D1EAD – Análise Estatística para Ciência de Dados 2021.1

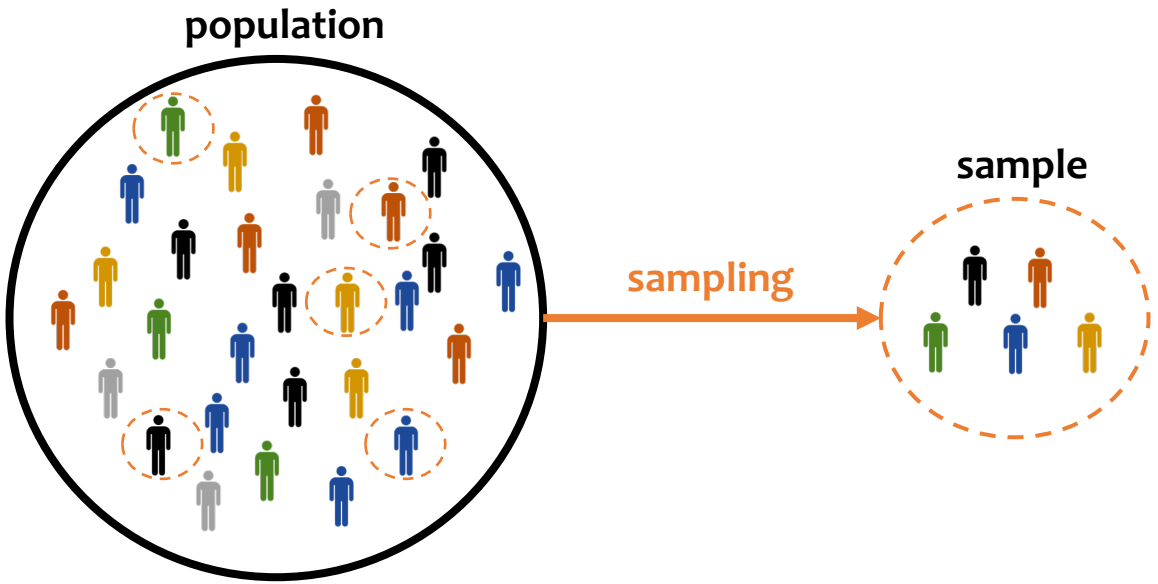
Data and Sampling Distributions

Prof. Ricardo Sovat
sovat@ifsp.edu.br
Prof. Samuel Martins (Samuka)
samuel.martins@ifsp.edu.br

