Observation Studies, Sampling Designs and Bias

Study / memorize this **Observation Study:** is a study wherein the researcher passively observes individuals or objects and measures / records some characteristic of the population – which may be a categorical or numeric variable – without seeking to influence the response. The researcher does not impose treatments upon the subjects.

Study this **What is the Objective of Observation Studies?**
The goal of Observation Studies is to examine if there is an association or relationship between 2 factors or variables [the Explanatory and Response variables]; they do not reveal a causal relationship because there are too many factors that aren't accounted or controlled for in an observation study. For example, *if* we found that a group of students that took Princeton Review did better than those that didn't, we may believe that the course is *associated* with better results.

Confounding occurs when the effects of an extraneous [~ outside] variable / factor upon the response variable gets mixed up with that of the explanatory variable so that the "contribution" / effect of the latter [i.e. the explanatory variable] upon the response cannot be isolated / separated / distinguished. As a result, we cannot infer causation since *if* the "treatment" groups differ in more ways than one, then we cannot attribute the differences in the response to the explanatory variable *alone* as the confounding variable / some combination of factors could be responsible!

CONCEPT CHECK: As a result, in observation studies, we cannot attribute causation to the explanatory variable since IF the "treatment" groups differed in more ways than one, then we cannot attribute the differences in the response to the explanatory variable ALONE as the confounding variable or some combination of factors could be responsible!

Experiment: is a study where the researcher actively imposes different treatments upon individuals / animals / objects and controls / manipulates the experimental circumstances / conditions to observe the response. Experiments permit cause-effect conclusions to be made. Each individual on which the experiment is performed is the experimental units/ *subject.*

What is the Objective of Experiments? How is it accomplished?
The objective is to observe cause-effect relationships between the Explanatory variable [i.e. the Treatments] and the Response. This is achieved by making the circumstances as alike as possible so that the only differences between the subjects are due to the Treatment differences [if any]!

What *can* one conclude from an experiment? Why / how?
That the effects in the Response variable were ***due to*** the Explanatory variable [Treatment] since the experimenter randomly assigns the subjects to the treatments [to minimize bias and average out the effects of unknown / uncontrollable sources of variation amongst the subjects] and seeks to create identical conditions amongst the treatment groups [similar subjects, similar circumstances] thereby minimizing the effects of potentially confounding factors upon the response. This way, any differences in the Response variable can be traced / attributed to the differences between the treatments [i.e. to the explanatory variable].

Study this Why can't causation be attributed in an Observation Study?
Because conditions for the 2 [or more groups] aren't identical or comparable, the individuals in the different "treatment" groups are likely to overtly – or covertly – in more ways than the explanatory variable alone! There are too many confounding [literally, "confusing"] variables: outside factors that may affect the response variable, too, in addition to the Explanatory variable...so that the effects of these extraneous factors – acting in concert amongst themselves upon the Response, or *with* the explanatory variable – shall get mixed up with that of the explanatory variable making isolation of the various contributory effects on the Response...impossible! E.g. church-goers having longer lives. Extraneous factors include: exercise, diet, moderation in "vices".

Caution: To show that a certain variable is a potentially confounding factor, you must
a) demonstrate a link between the explanatory variable and the confounding variable i.e. a clear direct / inverse relationship between the 2 variables
b) demonstrate a link between the confounding variable and the response variable i.e. a clear direct / inverse relationship between the 2 variables

The aim is to show *how* the effects of the confounding factor(s) [or combinations of these factors] upon the response variable could get mixed up with that of the explanatory variable so that the "contribution" / effect of the explanatory variable upon the response is unable to be isolated / separated / distinguished.

Example 1: Candy consumption [explanatory variable] results in longevity [response variable]. Candy consumers also, incidentally / coincidentally, exercise more [confounding factor 1] or have more sex [confounding factor 2] or smoke less [confounding factor 3], which lead to longer lives.

Example 2: Sleeping more [explanatory variable] results in less obesity [response

variable]. Sleeping more, incidentally / coincidentally, is also associated with exercising more [confounding factor 1] which leads to less obesity **Or** with strict / monitoring parents [confounding factor 2] who monitor diets…which leads to less obesity **Or** less stressful individuals [confounding factor 3] consume less food to relieve stress…which leads to less obesity.

Study this **Sampling Variability:** is the natural tendency of randomly drawn samples to differ from each other i.e. samples exhibit natural variation in their characteristics and numerical summaries.

Study this **Sample Size:** The larger the sample, the more precise the estimates of the population parameters obtained from the sample. Precision of estimates depends on Sample Size and *not* a fixed fraction / proportion of the Population i.e. larger populations do *not* require (proportionally) larger samples! The random selection process inherent in SRS allows for selection of a representative sample *irrespective* of the size of the population.

