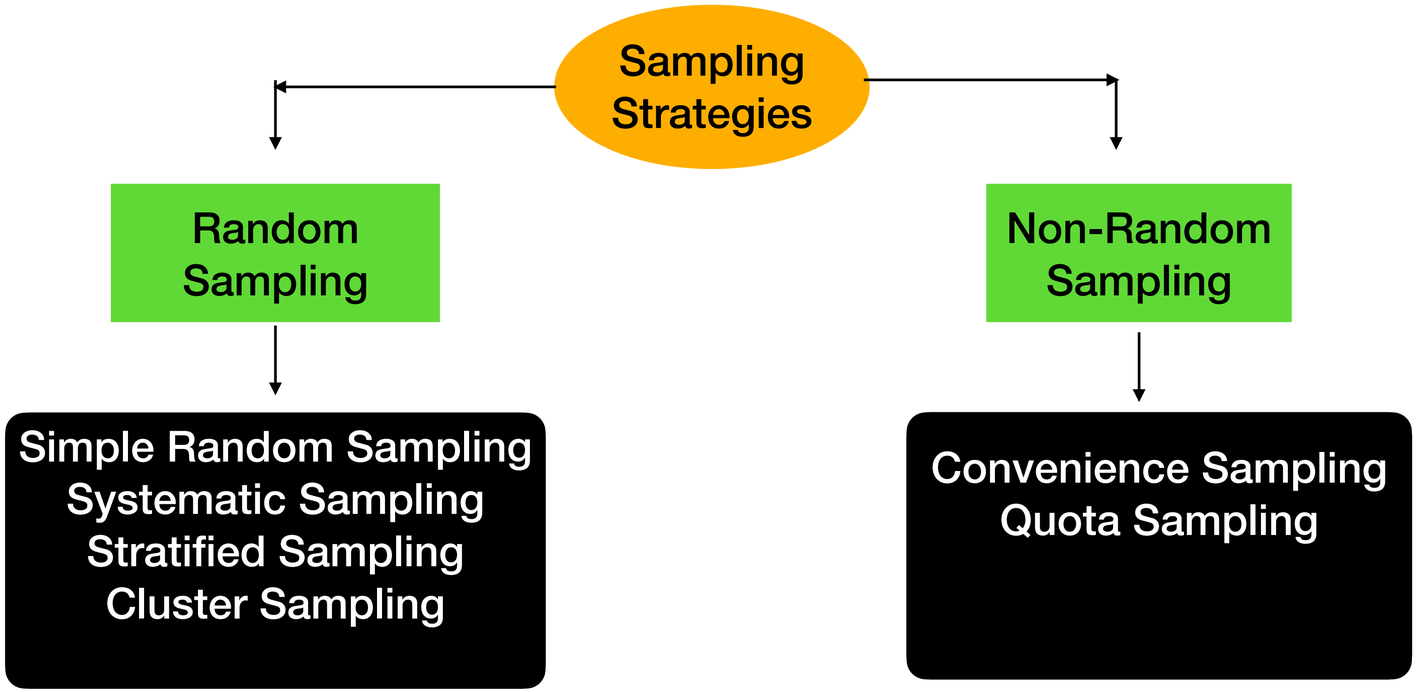
Overview of Sampling Strategies and CLT

Mainak Thakur

01.09.2020

Sampling Strategies

- Sampling strategies refer to different ways to choose members from the population. Biases may occur due to non-random selection leading to an unrepresentative sample of the population.
- Sampling strategies are broadly divided in two ways:



Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Random Sampling: Simple Random Sampling

- Every member of the population is equally likely to be selected in the sample. Random number generators are often used for this.
- Though this strategy is preferred but it is very difficult to carry out, time consuming and tedious.
- It provides a fairly good representative of the population.
- It is often used when the population members are similar in nature with respect to the study

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Random Sampling: Systematic Sampling

- Firstly, in an appropriate manner suitable for the study, population members are put in some order. Then a starting point is decided or randomly selected, and every n^{th} member is chosen in the sample.
- For example, if I want to do a survey among all IIITS students, I arrange them in alphabetical order and pick a random starting point, say 5th. Then after that I select every 8th student to take the survey.
- It is easier to implement than simple random sampling but it is less random than simple random sampling.

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Random Sampling: Stratified Sampling

- The population is divided into groups called strata. The overall sample consists of some members from every strata. The members from each strata are chosen randomly.
- For instance, a population consist of males and females. The study is regarding heights of human beings.
- It is used when population is heterogeneous and contains several different groups. It ensures that members from each group will be represented in the sample thereby guaranteeing good representation in the sample.
- In political survey, while choosing the sample voters it is essential to maintain the diversity of the population, the surveyor is required to include participants of various minority groups such as race or religion, according to their proportionality in the total population.

