SAMPLE DESIGNING

Dr. Asheesh Srivastava
Professor, Head & Dean
Department of Educational Studies
School of Education,
Mahatma Gandhi Central University,
Motihari, East Champaran, Bihar-845401
profasheesh@mgcub.ac.in

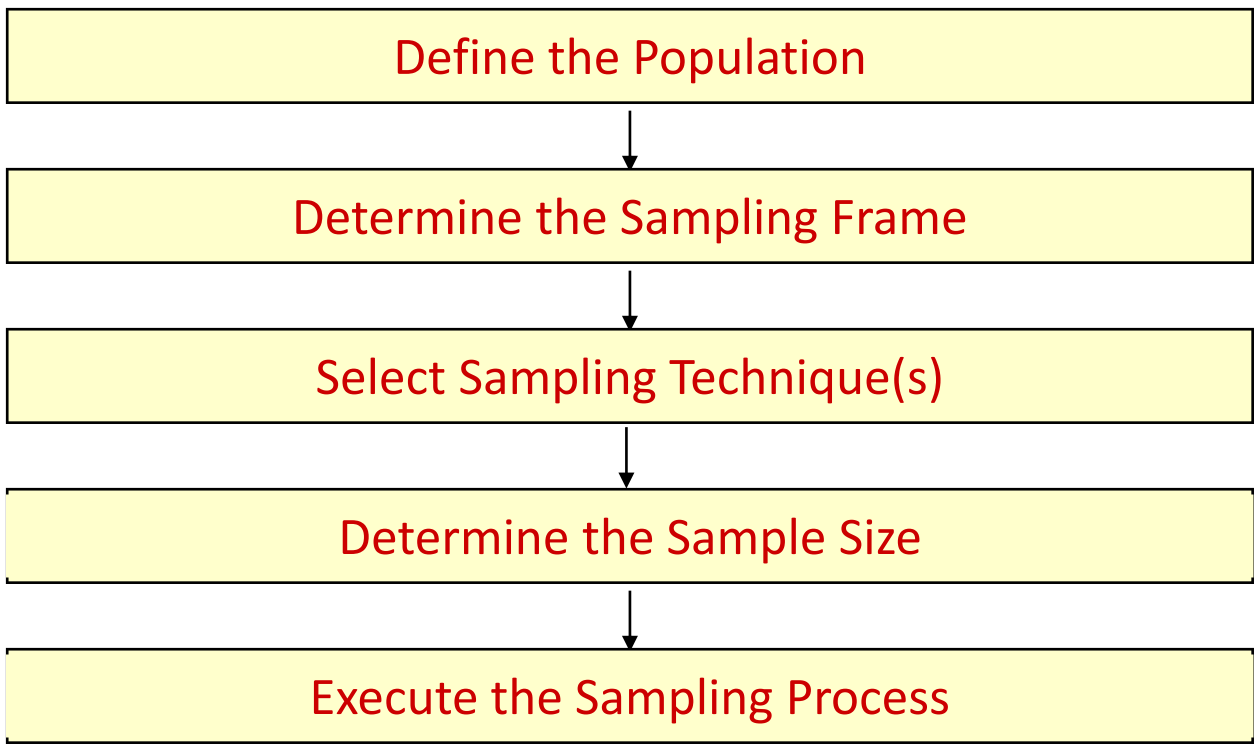
Key terms

- **Population** : the elements about which we wish to make some inferences
- **Census:** a census involves complete details of the elements of a population
- **Universe :** the universe is the entire group of items the researcher wish to study and about which they wish to generalize.
- **Population element**: the individual participant or object on which the measurement is taken
- **Population Parameter**: A **parameter** is a summary description of a fixed characteristic or measure of the target population. A parameter denotes the true value which would be obtained if a census rather than a sample was undertaken.
- **Target population:** the collection of elements or objects that possess the information about which inferences are to be made.
- **Sample**: a group of cases, participants, events, or records consisting of a portion of the target population, carefully selected to represent that population
- **Sample Statistic**: A **statistic** is a summary description of a characteristic or measure of the sample. The sample statistic is used as an estimate of the population parameter.
- **Sampling unit:** the basic unit containing the elements of the population to be sampled.
- **Sampling frame:** a representation of the elements of the target population.

Sample Vs. Census

Type of Study	Conditions Favoring the Use of Sample	Census
1. Budget	Small	Large
2. Time available	Short	Long
3. Population size	Large	Small
4. Variance in the characteristic	Small	Large
5. Cost of sampling errors	Low	High
6. Cost of nonsampling errors	High	Low
7. Nature of measurement	Destructive	Nondestructive
8. Attention to individual cases	Yes	No