Study this **Statistical Significance** An outcome [or statistic] is considered to be statistically significant [or simply, *significant*] if it is so "extreme" that its occurrence cannot be attributed to random chance / sampling variations, *were the hypothesized value of the parameter were true*.

Alternately, an observed result is statistically significant if it is "rare" i.e. if H0 were true, then we would *not* get the outcome merely due to sampling / chance variations. Think of it this way: if the statistic *is* rare, then it is significantly "different" from the hypothesized parameter value in Ho!

In the case of 2 or more samples, the differences are statistically significant [or simply, *significant*], assuming that there are indeed no significant differences between the populations [*re* Mean or Proportion], if the observed differences are so "extreme" or large that the outcomes cannot be attributed to random chance / sampling variations.

If an outcome is *not* rare, then it is statistically *not* significant [think of it this way: if the statistic is not rare, then it is *not* significantly "different" from the hypothesized parameter value in Ho!].

Study and memorize this **What is Sampling?**
Sampling is collecting information from a representative subset of the entire population and then drawing conclusions about the population from the sample. We may seek a fact [weight, income] or an opinion [preferences, interests]; the

characteristic may be numerical or categorical so we may calculate a statistic (sample mean or sample proportion) to estimate the corresponding population parameter.

Study and memorize this **What is the Objective of Surveys? How is it attained?**
The *objective* of a survey is to make inferences **[Psst!** Create Confidence Intervals...and perform Tests of Hypotheses] pertaining to the Parameter of Interest**!]** and to generalize the results to the population**.** We achieve this by selecting representative samples [≈ random-samples!] from the population of interest.

Study and memorize this **What 2 types of simple calculations can be performed using the sample data...that we subsequently use for Statistical Inference?**
Depending on the Q asked, we can calculate a Proportion, and a Mean. For example, from a survey of shampoo users, we may determine

- the Proportion of shampoo users satisfied with their current sampoo **AND**
- the Mean Satisfaction Rating of shampoo users [on a 1-5 scale?]

Study these terms:
Population: entire set or group of individuals or objects that we need information about **OR** the individuals to whom the results of a survey can be generalized. **(NOTE:** In Statistics, the population is <u>not</u> a number – 100,000 – but the set of actual individuals).
Units / Individuals / Cases: individual elements of the population
Population Size: Number of units in the population
Parameter: (Numerical) Characteristic of the *population* that needs to be estimated [Ex. True Mean / Proportion, True difference of Means / Proportions, True Mean Difference, True s.d., True slope of the LSRL, etc.]
Estimate / Statistic: (Numerical) Characteristic of the *sample* used to estimate a Parameter [Ex. Sample Mean / Proportion, Difference of Sample Means / Proportions, Sample Mean Difference, Sample s.d., Sample slope of the LSRL, etc.]
Census: collecting data from entire population**.** A Census – collecting data from the *entire* population – is time-consuming and expensive; it can also be destructive, and often unreliable.
Sample: A subset of the population selected to make inferences about a Parameter **OR** the subset of the population from which results are generalized. It also refers to the actual set of individuals from whom responses were obtained [/ whose characteristics were measured].
Sample design: method used to choose/ select the sample from the population.
Sampling Frame: list of individuals from which the sample is selected. It is the list of

population units.

Study this: Advantage of Random Samples
1. It overcomes bias due to sampling / selection as it permits neither *self-selection by respondents* nor *researcher-induced bias* (from convenience sampling / judgment sampling). Randomizing gives all individuals of population an equal chance of being chosen.
2. Randomizing accounts for factors / characteristics in the population the researcher is aware of (gender, race, ethnicity, religion, age) but also overcomes factors, hidden and not immediately obvious (disability, height, weight, wealth). Randomizing takes care of influences of all features, attributes and characteristics of the population and provides a representative sample.
3. Randomizing allows drawing conclusions and making generalized statements about the population by creating samples not very different from one another; therefore, the samples must not be very different from the population either, permitting inferences to be made about the population.
4. Randomizing allows estimation of error/ uncertainly in drawing conclusions from the sample about the population since the use of chance in randomizing leads to the outcomes obeying the laws of probability. Randomizing, by itself, does *not* guarantee a representative sample, but probability-based methods can quantify the risk of an unrepresentative sample and allows the generalization with a certain degree of confidence.

