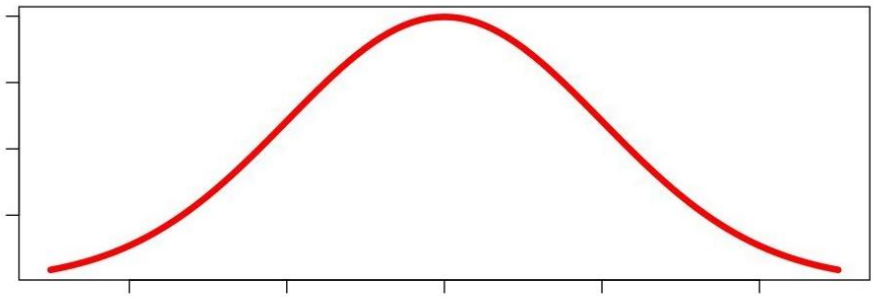
Recap

- What is a **variable** (say, y)?
- What is the **probability density function** ‘f(y)’ of the variable ‘y’ ?
- What is the **function g(y)** defined over the variable ‘y’?
- What is **‘sample’ vs ‘population**’ of ‘y’?
- What is the **‘expected value’** or expectation E[]? Say E[y] or E[g(y)]?

- **Given the PDF** of a variable ‘y’, how do we figure out **mean, median, mode, variance**, …

Standard Normal PDF



‘General Form’

$$f(y) = \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-b}{a}\right)^2\right), \qquad -\infty \le y \le \infty$$

Substitute, $z = \frac{y-\mu}{\sigma}$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \qquad -\infty \le z \le \infty$$

$$\mu = 0, \sigma = 1$$

Standard Normal Probabilities

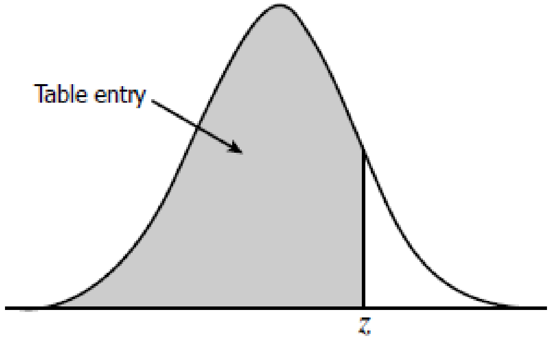
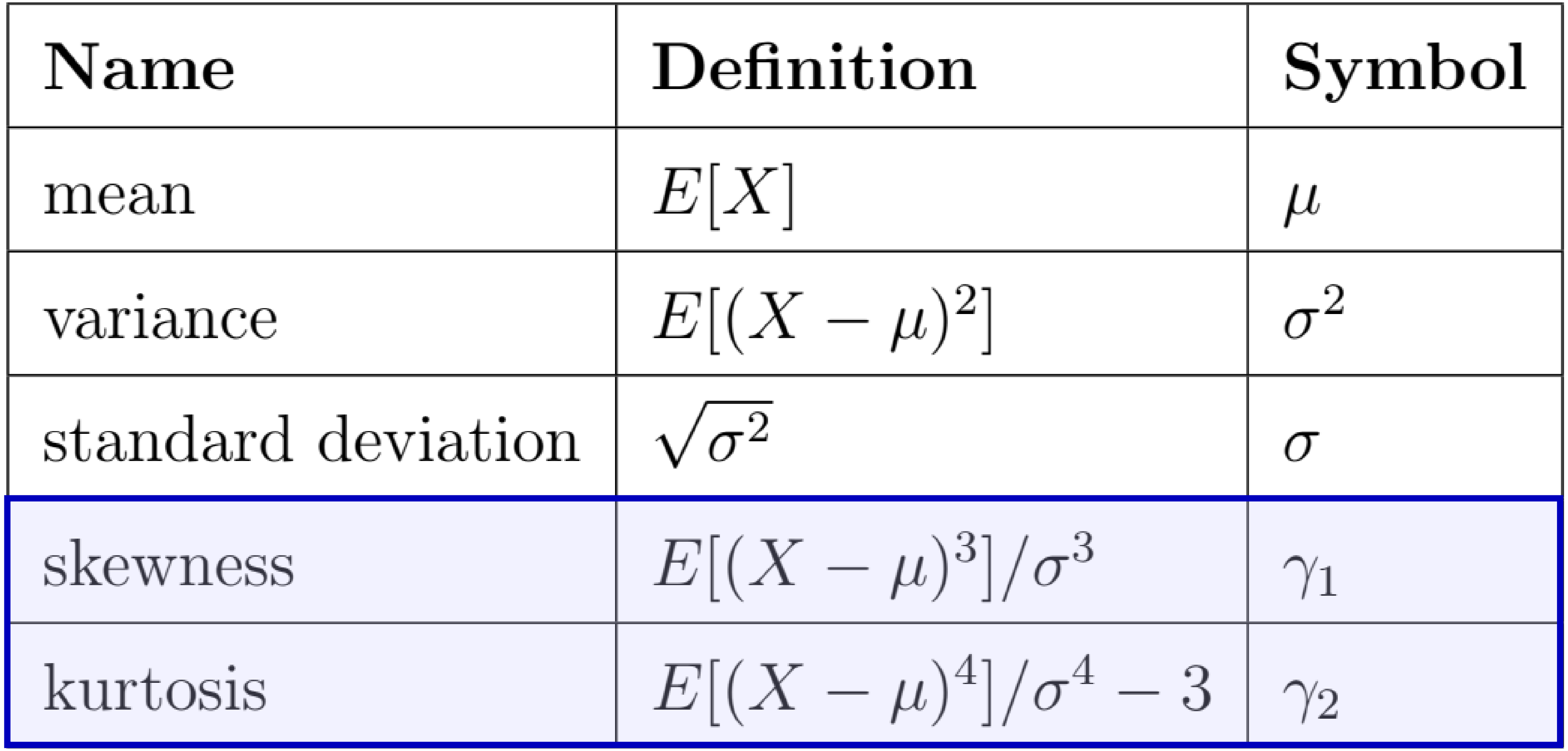