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas

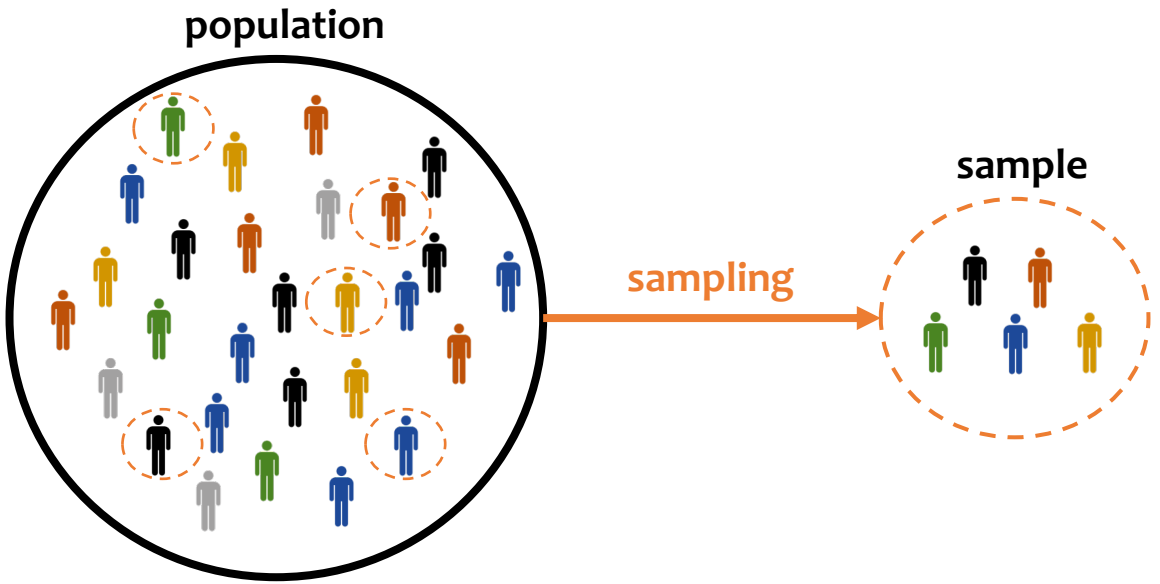
Population vs Sample

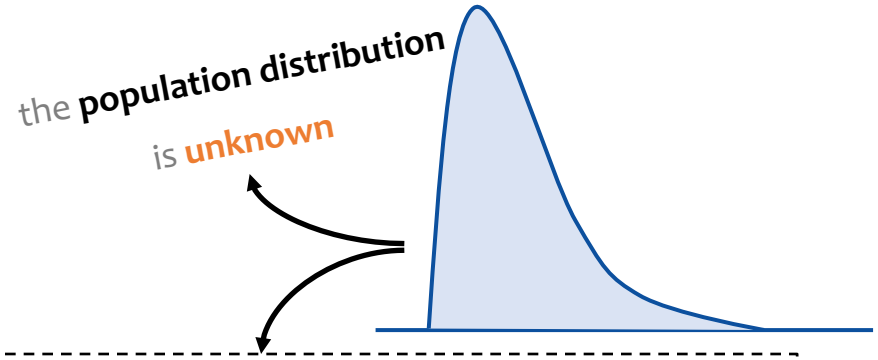
Even in the era of **big data**, **sampling** keeps being important and relevant



Population vs Sample

Even in the era of **big data**, **sampling** keeps being important and relevant

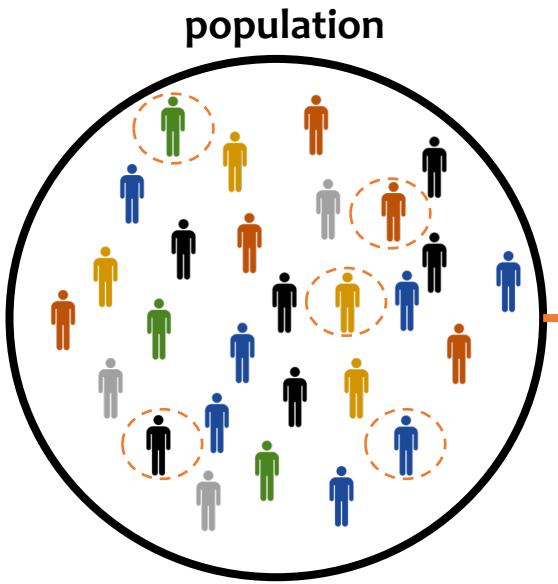


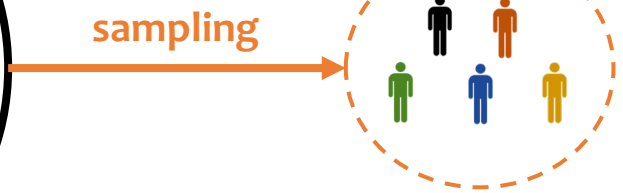


Frequentist stats focus on **asymptotic results**, i.e., as the **sample size** increases it should matches the **population distribution.**

Population vs Sample

Even in the era of **big data**, **sampling** keeps being important and relevant

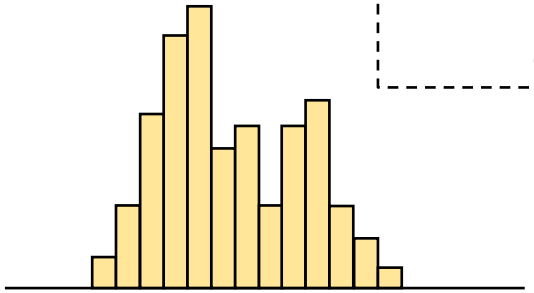




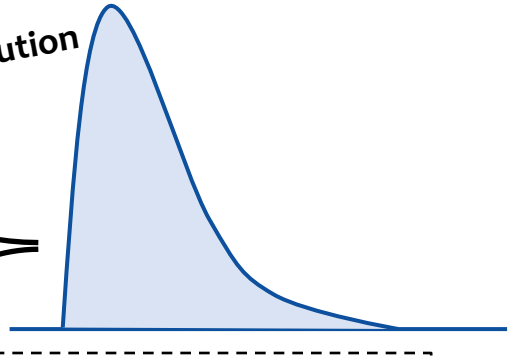
assumptions

are not needed

Modern methods relax this **assumption** and worry about the available **data sample.**

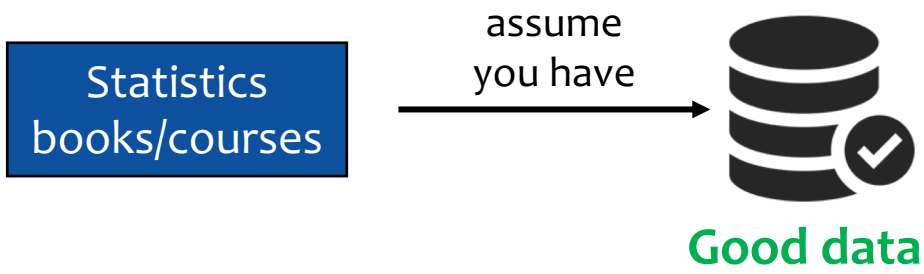


the **population distribution** is **unknown**



Frequentist stats focus on **asymptotic results**, i.e., as the **sample size** increases it should matches the **population distribution.**

