

Hand (Event)	**Spade and Club**	**Heart and Club**	**Diamond and Heart**	**Spade and Heart**	**Diamond and Club**	**Spade and Diamond**
Qualifying Ordered sequences	**[spade, club]** **[club, spade]**	**[heart, club]** **[club, heart]**	**[diamond heart]** **[heart, diamond]**	**[spade, heart]** **[heart, spade]**	**[diamond, club]** **[club, diamond]**	**[spade, diamond]** **[diamond, spade]**

Putting it all together: How to Calculate the Probability of a Sample Outcome, in which order does not matter, under Simple Random Sampling

The solution involves two calculations of # orderings (permutations)

(1) Calculate the total number of "ordered sequence" samples

$$\begin{pmatrix}\text{Total \#}\\ \text{ordered sequence samples}\end{pmatrix} = \begin{pmatrix}\text{\# ways to}\\ \text{draw 1st}\end{pmatrix}\begin{pmatrix}\text{\# ways to}\\ \text{draw 2nd}\end{pmatrix}$$

$$= (\text{N})(\text{N-1})$$

$$= (4)(3)$$

$$= 12$$

(2) Next, calculate the *number of ordered rearrangments (permutations) of the sample obtained.*

$$\begin{pmatrix}\text{\# orderings of}\\ \text{given sample}\end{pmatrix} = \begin{pmatrix}\text{\# choices for}\\ \text{position 1}\end{pmatrix}\begin{pmatrix}\text{\# choices for}\\ \text{position 2}\end{pmatrix}$$

$$= (\text{n})(\text{n-1})$$

$$= (2)(1)$$

$$= 2$$

Pr[each equally likely UNordered sample]

$$= \frac{\text{\# ordered rearrangements (permutations) of the sample obtained}}{\text{\# of ordered samples that could have been obtained}}$$

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

$$\begin{pmatrix}\text{Probability of}\\ \text{1 club and 1 heart}\end{pmatrix} = \frac{\text{\# permutations of n=2 cards}}{\text{\# of ordered samples of n=2 from N=4}} = \frac{(n)(n-1)}{(N)(N-1)} = \frac{(2)(1)}{(4)(3)} = \frac{2}{12}$$

Simple Random Sampling WITHOUT Replacement
Summary

IF Order DOES Matter

- ♦ Total # ordered sequence samples $= N(N-1)(N-2) \cdots (N-n+1)$

- ♦ Under simple random sampling,

$$\text{Probability [each ordered sequence sample]} = \frac{1}{(N)(N-1)...(N-n+1)}$$

IF Order Does NOT Matter

- ♦ Total # ordered samples $= N(N-1)(N-2) \cdots (N-n+1)$

- ♦ # re-arrangements of the sample obtained = $(n)(n-1)(n-2)\ldots(2)(1)$

- ♦ Under simple random sampling,

$$\text{Probability [each sample, regardless of order]} = \frac{(n)(n-1)(n-2)...(2)(1)}{(N)(N-1)...(N-n+1)}$$

- ♦ The total # of Unordered samples is the reciprocal of this.

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

b. How to Select a Simple Random Sample WITHOUT Replacement

(Using a random number table)

Step 1: List the subjects in the sampled population. ♣ This is the sampling frame.	**Example – Obtain a simple random sample of n=30 from a population of size N=500-** ♣ Make a list of all N=500

Step 2:
Number this listing from "1" to "N"
♣ where N = size of sampled population
	Example, continued -
	N = 500
	n = 30

Step 3: The size of "N" tells you how many digits in a random number to be looking at: ♣ For N ≤ 10 Need only read 1 digit ♣ For N between 10 and 99 Read 2 digits ♣ For N between 100 and 999 Read 3 digits etc	**Example, continued –** Since N=500 is between 100 and 999 and is 3 digits long. ♣ 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 ≤ 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.