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Other Statistical Parameters

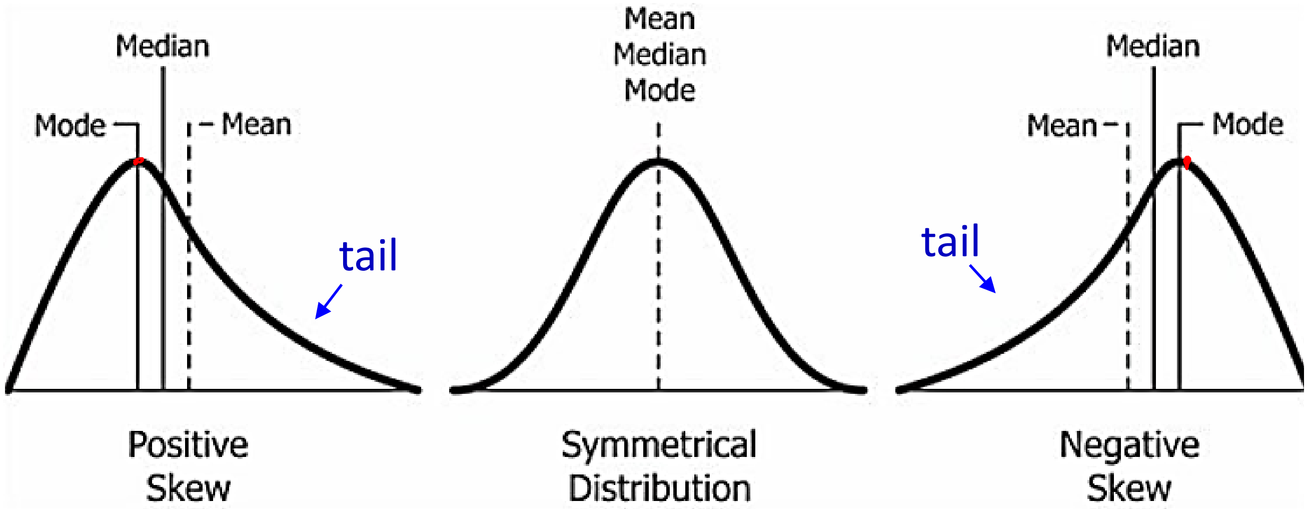


Name	Definition	Symbol
mean	$E[X]$	μ
variance	$E[(X-\mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X-\mu)^4]/\sigma^4-3$	γ_2

