Reasons for Sampling

Sampling is the process of selecting items from a population so that it can be used to make judgments or inferences about the population.

1. To contact the whole population would be time consuming

Example: A candidate for public office may wish to determine her chances for election. A sample poll using the regular staff and field interviews of a professional polling firm would take only one or two days. It could take years to contact all the voting population!

2. The cost of studying all the items in a population may be prohibitive

Example: Public opinion polls and consumer testing organizations usually contact only a small portion of the population since it is more cost effective than contacting the entire population.

Reasons for Sampling

3. The physical impossibility of checking all items in the population

Example: Some populations are infinite. It would be impossible to check all the water in Okanagan Lake for bacterial levels, so we select samples at various locations.

4. The destructive nature of some sample tests

Example: If the wine tasters in Niagara-on-the-Lake drank all the wine to evaluate the vintage, they would consume the entire crop, and none would be available for sale.

5. The sample results are adequate

Example: **T**he federal government uses a sample of grocery stores scattered throughout Canada to determine the monthly index of food prices.

Sampling Methods

When collecting sample data from a population, we define a frame.

The **frame** is a complete or partial listing of items that make up a population from which the sample will be selected.

If the frame excludes certain groups or portions of the population, then it can result in inaccurate or biased results.

One can select either a probability or non-probability sample.

Non-probability sample – items can be selected without knowing the probability (using a non-random selection process) of the sample

Probability sample – items are selected based on a known probability (using a random selection process)

Non-probability Sampling

The advantage of using non-probability samples is that it can be gathered quickly with low cost.

It can be used for making informal approximations or as pilot study.

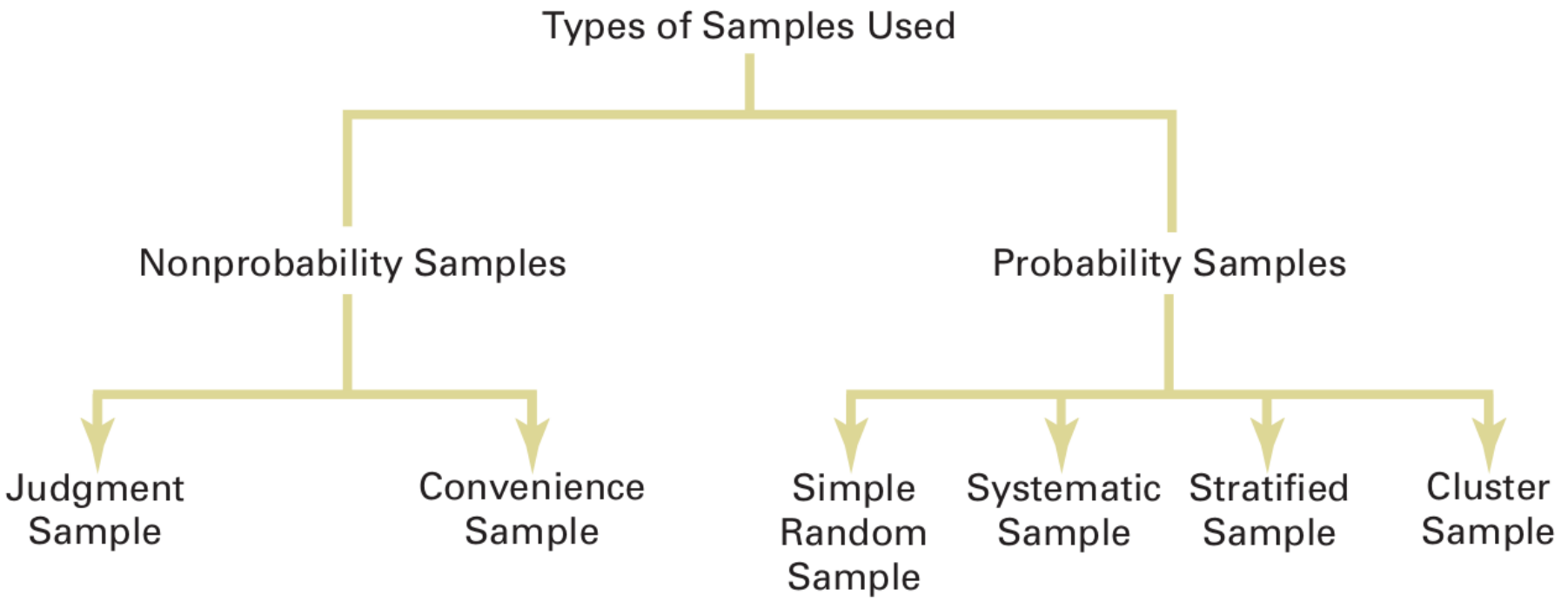
It cannot be used for statistical inference applications.