The Sampling Design Process

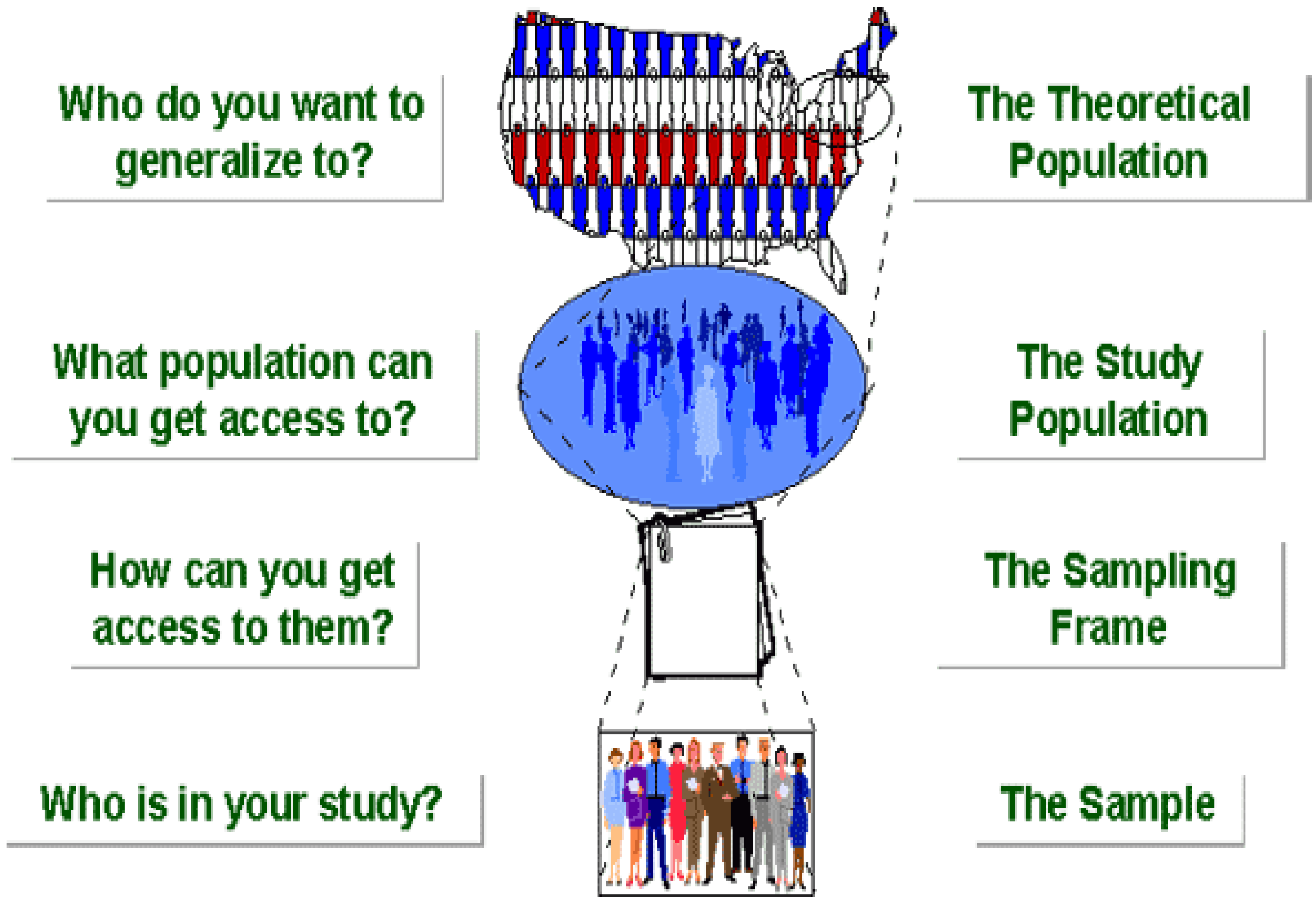


Define the Target Population

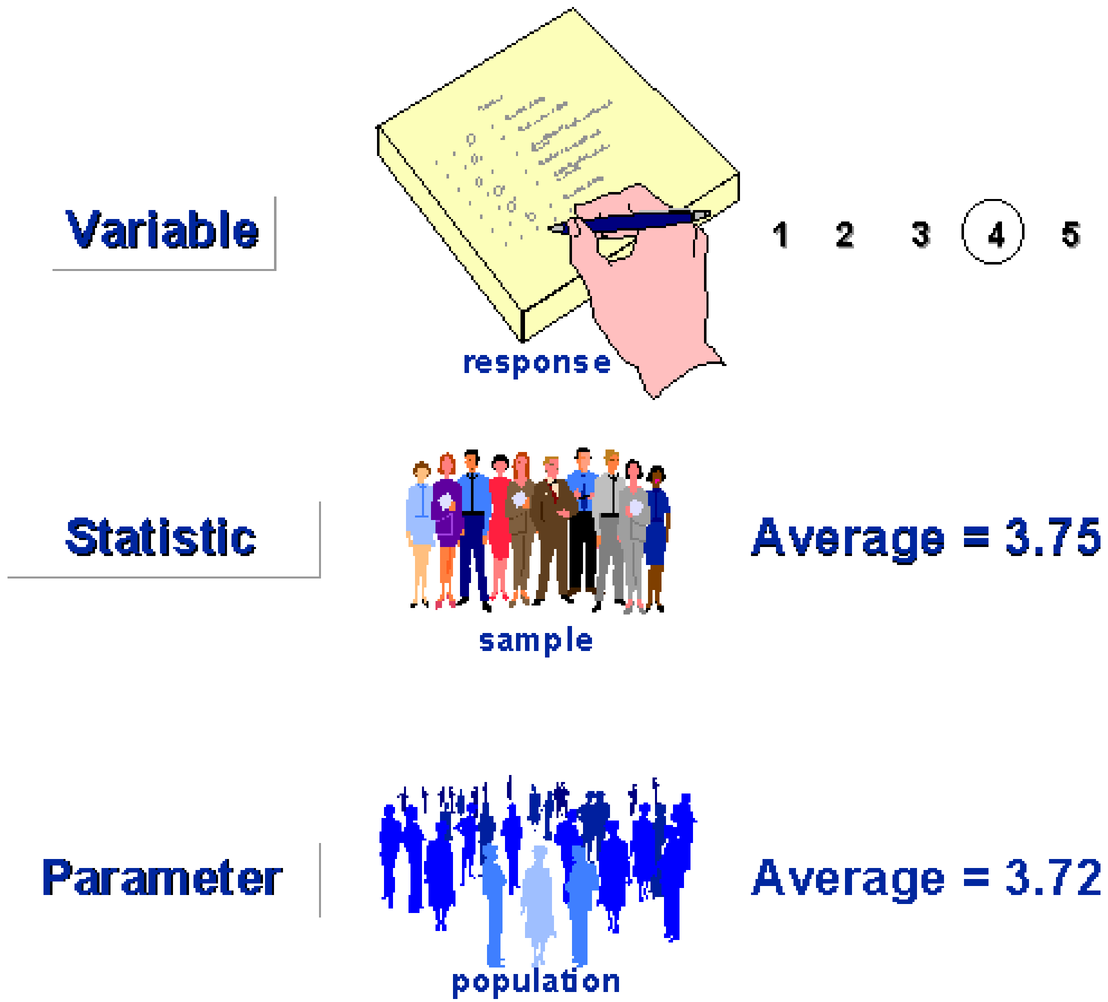
The target population should be defined in terms of elements, sampling units, extent, and time.

- An **element** is the object about which or from which the information is desired, e.g., the respondent.
- A **sampling unit** is an element, or a unit containing the element, that is available for selection at some stage of the sampling process.
- **Extent** refers to the geographical boundaries.
- **Time** is the time period under consideration.