Skewness

skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1

- Skewness can tell us about symmetry: . It measures the lack of symmetry in data distribution.
- It is the degree of distortion from the symmetrical bell curve or the normal distribution
- It differentiates extreme values in one versus the other tail
- A symmetrical distribution will have a skewness of 0



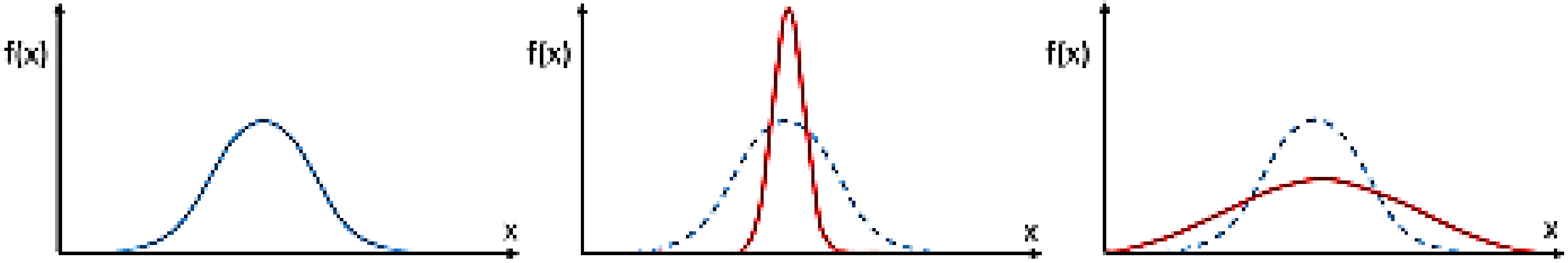
- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed), the data are moderately skewed.
- If the skewness is less than -1 (negatively skewed) or greater than 1 (positively skewed), the data are highly skewed.

NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

Kurtosis

kurtosis	$E[(X-\mu)^4]/\sigma^4 - 3$	γ_2

- Kurtosis is all about the tails of the distribution — the peakedness or flatness.
- It is used to describe the extreme values in one versus the other tail.
- It is actually the measure of outliers present in the distribution.
- High kurtosis data has heavy tails or outliers.
- Low kurtosis data has light tails or lack of outliers.