The Importance of Data



The Importance of Data

Statistics books/courses

assume you have

Good data

it's like

assume you have

good ingredients

The Importance of Data

Statistics books/courses

assume you have

Good data

it's like

assume you have

good ingredients

BUT

finest

+

spoiled ingredients

The Importance of Data

- Noise
- Missing values
- Bias
- Unrepresentative data

flawed data

+

many fancy analysis

flawed results/conclusions

The Importance of Data

- Noise
- Missing values
- Bias
- Unrepresentative data

flawed data

+

many fancy analysis

flawed results/conclusions

Garbage In → **Garbage Out**

The Importance of Data

- Noise
- Missing values
- Bias
- Unrepresentative data

flawed data

+

many fancy analysis

flawed results/conclusions

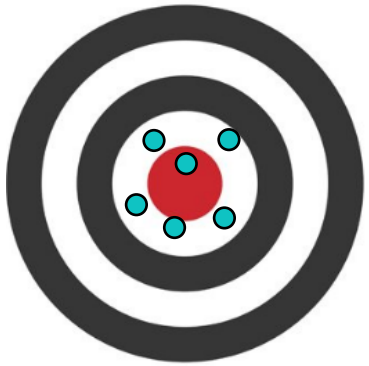
Garbage In → Garbage Out

Data quality often **matters more** than **data quantity** when making an estimate or a model based on a **sample**.