Basics of sampling

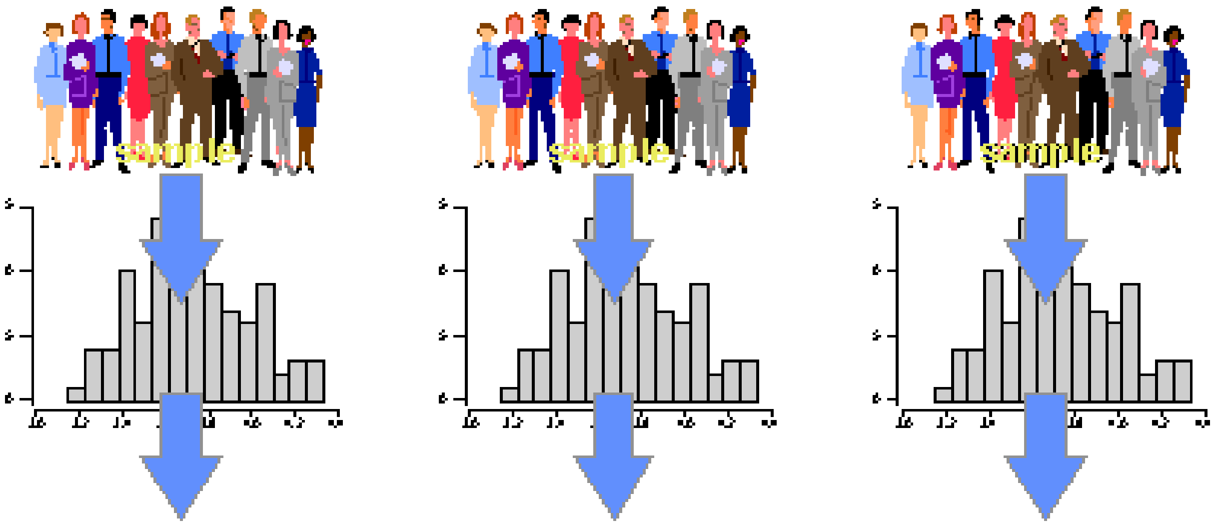


- The group that actually completes your study is a subsample of the sample -- it doesn't include nonrespondents or dropouts.



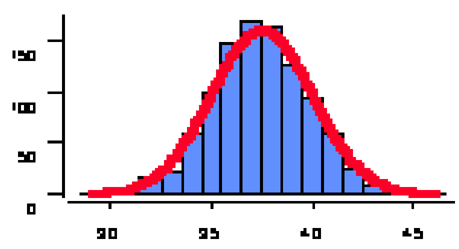
A response is a specific measurement value that a sampling unit supplies. In the figure, the person is responding to a survey instrument and gives a response of '4'.

The sampling Distribution



Average **Average** **Average**

The Sampling Distribution...



...is the distribution of a statistic across an infinite number of samples

- We don't ever actually construct a sampling distribution. Because to construct it we would have to take an *infinite* number of samples and infinite is not a number we know how to reach. So why do we even talk about a sampling distribution?

- The standard deviation of the sampling distribution tells us something about how different samples would be distributed. In statistics it is referred to as the **standard error**

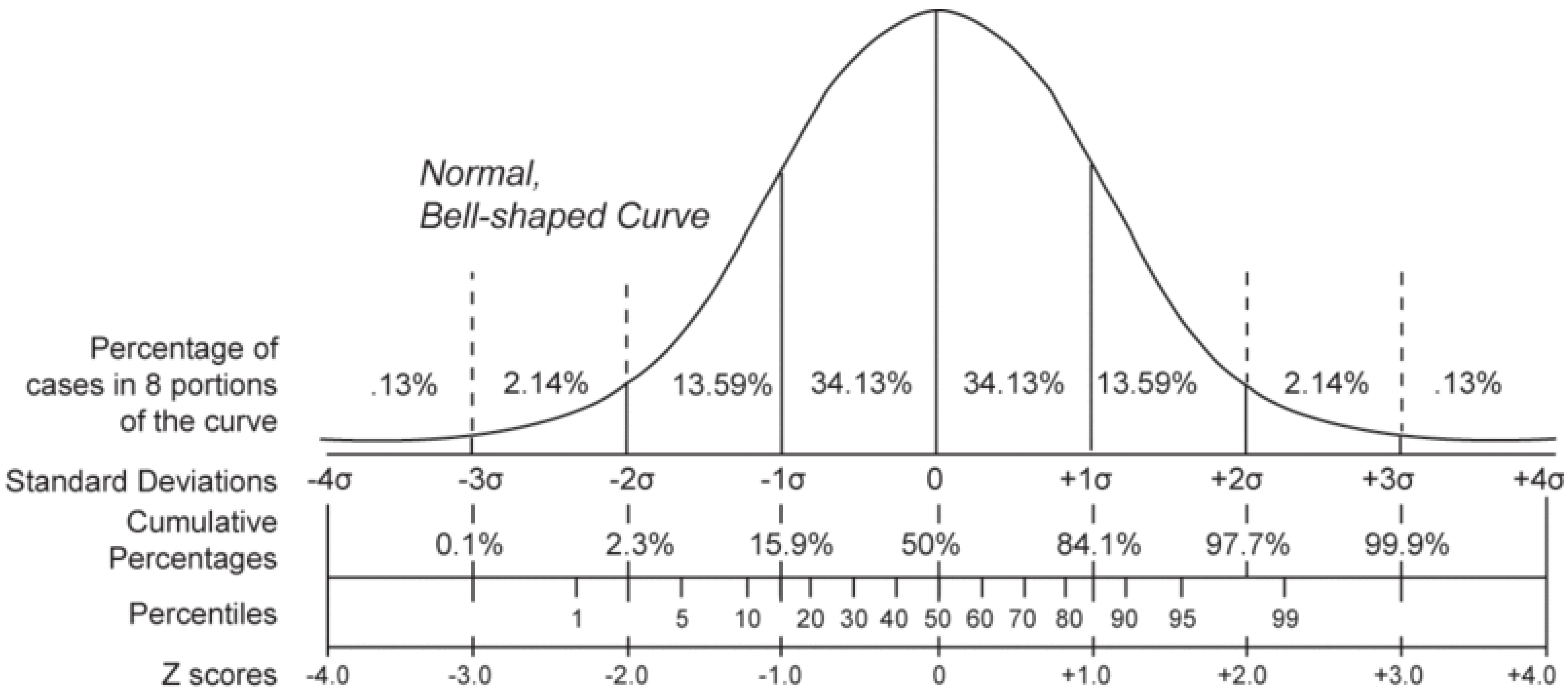
- A standard deviation is the spread of the scores around the average in a <u>single sample</u>.