Study and memorize this What are the 5 common Sampling Designs?
SRS, Stratified Sampling, Cluster Sampling, Multi-stage Sampling, Systematic Sampling

Study and memorize this Simple Random Sampling (SRS): is a method of choosing a sample such that every possible sample of a desired sample size has an equal chance of being chosen. For a sample of size *n*, each different subset of *n* must have an equal chance of being selected.

Study this To obtain a SRS of size, n, from a population of size, N:
1. Number the individuals in the population from 0 (or 00 or 000...) or 1 (or 01 or 001...) to N.
2. Choose *n* random numbers of 1 (or 2 or 3 …) digits in the domain specified in step **1.** and discard repeats.
3. Select the individuals corresponding to the random numbers to constitute the sample.

Study this Implementation: If N = 764 and n = 68, number the 764 individuals from

001 to 764. Choose 68 3-digit random numbers from 001-764, excluding repeats, from a random digit table or using a calculator, and select the corresponding individuals.

Study this Sampling With Replacement (SWR): is when the individual is placed back into the population pool after being chosen from the population and his information / characteristics are recorded. It allows the same observations to be repeated.
Sampling WithOut Replacement (SWOR): is when an individual who is chosen for inclusion into the sample, is not replaced / selected again. It leads to *n* distinct individuals in the sample from the population.

Note: When $N \geq 10n$, SWOR SWR i.e. if the sample size, *n*, is small relative to population size, N, there is little practical difference between SWR and SWOR.

Study and memorize this Stratified Random Sampling: consists of dividing the entire population into homogeneous groups or classes or categories called strata, with each stratum composed of individuals with similar characteristics or attributes or interests or opinions. Then, a SRS is taken from each stratum and pooled together.

Study this Examples of Stratification:

- for health survey → Smoker and Non-smokers → take SRS within both groups
- for unemployment rate, state of the economy, welfare dependence of adults → North, South, East, West, South-West, Mid-west regions → take SRS of adults within categories of states
- for divorce-rate, educational attainment, teenage pregnacy → Republican, Democrat-leaning states → take SRS of adults within categories of states
- for test-scores → public, private schools OR large, mid-sized, small schools OR unionized, non-unionized schools → take SRS of schools in each category
- for SAT scores → Prep classes, No Prep classes → take SRS of students in each category
- for happiness-level → frequency of Church attendance [<1, 1-3, >3 times per week] → take SRS of adults in each category
- in school → departments of Mathematics, Science, English, etc. → take SRS of teachers from each department

Study this Implementation: If N = 3800 with N1 = 1000 and N2 = 2800 with n1 = 100 and n2 = 280...number the individuals in Stratum 1 from 000-999; choose 100 3-digit random numbers, excluding repeats, from a random digit table and select the corresponding individuals. Repeat the process to select 280 individuals from the 2nd stratum population of 2800.

Study this **How should the Stratification Factor be chosen i.e. on what basis should the researcher choose to divide the population units?**
Choose a factor that is associated with the response (variable) such that those in different strata systematically differ in their responses or characteristics while those in the same strata have similar responses or characteristics. For example, if we seek the information from college students about a proposed fee increase, we may stratify by Major [Engineering, Medicine, Law, Arts, Sciences] or by Income level since those factors are related to the response of the students. Students of different majors or income levels may consistently differ in their attitude to the proposal.

Study this **When is a Stratified Random Sample preferable?**
A stratified random sample is preferable when the population *can* be divided into homogeneous groups with members of the same stratum responding similarly / sharing characteristics while those *between* groups responding differently to the question of interest. Then, with the variability within groups being minimized, the overall S.E.(Estimate) is lower than if a SRS were taken from the entire population. Also, stratifying permits a ready examination and comparison of estimates from different subsets of the population.

Study and memorize this **Advantages of Stratified Sampling over an SRS:**
1. It permits a ready comparison of the characteristics of different subsets of the population.
2. Using SRS, there is a possibility that certain samples may be unrepresentative as smaller subsets of the population may be entirely excluded, by chance. Using Stratified Sampling, *no* important subset of the population is underrepresented.
3. With stratified sampling, the homogeneity within each stratum reduces the variability of the response thereby decreasing the S.E.(Estimate), leading to more precise estimates. Consequently, we are likelier to find the observed differences to be statistically significant when they indeed are...thereby reducing a Type I Error and increasing the Power of the test.

Study and memorize this **To obtain a Stratified Random Sample from a population of size N**:
Divide the population into non-overlapping homogeneous strata. Then, simply draw an SRS from each stratum...and "pool" together to form the sample.

Study this **To obtain a Systematic Sample of size *n* from a population of size N:**
1. Number the individuals in the population. Calculate $k = N/n$, implicitly forming n successive lists or groups of k individuals.
2. Select 1 individual randomly from the first k individuals only.