Precision vs Accuracy

✔ Precision

✘ Accuracy



✘ Precision

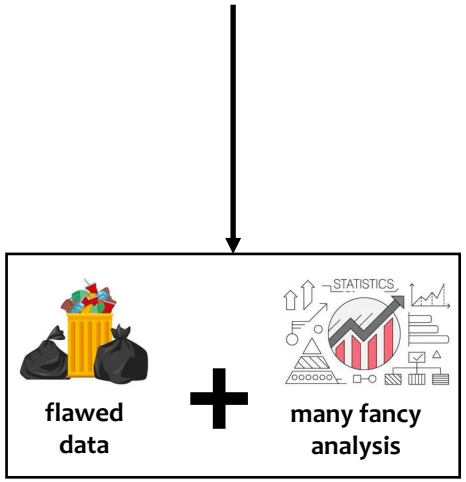
✔ Accuracy

✘ Precision

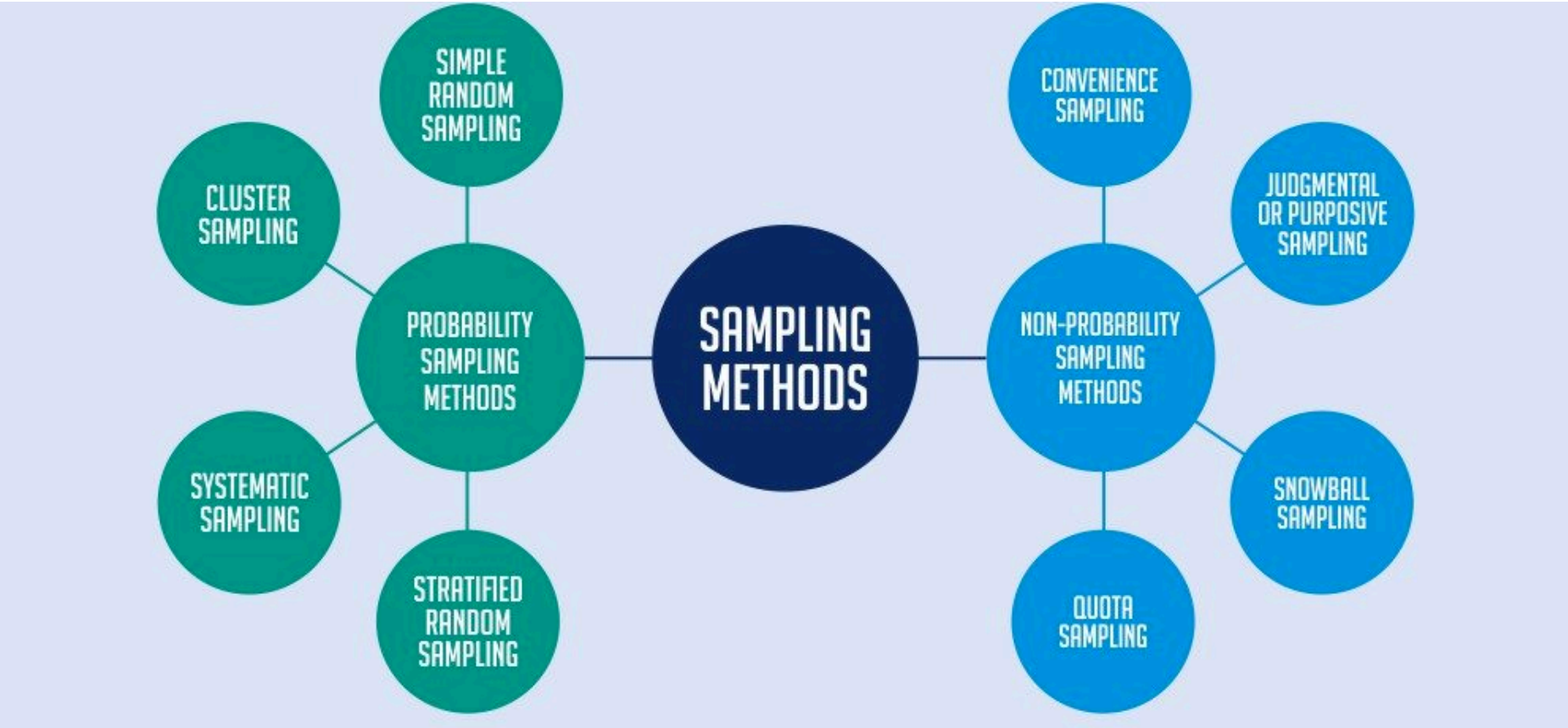
✘ Accuracy

✔ Precision

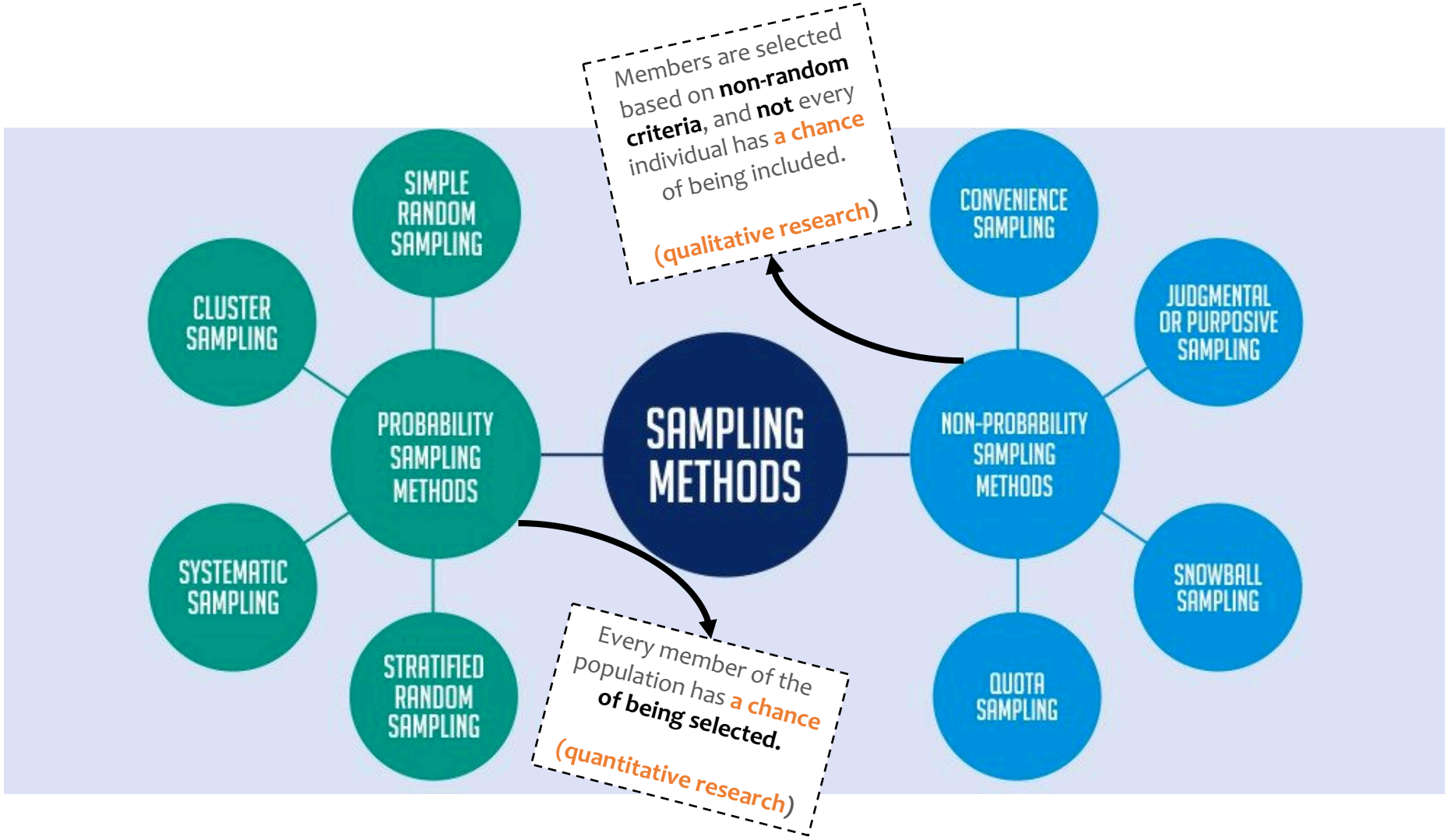