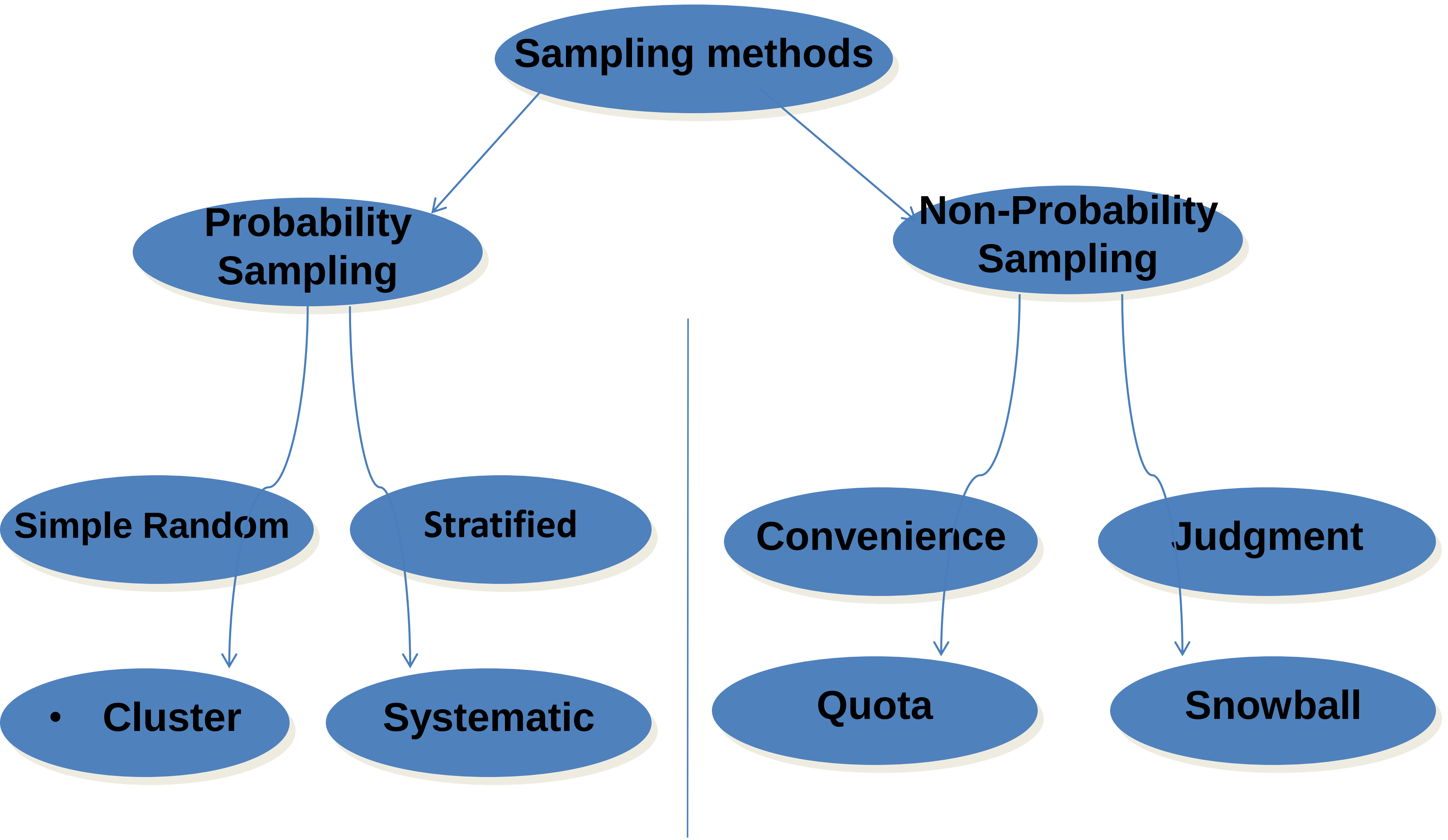
- The standard error is the spread of the averages around the average of averages in a<u>sampling distribution</u>.

The 68, 95, 99 percent rule

- There is a general rule that applies whenever we have a normal or bell-shaped distribution. Start with the average -- the center of the distribution.
- If you go up and down (i.e., left and right) one standard unit, you will include approximately 68% of the cases in the distribution (i.e., 68% of the area under the curve).
- If you go up and down two standard units, you will include approximately 95% of the cases.
- And if you go plus-and-minus three standard units, you will include about 99% of the cases. Notice that I didn't specify in the previous few sentences whether I was talking about standard deviation units or standard error units. That's because the same rule holds for both types of distributions (i.e., the raw data and sampling distributions).

The 68, 95, 99 percent rule




Sampling methods
Probability Sampling
Non-Probability Sampling
Simple Random
Stratified
Convenience
Judgment
• Cluster
Systematic
Quota
Snowball

Probability Sampling

- A **probability sampling** method is any method of sampling that utilizes some form of *random selection*.
- Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.
- **N** = the number of cases in the sampling frame
- **n** = the number of cases in the sample
- $_{N}C_{n}$ = the number of combinations (subsets) of n from N
- **f** = n/N = the sampling fraction

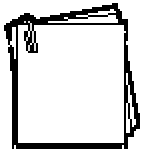
Random Selection & Assignment

- **Random *selection*** is how you draw the sample of people for your study from a population.
- **Random *assignment*** is how you assign the sample that you draw to different groups or treatments in your study.
- Random selection is related to sampling. Therefore it is most related to the external validity (or generalizability) of your results. After all, we would randomly sample so that our research participants better represent the larger group from which they're drawn.
- Random assignment is most related to design. In fact, when we randomly assign participants to treatments we have, by definition, an experimental design. Therefore, random assignment is most related to internal validity.

Simple Random Sampling

- The simplest form of random sampling is called **simple random sampling**.
- **Objective**: To select n units out of N such that each ${}_{N}C_{n}$ has an equal chance of being selected.
- **Procedure**: Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.

List of Clients



Random Subsample

- For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is $f = n/N = 100/1000 = .10$ or 10%.
- You would need three sets of balls numbered 0 to 9, one set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list of names from 1 to 1000 and then use the ball machine to select the three digits that selects each person.

Illustration of Simple Random Sampling

A	B	C	D	E
1	6	11	**16**	21
2	**7**	12	17	22
3	8	13	18	23
4	**9**	14	19	**24**
5	10	15	20	25

Select five random numbers from 1 to 25. The resulting sample consists of population elements 3, 7, 9, 16, and 24. Note, there is no element from Group C.

Stratified Random Sampling

- **Stratified Random Sampling**, also sometimes called *proportional* or *quota* random sampling, involves **dividing your population into homogeneous subgroups** and then taking a simple random sample in each subgroup. In more formal terms:

- **Objective**: Divide the population into non-overlapping groups (i.e., *strata*) N_1, N_2, N_3, ... N_i, such that $N_1 + N_2 + N_3 + ... + N_i = N$. Then do a simple random sample of $f = n/N$ in each strata.
- There are several major reasons why you might prefer stratified sampling over simple random sampling. First, it assures that you will be able to represent not only the overall population, but also **key subgroups** of the population, especially small minority groups.

