Observations vs Experiments

measures specific trait but does not modify subjects

Apply a treatment and then measure the effect on the subjects

objective of sampling is to generalize properties of the population.

Random - each member of the population have an equal probability of being picked or selected in the sample.
chance

a.) Simple Random Sampling.

Each group of size "n" has an equal chance of being selected.

4 common ways to achieve this -

① Convenience Sampling.
- use the results that are easy to get. (Not random)

eg:- what is peoples favourite movie right now? - you ask people on your phonelist
This is not going to be random.

② Systematic Sampling.
- put a population in order and select every "kth" member.

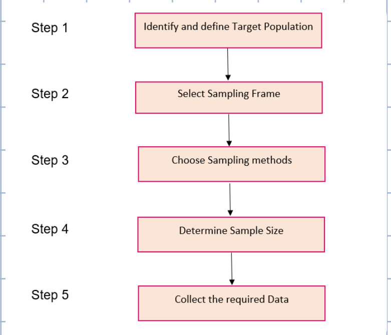
③ Stratified Sampling.   Strata = layers (stratosphere)
- In systematic sampling, there is a chance that you will miss out on a certain group.
In stratified, you divide the population into subgroups based on a characteristic, then sample each subgroup.

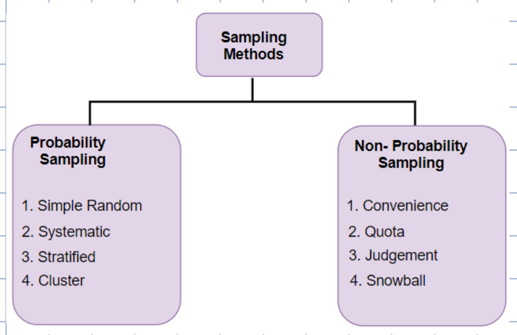
④ Cluster Sampling.
- divide population into groups that need not have any characteristics that are similar.
- Randomly selected a certain number of clusters and then sample or collect data from the entire cluster.
eg - Divide a class into 4 groups and pick 2 groups randomnly and then sample data.

Sampling Error -
↳ Difference in characteristics between your sample and the population.

https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

https://
ry.com/statistical-
resampling/


Step 1 Identify and define Target Population
Step 2 Select Sampling Frame
Step 3 Choose Sampling methods
Step 4 Determine Sample Size
Step 5 Collect the required Data


Sampling Methods

Probability Sampling
1. Simple Random
2. Systematic
3. Stratified
4. Cluster

Non- Probability Sampling
1. Convenience
2. Quota
3. Judgement
4. Snowball

machinelearningmaste
sampling-and-

sampling and resampling ?

Sampling with replacement and without replacement .

With replacement, → same observation can repeat . If sampling 2 events the occurence of first event does not impact other event . covariance is zero .
i-e samples are independent .

eg. pick 2 samples from a set with below numbers. {1,2,3, 4,5,6}

(1,1) can appear --→ repeated values .

$P = 1/6 \times 1/6$.

Without replacement, samples cannot repeat . covariance is non-zero . events depends on one another -

eg:- $P(1,2)$

$P = 1/6 \times 1/5$

basically, when you pick the first observation, you remove it from the original population .

Random - number generators are used to create random samples . There is no such thing as "truly" random . Hence called pseudo-random number generators (PRNG)

"seed" variable is used to initialize this PRNG algorithms.

Python random() module implements these algorithms .

One thing that traditional computer systems aren't good at is coin flipping," says Steve Ward, Professor of Computer Science and Engineering at MIT's Computer Science and Artificial Intelligence Laboratory. "They're deterministic, which means that if you ask the same question you'll get the same answer every time. In fact, such machines are specifically and carefully programmed to eliminate randomness in results. They do this by following rules and relying on algorithms when they compute."

You can program a machine to generate what can be called "random" numbers, but the machine is always at the mercy of its programming. "On a completely deterministic machine you can't generate anything you could really call a random sequence of numbers," says Ward, "because the machine is following the same algorithm to generate them. Typically, that means it starts with a common 'seed' number and then follows a pattern." The results may be sufficiently complex to make the pattern difficult to identify, but because it is ruled by a carefully defined and consistently repeated algorithm, the numbers it produces are not truly random. "They are what we call 'pseudo-random' numbers," Ward says.

For most applications, a pseudo-random number is sufficient, he adds. "For example, if you want to do a random sampling of a large set of data, you'll need numbers to feed into the program so that the samples are more or less evenly distributed. Using pseudo-random numbers is perfectly acceptable in this case because there's no quantitative advantage in the degree of randomness." Similarly, a CD player in "random" mode is probably really playing in pseudo-random mode, with a pattern that is discernible if you listen carefully enough.

Not all randomness is pseudo, however, says Ward. There are ways that machines can generate truly random numbers. And the importance of true randomness is not to be underestimated, he adds. "If you go to an online poker site, for example, and you know the algorithm and seed, you can write a program that will predict the cards that are going to be dealt." Truly random numbers make such reverse engineering impossible, he adds. There are devices that generate numbers that claim to be truly random. They rely on unpredictable processes like thermal or atmospheric noise rather than human-defined patterns. The results might still be slightly biased towards higher numbers or even numbers, but they're not generated by a deterministic algorithm

https://medium.com/swlh/random-functions-a4f36b1dfd8f

Linear congruential method implemented in random() function .

$$A_{n+1} = (Z \times A_n + I) \bmod M$$

Z, I, M all constant numbers .

This process is fairly deterministic / predictable .