✔ Accuracy



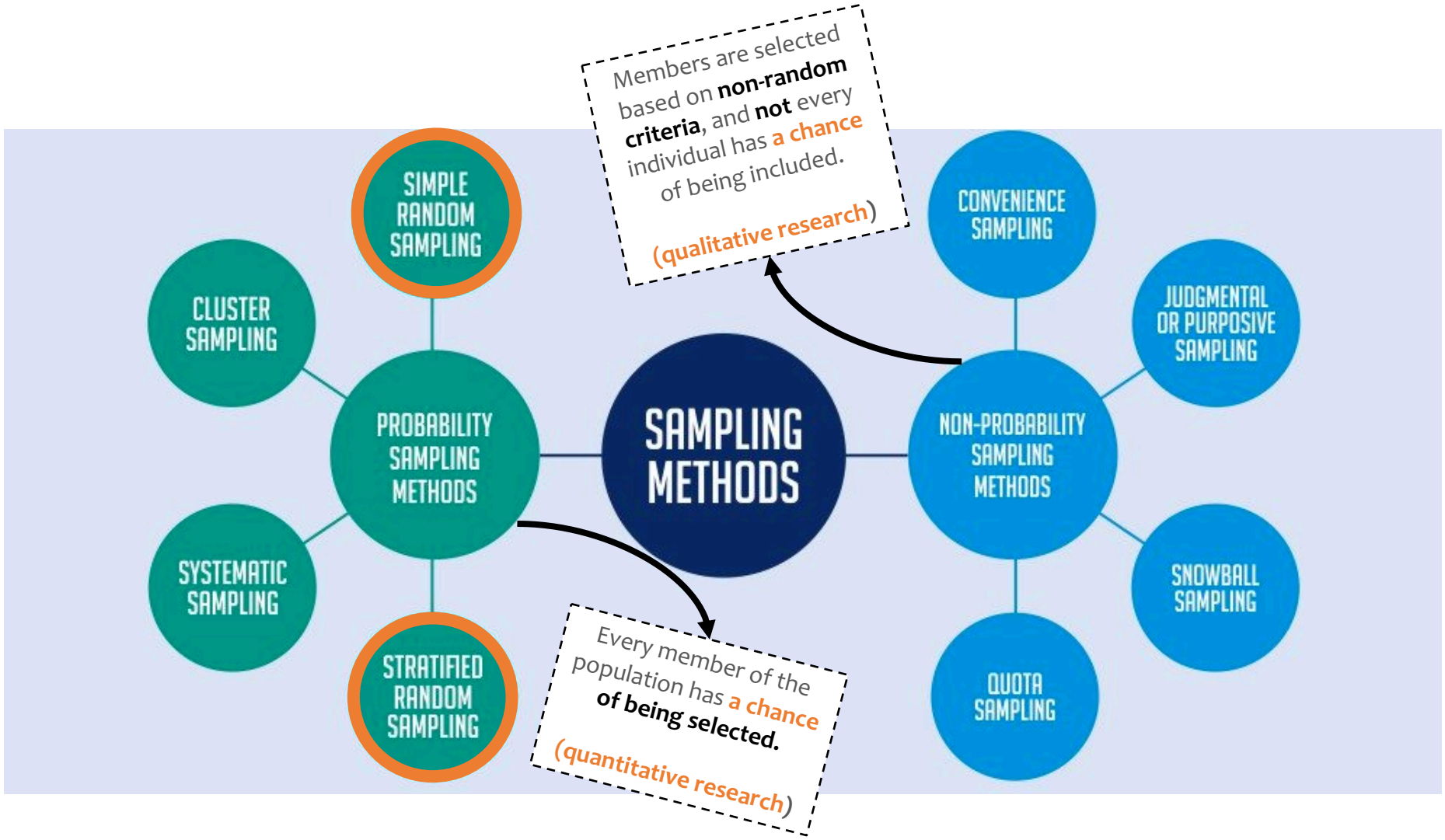
Types of Sampling



https://www.questionpro.com/blog/types-of-sampling-for-social-research/



https://www.questionpro.com/blog/types-of-sampling-for-social-research/