3. Choose every *k*th individual beyond automatically.

Study this Implementation: To take a systematic sample of n = 150 students from N = 1500 exiting the Gym, form 150 groups of size 1500/150 = 10 students each. Select 1 randomly from the 1st group of 10 students only [using a random digit table]. Then, automatically choose every 10th student exiting...since the others will fall in each of the other groups → systematic sample of 150 students!

Study and memorize this Cluster Sampling: consists of dividing the population into non-overlapping relatively heterogeneous groups called clusters – each cluster reflecting the population. A SRS of certain number of clusters is taken, and then all members of the chosen clusters is automatically included in the sample. Ideally, clusters mirror the population characteristics, so that a small number of clusters leads to representative samples. The larger the size of each cluster, fewer clusters need to be randomly selected.

Study this Examples:

- in schools → certain classrooms are clusters → take SRS of classrooms and all students within chosen classrooms
- in school → WASC teams [with teachers from all departments!] are clusters → take SRS of teams and choose all teachers within selected teams
- in large used-car lots → sections of lot [A, B, C...] are clusters → take SRS of sections and all cars within selected sections
- in my bookcases at home → certain shelves consist of fiction books of all genres...reflecting the "population" of all fiction books → take SRS of shelves ["clusters"] and sample all books from the selected shelves

Study this Implementation: I have 16 bookcases, each with 6-7 shelves = ~100 shelves. To estimate the total number of books, number the shelves ["clusters"] from 00-99 and choose 10 2-digit random numbers, discarding repeats, from a random digit table and select all the books in the corresponding shelves to constitute the sample.

Study and memorize this Key Differences between Stratified and Cluster Sampling:

- Strata are homogeneous; clusters are heterogeneous.
- We take a SRS within each strata; we take SRS of clusters and select all units within chosen clusters.
- Strata are usually man-made; clusters tend to occur naturally

Study this Two-stage / Multi-stage Sampling: consists of randomly selecting

heterogeneous clusters and then choosing a SRS of individuals within each cluster. *Alternately,* it may also refer to *any* multi-step random sampling procedure with different steps involving different sampling designs. E.g. a SRS of clusters followed by Systematic Random Sampling of individuals.

Study this Are ALL Random Sampling Schemes equally "good"?
All random sampling methods – SRS, Stratified Random Sampling, Systematic Random Sampling, Cluster Random Sampling and Multi-Stage Random Sampling – yield unbiased estimates of the Parameter of interest. That is, in repeated samples, the "average" of all the estimates *would* be the Parameter: E(Estimate) = Parameter.

With random sampling, there is no under-estimation or over-estimation of the parameter, assuming, of course, the absence of *other* biases: Selection Bias (no subsets of the population are being excluded), Measurement / Response bias (flaws in the question, interviewer) and Non-response (refusal of a significant proportion of the population to respond).

Still, depending on characteristics of the population, the variability of the estimates would differ. In other words, some random sampling designs produce less variable estimates than others [SE(Estimate)]. Other random sampling designs may be easier to implement. And certain others might be cheaper to perform!

There is no *one* "best" random sampling design for all situations!

Study and memorize this: Bias: A sampling design is biased when it systematically / consistently / predictably / reliably favours certain outcomes on account of subsets of the population being over- or under-represented. This leads to non-representative samples so that the parameter of interest [% / μ] is over-/under-estimated.

Concept Note: *It's hard to judge if a particular sample is representative but it is possible to determine if a sampling method is biased.*

Study and memorize this What are 3 main sources of Bias?
There are 3 main sources of Bias:
Sampling or Selection Bias
Measurement or Response Bias
Non-Response Bias

Study and memorize this What is Sampling / Selection Bias?
Sampling or Selection Bias occurs due to a flaw in the sample *design.*

Study and memorize this **What is the source of Sampling or Selection Bias?**
There are 3 basic types of Sampling or selection bias: convenience sampling, voluntary response and under-coverage.

Study and memorize this **How does Sampling or Selection Bias arise?**
A. Sampling Bias may arise from using a non-random sample:

- **Convenience Sampling**: occurs when individuals close to the researcher are selected and differ from those that weren't, leading to an unrepresentative sample.
- **Voluntary Response Sampling**: is a type of sampling / selection bias wherein motivated or passionate or opinionated individuals – those with strong opinions – volunteer their opinions [e.g. telephone call-in; volunteer for a taste/smell test, etc] so that those that do respond are likely to systematically differ from those that chose not to respond.