Stratified sampling

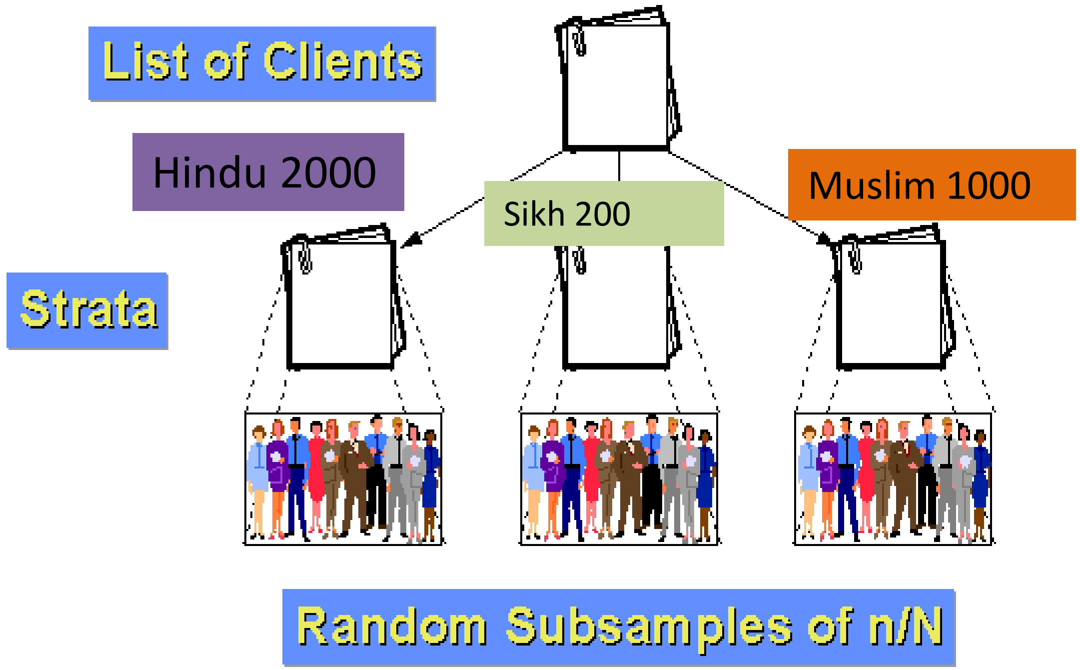


Illustration of Stratified Sampling

A	B	C	D	E
1	6	11	16	**21**
2	**7**	12	17	22
3	8	**13**	18	23
4	9	14	**19**	24
5	10	15	20	25

Randomly select a number from 1 to 5 for each stratum, A to E. The resulting sample consists of population elements 4, 7, 13, 19 and 21. Note, one element is selected from each column.

Systematic Random Sampling

- Here are the steps you need to follow in order to achieve a **systematic random sample**:
- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- k = N/n = the interval size
- randomly select an integer between 1 to k
- then take every kth unit

Systematic random sampling

N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Systematic Sampling

A	B	C	D	E
1	6	11	16	21
2	**7**	**12**	**17**	**22**
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

Select a random number between 1 to 5, say 2. The resulting sample consists of population 2, (2+5=) 7, (2+5x2=) 12, (2+5x3=)17, and (2+5x4=) 22. Note, all the elements are selected from a single row.

Cluster (Area) Random Sampling

- Elements within a cluster should be as heterogeneous as possible, but clusters themselves should be as homogeneous as possible. Ideally, each cluster should be a small-scale representation of the population.

- In **probability proportionate to size sampling**, the clusters are sampled with probability proportional to size. In the second stage, the probability of selecting a sampling unit in a selected cluster varies inversely with the size of the cluster.

- In cluster sampling, we follow these steps:
- divide population into clusters (usually along geographic boundaries)
- randomly sample clusters
- measure <u>all</u> units within sampled clusters


I N D I A
POLITICAL MAP
LEGEND

Cluster Sampling (2-Stage)

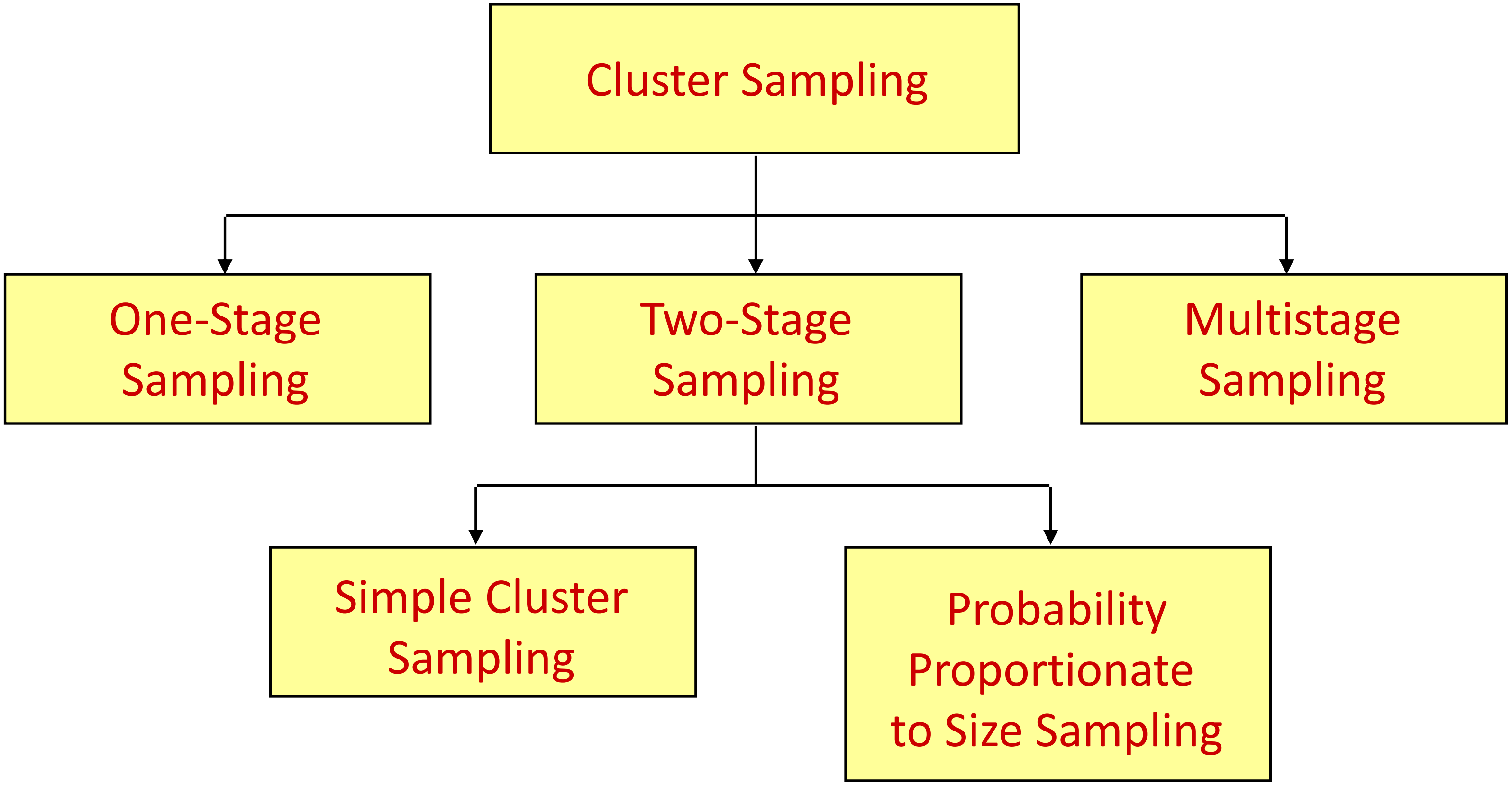
A	B	C	D	E
1	6	11	16	**21**
2	**7**	12	17	22
3	8	13	**18**	**23**
4	9	14	19	24
5	10	15	**20**	25

Randomly select 3 clusters, B, D and E. Within each cluster, randomly select one or two elements. The resulting sample consists of population elements 7, 18, 20, 21, and 23. Note, no elements are selected from clusters A and C.