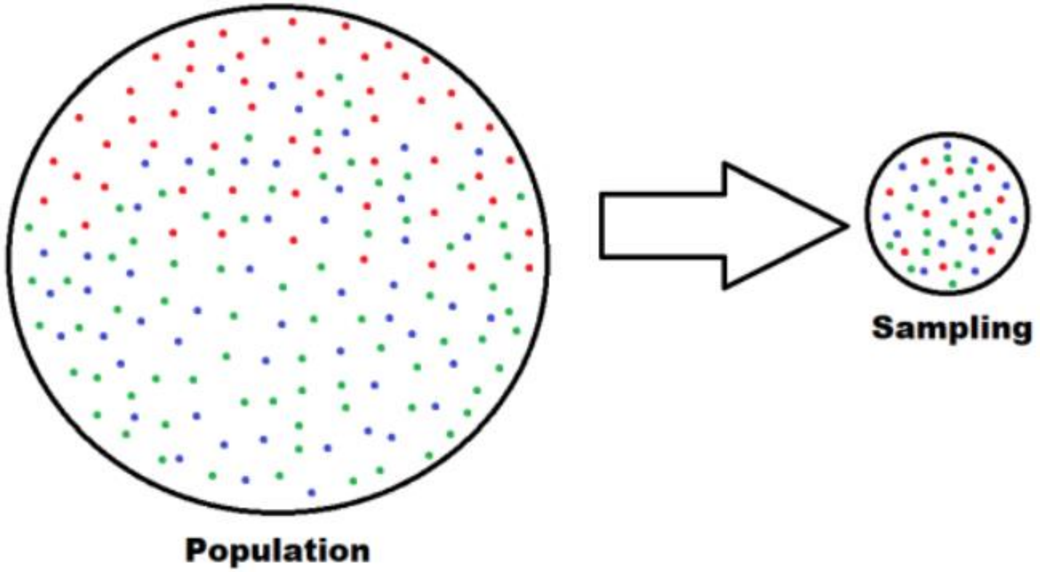
Zero kurtosis
Gaussian distribution

Positive kurtosis

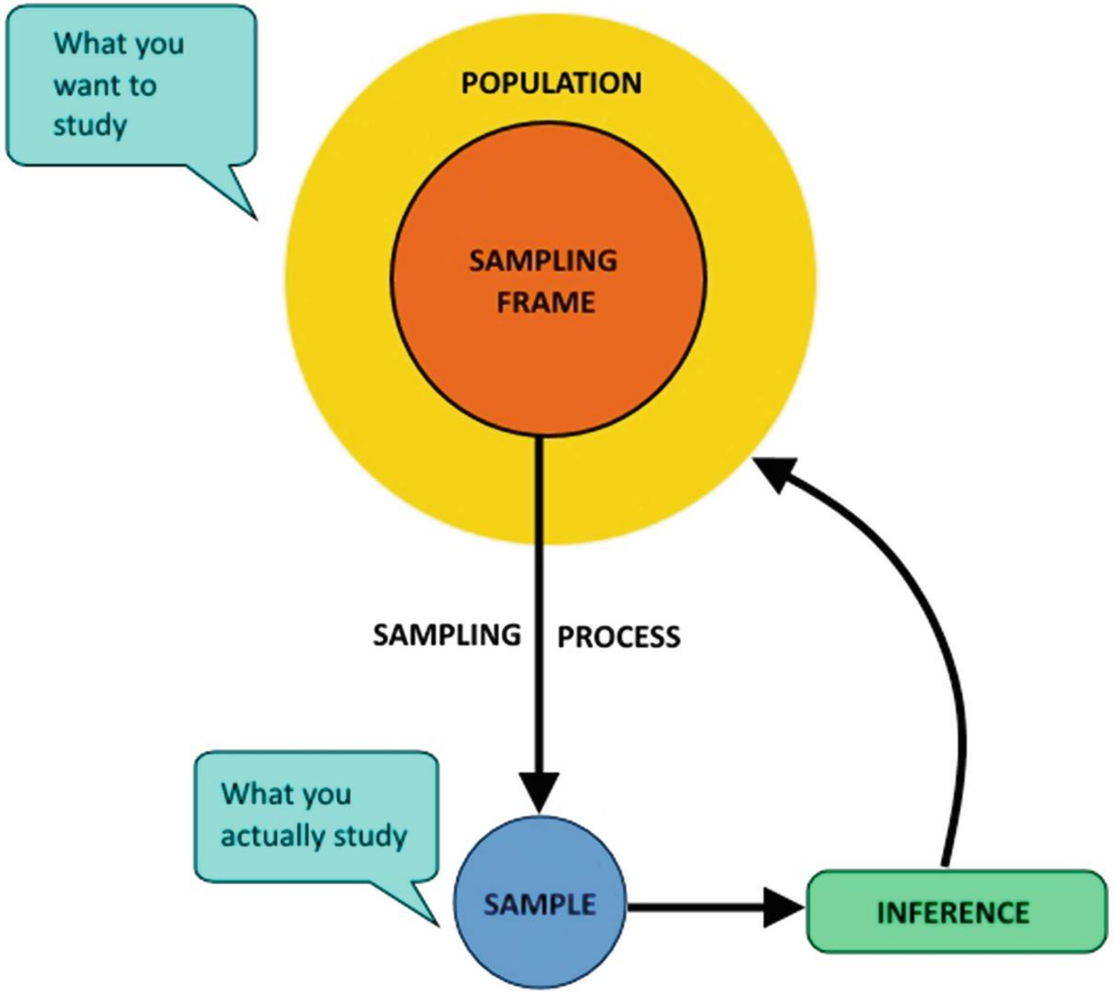
Negative kurtosis

Data Sampling

What is Sampling?