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Random Sampling: Cluster Sampling

- Population is divided into clusters (usually geographical) where the population within a cluster should ideally be as heterogeneous as possible, but there should be homogeneity between clusters. Each cluster should be a small-scale representation of the total population. Some clusters are chosen using simple random sampling. The members in each selected cluster are then sampled. Sometimes simple random subsample of members is chosen within each of these clusters.

- Cluster sampling is often used to estimate mortalities due to epidemic, wars etc.

- It is easy and convenient.

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Non-random Sampling: Convenience Sampling

- Convenience sampling is quite easy to perform, but it is discouraged. Mostly readily available data is used in a non-random manner such as the first person the surveyor runs into.
- For example, a media group is collecting the opinion of the residents of the city on some issue and the journalists ask people whoever they find on the street.
- It is convenient and cheap but the degree of representativeness is questionable.

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Non-random Sampling: Quota Sampling

- There should be subgroups in the population and the surveyor is provided some quota to fill from particular subgroups. The surveyor needs to ensure that the final sample meets the required quota criteria.
- Quota in terms of say 10 males and 20 females teenager are required to be surveyed.
- In quota sampling, non-random sample selection is used such as convenient sampling and this can be quite unreliable.

Ref: http://www2.hawaii.edu/~cheang/Sampling%20Strategies%20and%20their%20Advantages%20and%20Disadvantages.htm
https://people.richland.edu/james/lecture/m170/ch01-not.html
http://www.statstutor.ac.uk/resources/uploaded/13samplingtechniques.pdf
https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.628.7338&rep=rep1&type=pdf
wikipidea

Sample variance is unbiased estimator for population variance

Let $Y_1, Y_2, \ldots, Y_n$ are n i.i.d random sample from an unknown population with mean μ and variance σ^2. Then $S^2 = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2$ is an unbiased estimator of σ^2.

$$S^2 = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)$$

$$= \frac{1}{n-1}\left(\sum Y_i^2 - 2\bar{Y}\sum Y_i + n\bar{Y}^2\right) = \frac{1}{n-1}\left(\sum Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2\right)$$

$$= \frac{1}{n-1}\left(\sum Y_i^2 - n\bar{Y}^2\right) = \frac{1}{n-1}\sum Y_i^2 - \frac{n}{n-1}\left(\frac{1}{n}\sum Y_i\right)^2$$

$$= \frac{1}{n-1}\sum Y_i^2 - \frac{1}{n(n-1)}\left(\sum Y_i\right)^2$$

Ref: https://web.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf

Contd.

$$E(S^2) = \frac{1}{n-1}\sum E(Y_i^2) - \frac{1}{n(n-1)}E[(\sum Y_i)^2]$$

$$= \frac{1}{n-1}\sum (Var(Y_i) + [E(Y_i)]^2) - \frac{1}{n(n-1)}E[(\sum Y_i)^2]$$

Let, $Z = \sum Y_i$, $E(Z^2) = Var(Z) + [E(Z)]^2$

$$= \frac{1}{n-1}\sum (Var(Y_i) + [E(Y_i)]^2) - \frac{1}{n(n-1)}\{Var[(\sum Y_i)] + (E[\sum Y_i])^2\}$$

Ref: https://web.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf

Contd.

$$= \frac{1}{n-1} \sum (\sigma^2 + \mu^2) - \frac{1}{n(n-1)} \{ \sum Var(Y_i) + (\sum E[Y_i])^2 \}$$

[by independence of samples]

$$= \frac{1}{n-1} (n\sigma^2 + n\mu^2) - \frac{1}{n(n-1)} \{ n\sigma^2 + n^2\mu^2 \}$$

$$= \frac{1}{n(n-1)} (n^2\sigma^2 + n^2\mu^2 - n\sigma^2 - n^2\mu)^2 \} = \sigma^2$$

Ref: https://web.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf

Central Limit Theorem

- Idea: Whatever the population distribution is, if we draw large number of i.i.d samples, the distribution of the sample mean tends to be Gaussian with mean same as the population mean, and its standard deviation shrinks as n increases.

- Question: Take i.i.d sample $Y_1, Y_2, \ldots, Y_n$ from a population with mean μ and variance σ^2. Find $E(\bar{Y})$ and $Var(\bar{Y})$.

Ref: http://homepages.math.uic.edu/~bpower6/stat101/Sampling%20Distributions.