Multi-Stage Sampling

- The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

Types of Cluster Sampling



Non-Probability Sampling

- The difference between nonprobability and probability sampling is that nonprobability sampling does not involve *random* selection and probability sampling does. Does that mean that nonprobability samples aren't representative of the population? Not necessarily. But it does mean that nonprobability samples cannot depend upon the rationale of probability theory.

Convenience Sampling

Convenience sampling attempts to obtain a sample of convenient elements. Often, respondents are selected because they happen to be in the right place at the right time.

- use of students, and members of social organizations
- mall intercept interviews without qualifying the respondents
- department stores using charge account lists
- “people on the street” interviews

Illustration of Convenience Sampling

A	B	C	D	E
1	6	11	**16**	21
2	7	12	**17**	22
3	8	13	**18**	23
4	9	14	**19**	24
5	10	15	**20**	25

Group D happens to assemble at a convenient time and place. So all the elements in this Group are selected. The resulting sample consists of elements 16, 17, 18, 19 and 20. Note, no elements are selected from group A, B, C and E.

Judgmental Sampling

Judgmental sampling is a form of convenience sampling in which the population elements are selected based on the judgment of the researcher.

- test markets
- purchase engineers selected in industrial marketing research
- expert witnesses used in court

Illustration of Judgmental Sampling

A	B	C	D	E
1	6	**11**	16	21
2	7	12	17	22
3	**8**	**13**	18	23
4	9	14	19	**24**
5	**10**	15	20	25

The researcher considers groups B, C and E to be typical and convenient. Within each of these groups one or two elements are selected based on typicality and convenience. The resulting sample consists of elements 8, 10, 11, 13, and 24. Note, no elements are selected from groups A and D.

Quota Sampling

Quota sampling may be viewed as two-stage restricted judgmental sampling.

- The first stage consists of developing control categories, or quotas, of population elements.
- In the second stage, sample elements are selected based on convenience or judgment.

Control Characteristic	Population composition Percentage	Sample composition Percentage	Sample composition Number
Sex			
Male	48	48	480
Female	52	52	520
	100	100	1000

Illustration of Quota Sampling

A	B	C	D	E
1	**6**	11	16	21
2	7	12	17	**22**
3	8	**13**	18	23
4	9	14	19	24
5	10	15	**20**	25

A quota of one element from each group, A to E, is imposed. Within each group, one element is selected based on judgment or convenience. The resulting sample consists of elements 3, 6, 13, 20 and 22. Note, one element is selected from each column or group.

Snowball Sampling

In **snowball sampling**, an initial group of respondents is selected, usually at random.

- After being interviewed, these respondents are asked to identify others who belong to the target population of interest.
- Subsequent respondents are selected based on the referrals.

Illustration of Snowball Sampling

Random <u>Selection</u>

<u>Referrals</u>

A	B	C	D	E
1	6	11	16	21
2	7	**12**	17	22
3	8	**13**	**18**	23
4	**9**	14	19	24
5	10	15	20	25

Elements 2 and 9 are selected randomly from groups A and B. Element 2 refers elements 12 and 13. Element 9 refers element 18. The resulting sample consists of elements 2, 9, 12, 13, and 18. Note, there are no element from group E.

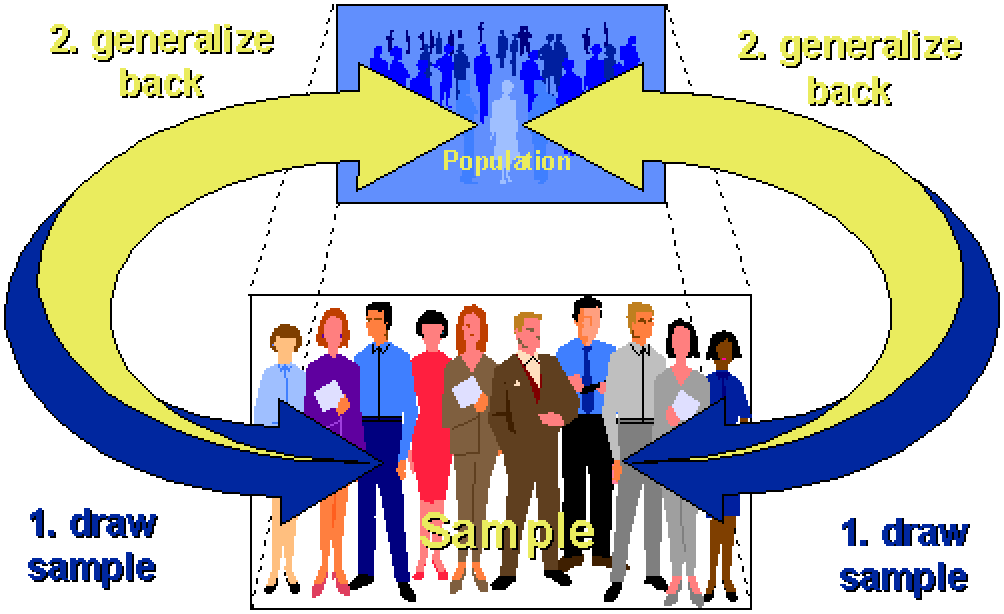
Expert Sampling

- Expert sampling involves the assembling of a sample of persons with known or demonstrable experience and expertise in some area. Often, we convene such a sample under the auspices of a "panel of experts." There are actually two reasons you might do expert sampling. First, because it would be the best way to elicit the views of persons who have specific expertise. In this case, expert sampling is essentially just a **specific subcase of purposive sampling**.