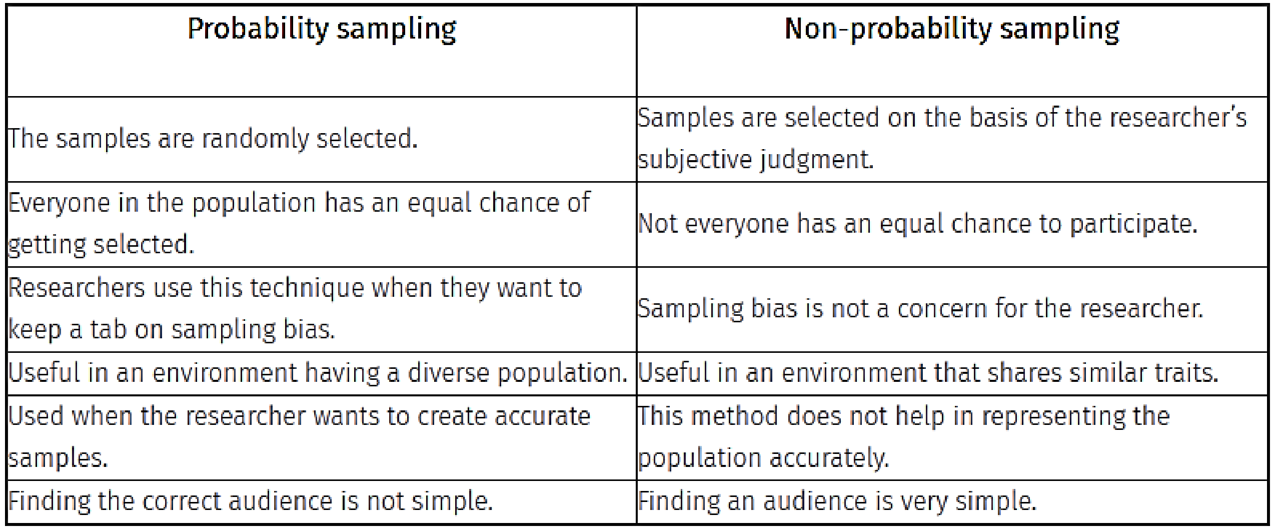
https://images.app.goo.gl/cvN1oMakejPsdhva7



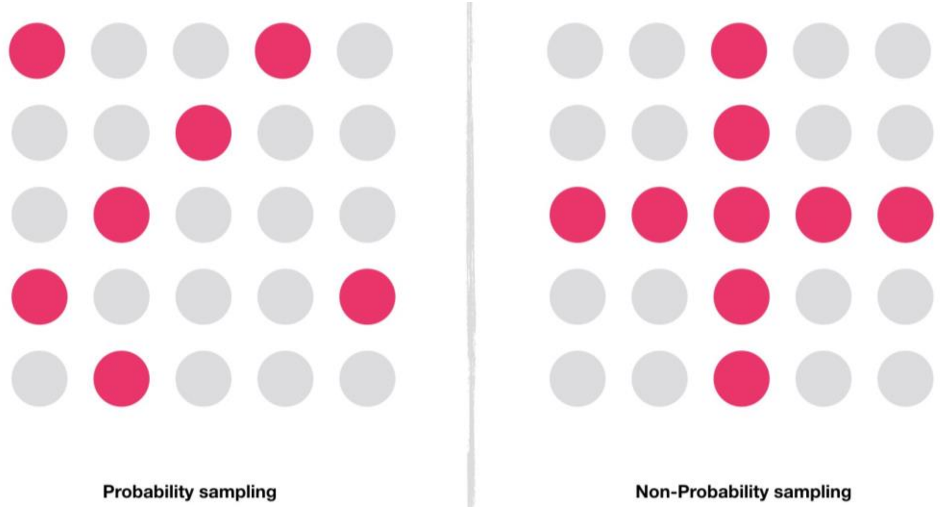
NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

Sampling Methods

- **Probability Sampling**
  - When *each entity* of the population has *a finite, non-zero probability* of being into the sample
  - Sampling procedure involves random sampling and without bias
- **Non-probability Sampling**
  - Some units of the population have zero chance of selection
  - OR probability of selection cannot be determined accurately