Non-probability samples can be subdivided into convenience samples or judgment samples.

Convenience (grab sampling, accidental sampling, opportunity sampling) data taken from a group of people easy to contact and reach...best used for pilot studies.

Judgment samples are opinions from pre-selected experts in a subject matter...cannot generalize results to a population.

Summary of Types of Samples



Simple Random Sample

The most widely used sample is the simple random sample.

In a simple random sample, a sample is selected so that each item or person in the population has the same chance of being included.

To select the sample we can use a spreadsheet or other statistical software to assign a random number for each item and then select a number of those random items for the sample.

If software is unavailable, then we can select a simple random sample by using a table of random numbers from the Appendix section of most statistic books.

Major disadvantage of this method is the numbering of all items in the population from 1 to N. Each item has an equal chance (1/N) of being selected.

Example – Simple Random Sample

Suppose a population consists of 650 employees in a company. A sample of 4 employees is to be selected from that population randomly with the help of random table.

You can use any technique to select n number of samples from the N population using the random number table.

52454	61352	76065	87641	45614	86752	29683
17001	47577	13445	08613	29744	33445	26874
60273	07573	22235	24833	34424	07360	44401
01959	60470	66829	42564	78863	41258	46837
21772	05860	31090	64660	36905	69138	96691
51552	14875	99400	91654	16830	39168	06374

Random Numbers and Generators

Random numbers can be made from a random number generator (RNG).

There are 2 types of RNGs...pseudo random number generators (PRNG) and real random number generators (RRNG).

The difference between the 2 type of generators is the SEED number. This seed number is used to initiate the RNG.

In PRNGs, the seed number is a very large number and is known by the programmer and can be modified at any time. This is useful for fields of simulations (Monte Carlo) or cryptography where you might want to reproduce your results.

In RRNG, the seed number used comes from real time data from physical phenomena (atmospheric or thermal noise) at any time or place on Earth.

Another way to seed the RRNG would be to use the system time (Unix epoch time Jan. 1, 1970...to present) which keeps changing every second.

In Excel, the Randbetween(a,b) function uses the system time as the seed to produce random numbers

Systematic Random Sampling

In some cases, even the simple random sampling technique maybe too time consuming. (Ex. Numbering each item before a random value can be used to pick a random item...numbering each item is the time consuming part)

Instead a systematic random sampling technique can be used.

A starting point is selected randomly and then every a^{th} member of the population is selected for the sample.

a can be calculated as $a = N/n$

For example, if the items are arranged in order of increasing size or other characteristic, then this method should not be used.

Statistics Canada uses this method to select households to fill out the long-form census every 5 years.

Example – Systematic Random Sampling

The following class roster lists the students enrolling in an introductory course in statistics. 3 students are to be randomly selected and asked various questions regarding course content and method of instruction.

Suppose a systematic random sample will select every 7^{th} student enrolled in the class (a number).

Initially, the 3^{rd} student on the list was selected at random.

Remembering that the random numbers start with 00, which students will be chosen to be members of the sample?