External Validity

- external validity is the degree to which the conclusions in your study would hold for other persons in other places and at other times.
- In science there are two major approaches to how we provide evidence for a generalization. Called as **Sampling Model**.
- In the sampling model, you start by identifying the population you would like to generalize to. Then, you draw a fair sample from that population and conduct your research with the sample. Finally, because the sample is representative of the population, you can automatically generalize your results back to the population.
- The second approach to generalizing is the **Proximal Similarity Model**. 'Proximal' means 'nearby' and 'similarity' means 'similarity'

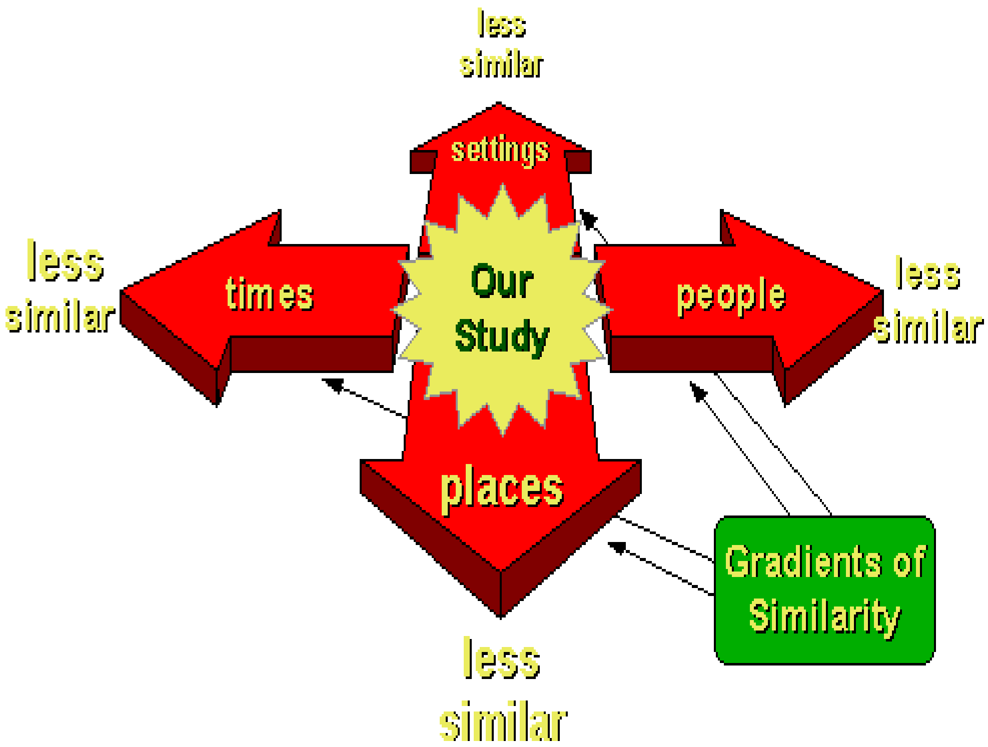
Generalisation



Threats to External Validity

- There are three major threats to external validity because there are three ways you could be wrong –
- **people,**
- **places or**
- **times**.
- Your critics could come along, for example, and argue that the results of your study are due to the **unusual type of people** who were in the study. Or, they could argue that it might only work because of the **unusual place** you did the study in (perhaps you did your educational study in a college town with lots of high-achieving educationally-oriented kids). Or, they might suggest that you did your study in a **peculiar time**.
- For instance, if you did your **smoking cessation study** the week after the Surgeon General issues the well-publicized results of the **latest smoking and cancer studies**, you might get different results than if you had done it the week before.

Threats to external validity



Strengths and Weaknesses of Non-Probability Sampling Techniques

Technique	Strength	Weakness
Convenience sampling	**Least expensive; most convenient**	**Selection bias; sample not representative; not recommended for descriptive or causal research**
Judgmental sampling	**Low cost; convenient**	**Not generalizable; subjective**
Quota sampling	**Control over key variables**	**Selection bias; not representative**
Snowball sampling	**Useful when dealing with low incidence**	**Time-consuming**

Strengths and Weaknesses of Probability Sampling Techniques

Technique	Strength	Weakness
Simple random sample	Easy to understand; results generalizable	Sample frame construction a problem; may not be representative
Systematic sampling	May increase representativeness; easier to implement	May decrease representativeness
Stratified sampling	Includes all important sub-groups	Many variables defy stratification; variable selection is crucial
Cluster sampling	Easy to implement; very cost effective	Can be imprecise

Choosing a Sampling Technique

Factors	Conditions favoring Nonprobability Sampling	Conditions favoring Probability Sampling
Nature of Research	Exploratory	Conclusive
Relative magnitude of sampling and nonsampling errors	Nonsampling larger	Sampling larger
Variability in the population	Low	High
Statistical considerations	Unfavorable	Favorable
Operational considerations	Favorable	Unfavorable

Determining the sample size

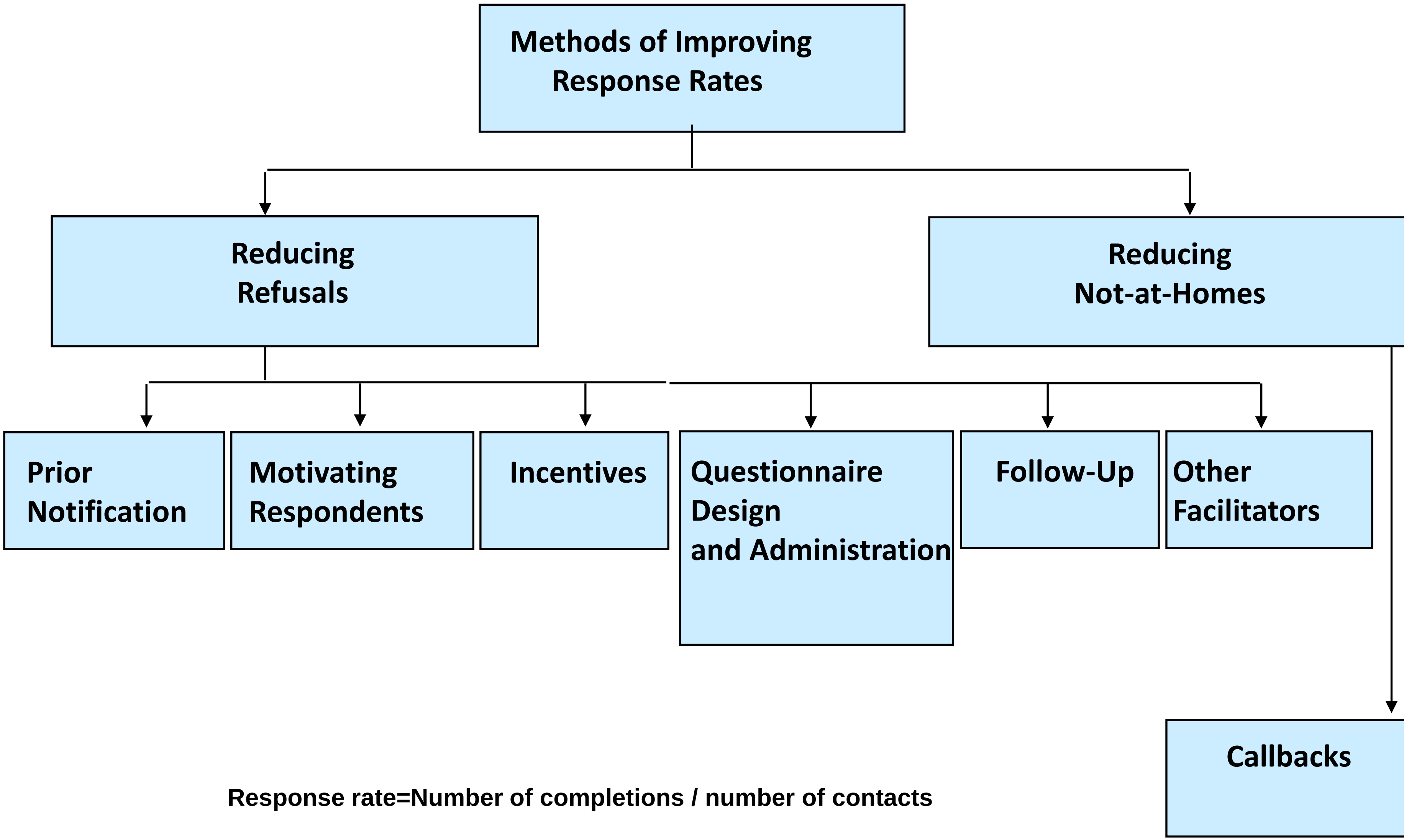
Important qualitative factors in determining the sample size are:

- the importance of the decision → More complex decision- more information required so larger sample
- the nature of the research → Exploratory- small sample; → Descriptive-large sample
- the number of variables → More variables-larger sample
- the nature of the analysis → Sophisticated analysis i.e. multivariate techniques-larger sample
- sample sizes used in similar studies
- incidence rates → The rate of occurrence of persons eligible to participate in the study expressed as percentage
- completion rates → The percentage of qualifying respondents who complete the interview
- resource constraints

Symbols for Population and Sample Variables

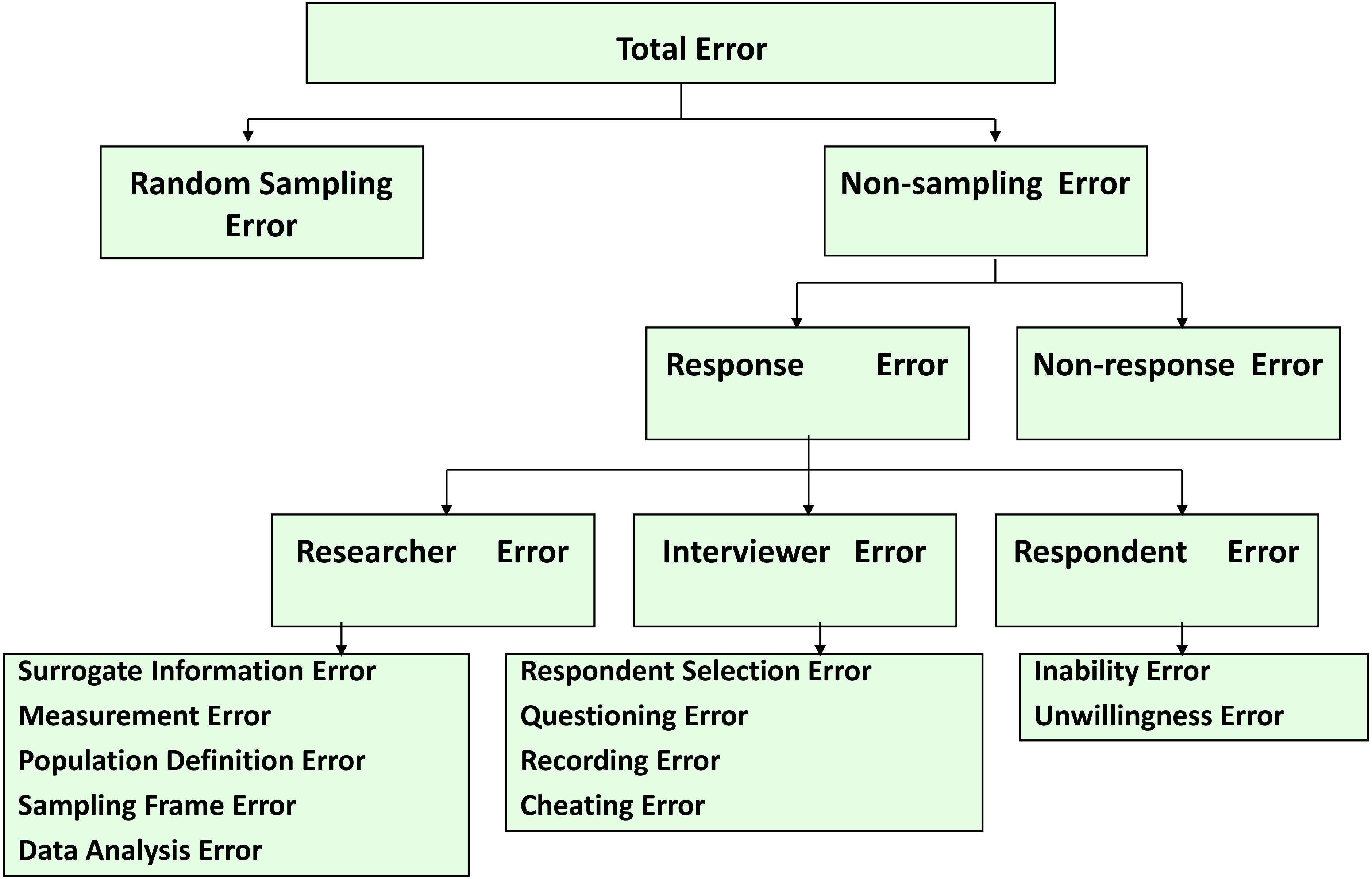
Variable	Population	Sample
Mean	μ	$\bar{X}$
Proportion	π	p
Variance	σ^2	s^2
Standard deviation	σ	s
Size	N	n
Standard error of the mean	$\sigma_{\bar{X}}$	$s_{\bar{X}}$
Standard error of the proportion	σ_p	s_p
Standardized variate *(z)*	$\frac{X - \mu}{\sigma}$	$\frac{X - \bar{X}}{S_X}$

Improving Response Rates



Response rate=Number of completions / number of contacts

Potential Sources of Error in Research Designs



THANK YOU