Central Limit Theorem

- Idea: Whatever the population distribution is, if we draw large number of iid samples, the distribution of the sample mean tends to be Gaussian with mean same as the population mean, and its standard deviation shrinks as n increases.

- Question: Take i.i.d sample $Y_1, Y_2, \ldots, Y_n$ from a population with mean μ and variance σ^2. Find $E(\bar{Y})$ and $Var(\bar{Y})$.

- $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \sigma^2/n$

Ref: http://homepages.math.uic.edu/~bpower6/stat101/Sampling%20Distributions.pdf

CLT Statement

Let $Y_1, Y_2, \ldots, Y_n$ are i.i.d random variables each having mean μ and variance σ^2. Suppose, $\bar{Y}_n$ be the average of $Y_1, Y_2, \ldots, Y_n$. For large n, $\bar{Y}_n$ tends to follow $N(\mu, \sigma^2/n)$. $Z_n = \dfrac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$ tends to follow standard normal distribution.

- $\bar{Y}_n$ is approximately normal with same mean as Y but has a smaller variance
- CLT enables us to approximate average of any large i.i.d random variables by a normal distribution
- If $S_n = Y_1 + Y_2 + \ldots + Y_n$, then what is the distribution of S_n? E(S_n)=? Var(S_n) =?

Ref: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6b.pdf

CLT Statement

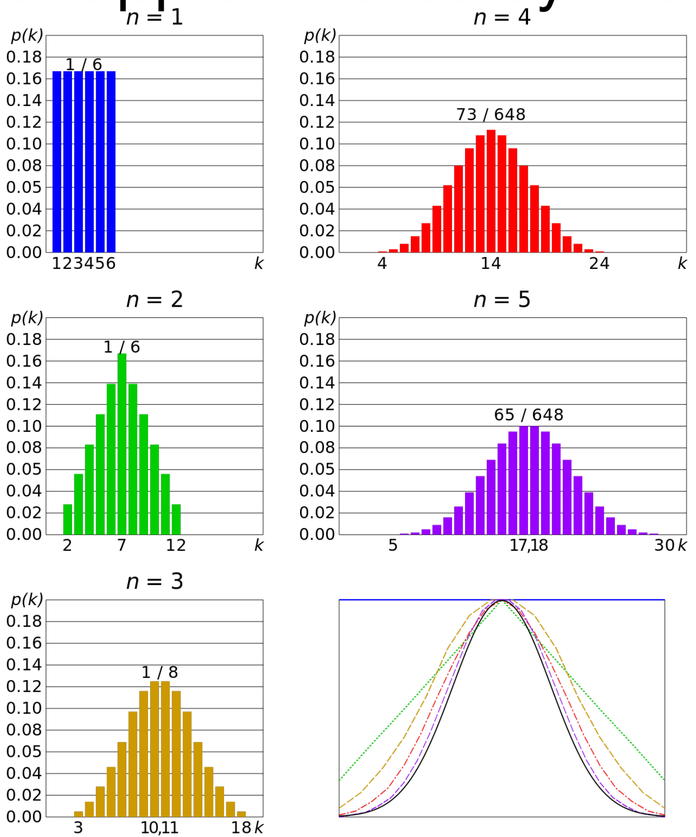
- Let $Y_1, Y_2, \ldots, Y_n$ are i.i.d random variables each having mean μ and variance σ^2. Suppose, $\bar{Y}_n$ be the average of $Y_1, Y_2, \ldots, Y_n$. For large n, $\bar{Y}_n$ tends to follow $N(\mu, \sigma^2/n)$.
  $Z_n = \dfrac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$ tends to follow standard normal distribution.

- $\bar{Y}_n$ is approximately normal with same mean as Y but a smaller variance
- CLT enables us to approximate average of any large i.i.d random variables by a normal distribution
- If $S_n = Y_1 + Y_2 + \ldots + Y_n$, then for large n, S_n follows $N(n\mu, n\sigma^2)$.

Ref: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6b.pdf

CLT: Example

You can think of a simple example of the central limit theorem. Roll as many as identical, unbiased dice and note down the sum (or average). Repeat this procedure for a long time and you have a distribution of sum (or average). Put that into a histogram and see. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution.



Ref: Wikipedia

CLT: Example

- A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ = 205 pounds and standard deviation σ = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Ans: n = 49, μ = 205, σ = 15. The probability that the total weight of these 49 boxes is less than 9800 pounds is $P(S_n < 9800) = P(Z < \frac{9800 - 49 \times 205}{\sqrt{49} \times 15})$ = P(Z<-2.33) =0.0099

.

Ref: http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_central_limit_theorem.pdf