Example – Systematic Random Sampling

CSPM 264 01 BUSINESS & ECONOMIC STAT

8:00 AM 9:40 AM MW ST 118 LIND D

RANDOM NUMBER	NAME	CLASS RANK	RANDOM NUMBER	NAME	CLASS RANK
00	ANDERSON, RAYMOND	SO	23	MEDLEY, CHERYL ANN	SO
01	ANGER, CHERYL RENEE	SO	24	MITCHELL, GREG R	FR
02	BALL, CLAIRE JEANETTE	FR	25	MOLTER, KRISTI MARIE	SO
03	BERRY, CHRISTOPHER G	FR	26	MULCAHY, STEPHEN ROBERT	SO
04	BOBAK, JAMES PATRICK	SO	27	NICHOLAS, ROBERT CHARLES	JR
05	BRIGHT, M. STARR	JR	28	NICKENS, VIRGINIA	SO
06	CHONTOS, PAUL JOSEPH	SO	29	PENNYWITT, SEAN PATRICK	SO
07	DETLEY, BRIAN HANS	JR	30	POTEAU, KRIS E	JR
08	DUDAS, VIOLA	SO	31	PRICE, MARY LYNETTE	SO
09	DULBS, RICHARD ZALFA	JR	32	RISTAS, JAMES	SR
10	EDINGER, SUSAN KEE	SR	33	SAGER, ANNE MARIE	SO
11	FINK, FRANK JAMES	SR	34	SMILLIE, HEATHER MICHELLE	SO
12	FRANCIS, JAMES P	JR	35	SNYDER, LEISHA KAY	SR
13	GAGHEN, PAMELA LYNN	JR	36	STAHL, MARIA TASHERY	SO
14	GOULD, ROBYN KAY	SO	37	ST. JOHN, AMY J	SO
15	GROSENBACHER, SCOTT ALAN	SO	38	STURDEVANT, RICHARD K	SO
16	HEETFIELD, DIANE MARIE	SO	39	SWETYE, LYNN MICHELE	SO
17	KABAT, JAMES DAVID	JR	40	WALASINSKI, MICHAEL	SO
18	KEMP, LISA ADRIANE	FR	41	WALKER, DIANE ELAINE	SO
19	KILLION, MICHELLE A	SO	42	WARNOCK, JENNIFER MARY	SO
20	KOPERSKI, MARY ELLEN	SO	43	WILLIAMS, WENDY A	SO
21	KOPP, BRIDGETTE ANN	SO	44	YAP, HOCK BAN	SO
22	LEHMANN, KRISTINA MARIE	JR	45	YODER, ARLAN JAY	JR

Stratified Random Sampling

If the population can be subdivided into groups based on some characteristic, then a stratified random sample can be used.

It guarantees that each group is represented in the sample. The groups are called strata.

Once the strata are defined, we can apply simple random sampling within each group or strata to collect the sample.

Example – Stratified Random Sampling

For the stratified random sample, in order to sample 50 from the 200 companies, it should be proportional to the number of items in each firm.

For stratum 1, (0.02)(50) = 1 sample
For stratum 2, (0.10)(50) = 5 samples
For stratum 3, (0.54)(50) = 27 samples
For stratum 4, (0.33)(50) = 16.5 = 16 samples
For stratum 5, (0.01)(50) = 0.5 = 1 sample
Total = 50

Stratum	Profitability (return on equity)	Number of Firms	Relative Frequency	Number Sampled
1	30% and over	4	0.02	1
2	20 up to 30%	20	0.10	5
3	10 up to 20%	108	0.54	27
4	0 up to 10%	66	0.33	16
5	Deficit	2	0.01	1
Total		**200**	**1.00**	**50**

Cluster Sampling

Cluster sampling is often used to reduce the cost of sampling a population scattered over a large geographic area.

First, a population is divided into clusters using naturally occurring geographic or other boundaries. These are called primary units.

Then, individual units are randomly selected.

Then within each unit, a random sample of items can be made.

Cluster sampling is combination of stratified and simple random sampling.