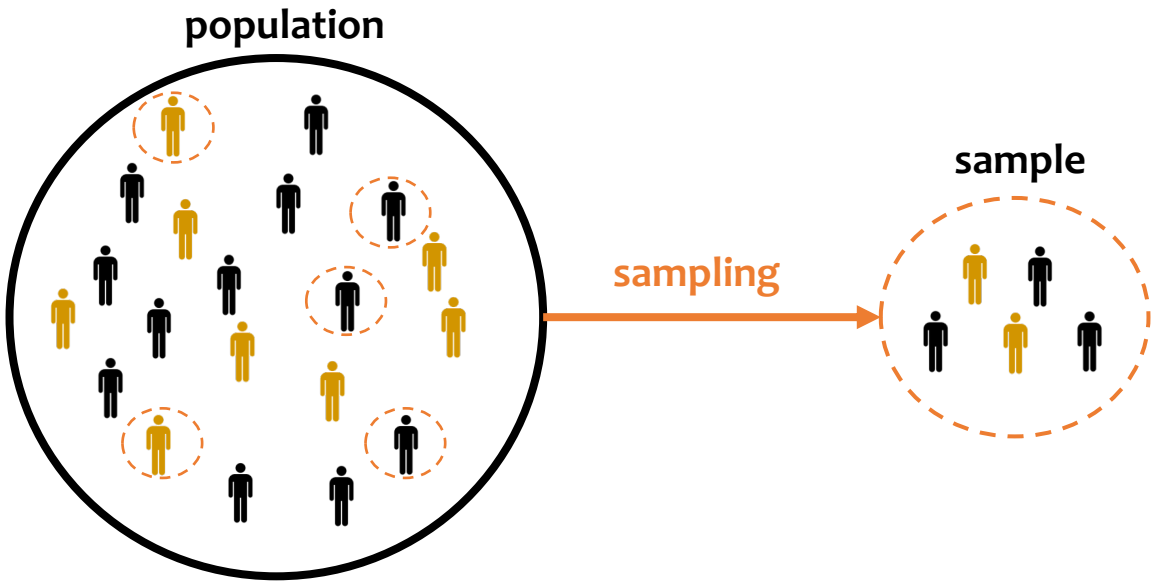


https://www.questionpro.com/blog/types-of-sampling-for-social-research/

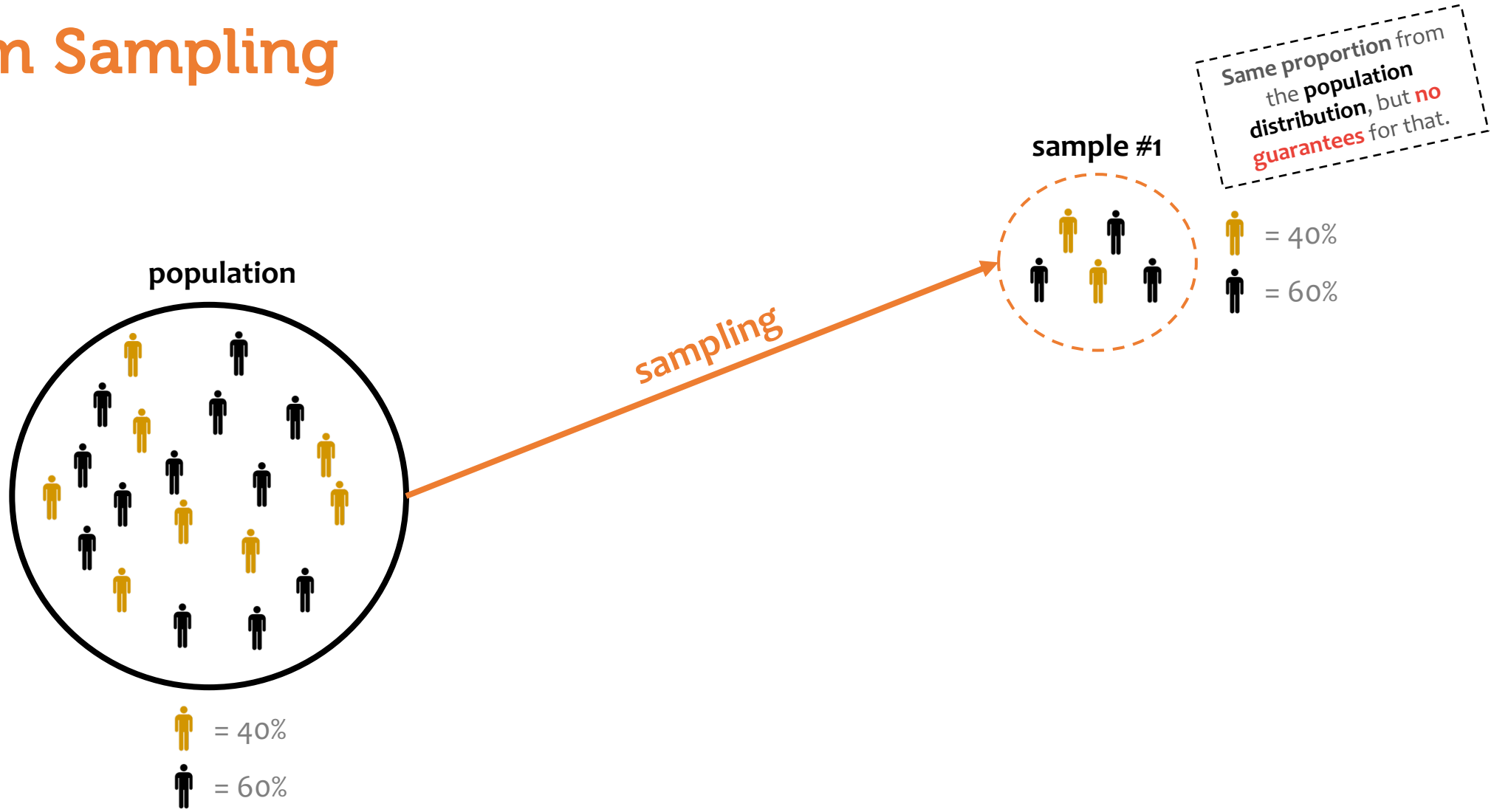
Random Sampling