  Caution! The term is usually <u>not</u> used when individuals are contacted, by, say, a questionnaire / survey in the mail / by phone since in *that* case, the respondents are <u>not</u> "volunteering" their opinions.

B. Sampling Bias may also arise from using a random-sample but excluding a subset of the population whose responses or characteristics differ from those selected. This is also called **Under-coverage.** E.g. randomly contacting individuals on their cell-phones but incidentally discarding those without cell-phones [those with land-lines alone, the poor, those in rural areas, the old, etc].

Study and memorize this **What is Measurement / Response Bias?**
Measurement / Response Bias occurs when the responses are biased / skewed in a certain direction on account of

- question characteristics [confused wording, insufficient choices, leading Q]
- interviewer characteristics [ethnicity, religion, clothing, appearance, gender]
- dishonesty from the issue being "sensitive" and bearing positive or negative connotations [smoking (negative), voting (positive), cheating (negative), community service (positive)

Study and memorize this **What is Non-Response Bias?**
Non-response Bias: occurs when a significant % of those contacted [at home, telephonically, via a mail-in survey] do not answer [outright refusal, inability to be reached], and their characteristics [potentially] differ from those from whom responses were obtained.

Note: While Sampling / Selection Bias occurs from flaw in Sampling Design leading to non-representative samples , Measurement / Response Bias occurs *after* the samples have been selected and arise from a flaw in collecting accurate responses from the individuals. This implies that bias may arise even if the sampling *method* is unbiased i.e. a random sample was selected, but the bias arose *subsequently* e.g. response or non-response biases. **Random samples do *not* insure absence of *all* biases!**

For Qs about describing Bias, follow the rubric below:
1. Identify the Population →
2. Identify the Sample →
3. Identify and Explain the bias(es) →
- Sampling / Selection Bias: Random or Non-Random (Voluntary Response / Convenience). **Identify the *excluded* subsets of the Population, if any**
- Measurement / Response: "flaws" in the Q or interview process
- Non-Response: insufficient % of the randomly *contacted* sample responding

4. Explain, if possible, the *direction* of the bias.

1a) When is a sampling method said to be biased?
Solution. When certain outcomes are systematically favoured **OR** when the method systematically favours certain outcomes...due to a subset of the population being over-represented **OR** due to a flaw in the response-collection...resulting in the parameter being under- or over-estimated.
b) What are 3 broad types of bias?
Solution. Sampling, Measurement, Non-response
c) What are the 3 major types of Sampling / Selection bias?
Solution. Undercoverage, Convenience, Voluntary-response
d) What are the major sources of Measurement / Response bias?
Solution. Flawed Q characteristics; Flawed interviewer characteristics; Sensitive issue or one having positive or negative connotations
e) What is non-response bias? Be detailed and specific: dont merely define "non-response" → explain the *bias*.
Solution. When a minuscule proportion of individuals respond and *likely systematically differ from those that did not*.
f) What is undercoverage bias? Be detailed and specific: dont merely define "undercoverage" → explain the *bias*.
Solution. When a subset of the population is excluded or underrepresented and *likely systematically differs from those that weren't*.
g) What is convenience sampling?

Solution. When opinions / characteristics of those in proximity to the researcher are selected non-randomly and *systematically differ from those that weren't chosen*.
h) What is the objective of conducting surveys or sampling? Why conduct surveys?
Solution. To use the statistic derived from the sample to estimate the parameter of interest **OR** to generalize the results to the population of interest **OR** to draw conclusions / make inferences about the parameter.

i) What is the objective of observation studies?

Solution. To determine if there is an association between the explanatory and response variables.
j) What is the objective of experiments?
Solution. To determine if there is a causal relationship between the explanatory and response variables.
k) How is CRD implemented?
Solution. Randomly assign the *n* subjects to the *k* different treatments.
l) How is an RBD implemented?
Solution. Divide the subjects into non-overlapping homogeneous blocks and then randomly assign the treatments to the subjects within each block.
m) What are the 2 key objectives while designing an experiment? [**Tip!** Friday's CW Notes / Email Notes above!]
Solution. To insure that the conditions across the treatment groups are comparable and to reduce the S.E.(Estimate) **OR** reduce the variability of the responses.
n) What are the 4 aspects of a well-designed experiment?
Solution. Randomization, Replication, Direct Control and Use of a Comparison Group.
o) What is voluntary-response bias? Be detailed and specific: dont merely define "voluntary-response" → explain the bias i.e. the specific *problem*.
Solution. Voluntary-response bias occurs when individuals volunteer their responses so that the sampling is non-random and those with strong opinions about the issue are likeliest to respond and systematically differ from those that don't believe as passionately.