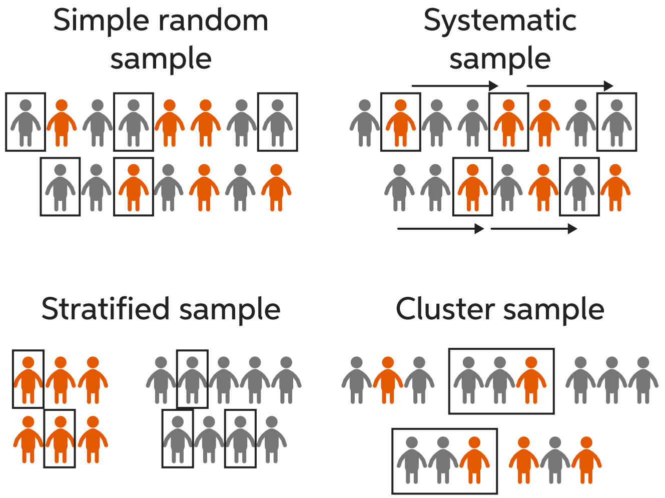
Probability sampling	Non-probability sampling
The samples are randomly selected.	Samples are selected on the basis of the researcher's subjective judgment.
Everyone in the population has an equal chance of getting selected.	Not everyone has an equal chance to participate.
Researchers use this technique when they want to keep a tab on sampling bias.	Sampling bias is not a concern for the researcher.
Useful in an environment having a diverse population.	Useful in an environment that shares similar traits.
Used when the researcher wants to create accurate samples.	This method does not help in representing the population accurately.
Finding the correct audience is not simple.	Finding an audience is very simple.



https://www.questionpro.com/blog/probability-sampling/

Probability Sampling

- **Simple Random Sampling**
  - Each subject/unit selected at random, independent from each other
  - Typically done when the population is large
- **Systematic Sampling**
  - Arrange the population in some order, and pick a unit at regular intervals from the list
  - When population is logically homogenous
  - E.g. You ask every 10th customer entering a shop about his purchase habits
- **Stratified Sampling**
  - Population divided into groups/stratas based on some characteristics
  - Then population is sampled randomly within each strata
  - E.g. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated subset of the population
- **Cluster Sampling**
  - Random sample is drawn from a cluster of data, rather than individual samples
  - E.g. An NGO wants to create a sample of girls across five neighboring towns to provide education. Using single-stage sampling, the NGO randomly selects towns (clusters) to form a sample and extend help to the girls deprived of education in those towns.



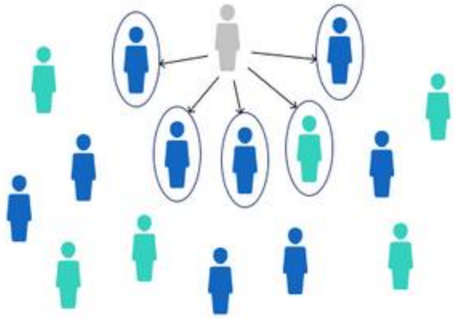
www.chegg.com

NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

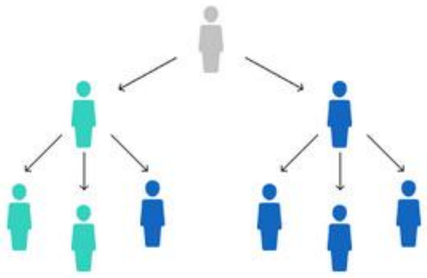
Non-Probability Sampling

- **Convenience Sampling**
  - Each subject/unit is selected on the basis of convenience, availability, reach, etc.
  - Typically during preliminary research

Convenience sample



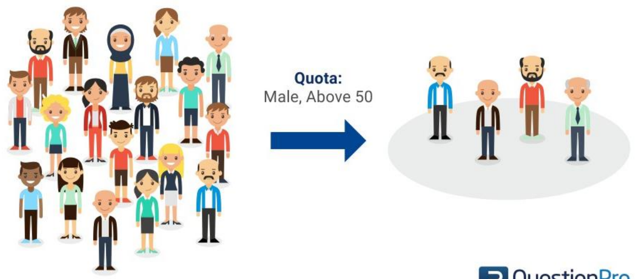
Snowball sample



- **Snowball Sampling**
  - One unit refers you to the next unit
  - Costs of sampling are lower

- **Quota Sampling**
  - Population divided into mutually exclusive subgroups and non-random set of observations chosen from each subgroup

Quota Sampling



QuestionPro

Distribution of Sample Means

CEP2022_Notebook (1.6)

- We say that the sample mean ($\bar{y}$) gives us *an estimate* of the population mean μ
- But, since the sample is only a small subset of the entire population, *$\bar{y}$ is an uncertain estimate of μ*

- What if, we sample the population *several times*, each time calculating the sample mean $\bar{y}$
- Let's say we do it 'k' times, we get sample means as $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \ldots.., \bar{y}_k$
- How would these sample means behave?
  - How close are they from *μ?*
  - What's the *mean of sample means*?
  - What's the *standard deviation of sample means*?
  - More importantly, *what's their frequency distribution*?