The common way for **sampling** a **population** and to **avoid** **selection bias** (we'll se soon!).

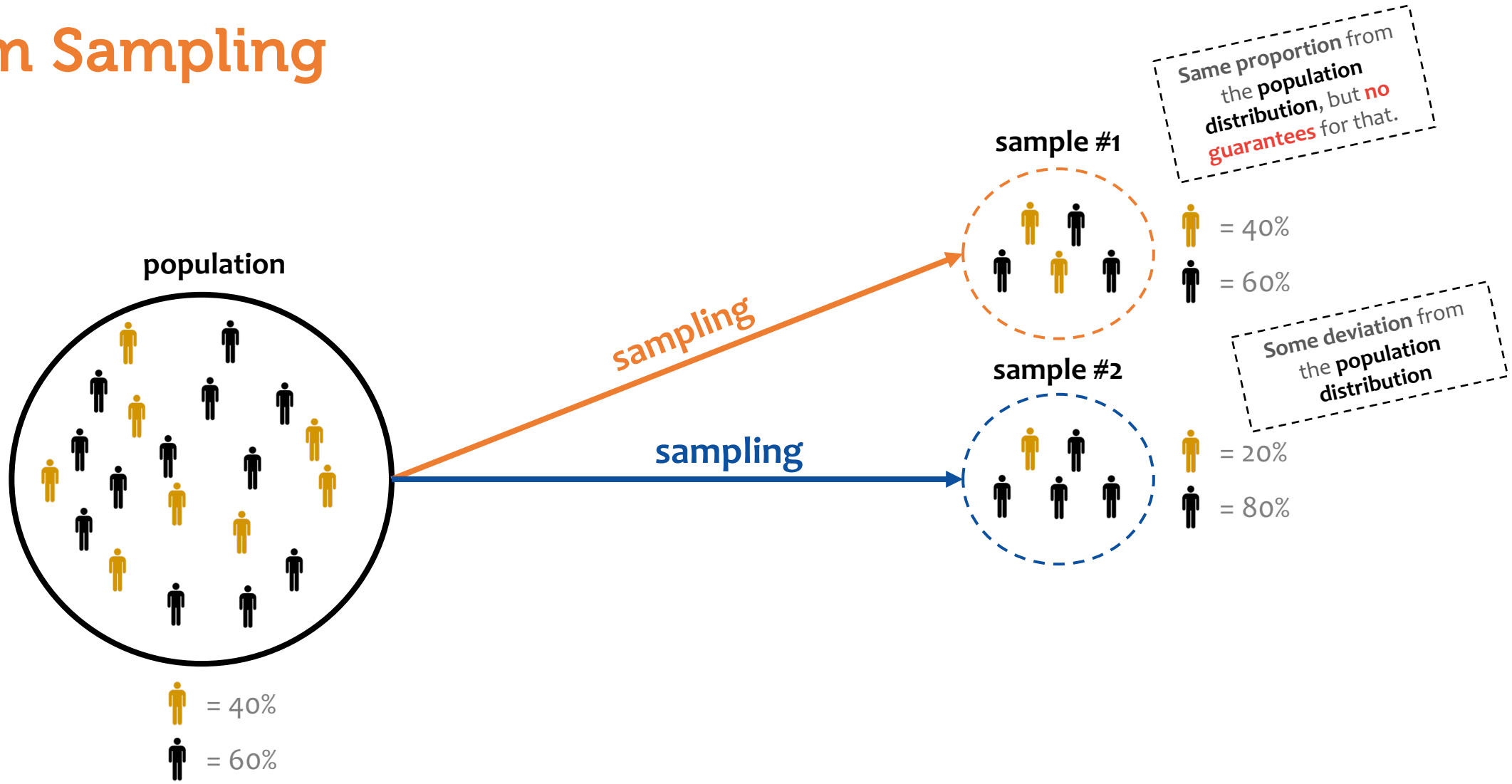
Each **member/observation** has **an equal chance** of being chosen for the **sample** at each draw.



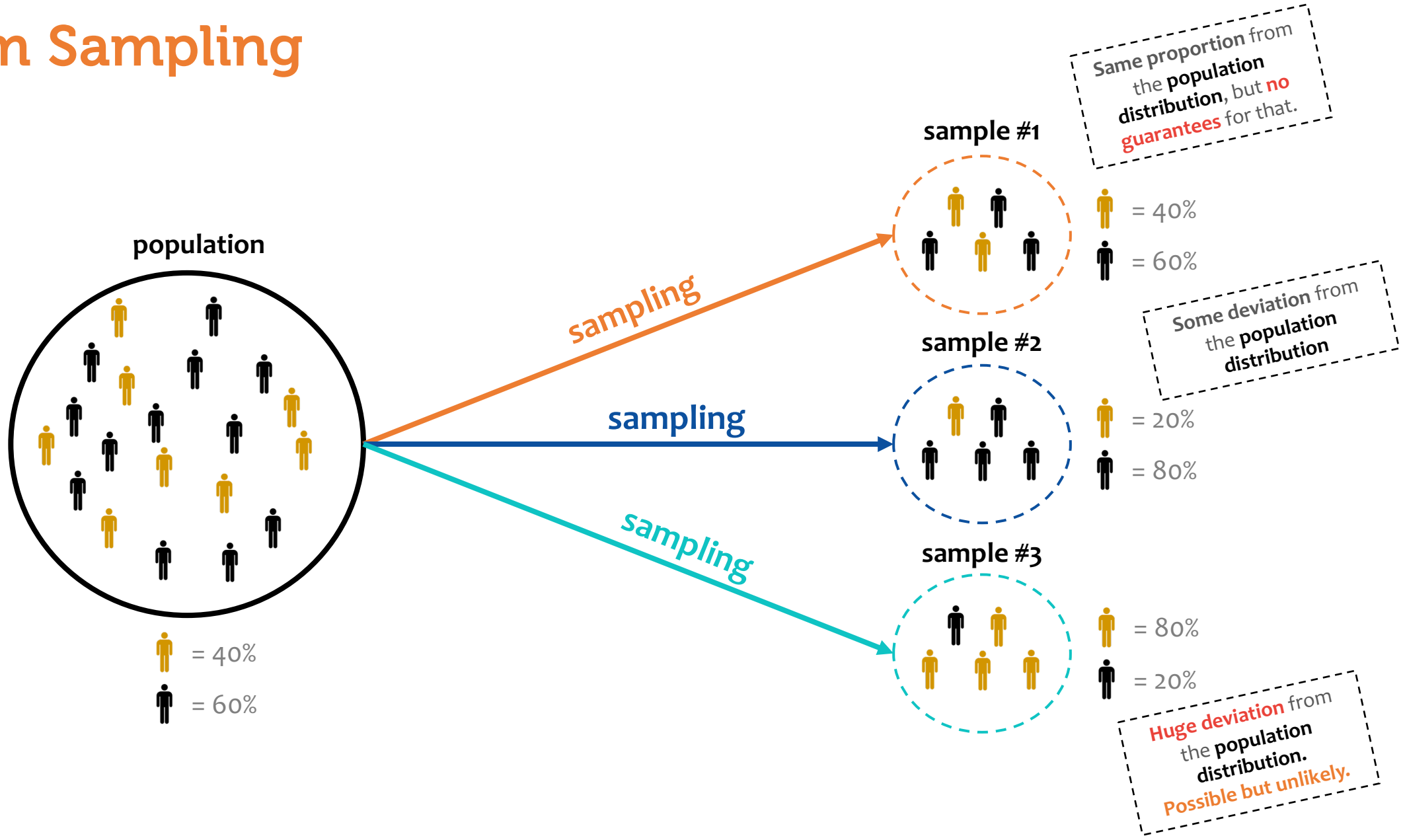
Random Sampling



Random Sampling



Random Sampling



Random Sampling – Pandas

App	Rating	Type
Duolingo: Learn Languages Free	4.7	Free
Z City	4.3	Free
EF Coach	4.8	Free
Dating Network	4.0	Free
Chess of Blades (BL/Yaoi Game) (No VA)	4.8	Paid
I am rich	3.8	Paid
Gangster Town	4.1	Free
Safest Call Blocker	4.4	Free
BJ's Bingo & Gaming Casino	4.5	Free
Chess School for Beginners	4.3	Free

Free: 10039 (92.62%)

Paid: 800 (7.38%)

Total: 10839 observations

random sampling

```
sample = population.sample(100)
```

App	Rating	Type
Badoo - Free Chat & Dating App	4.3	Free
What was I in my Past Life	3.7	Free
Diabetes & Diet Tracker	4.6	Paid
ez Share Android app	3.3	Free
IP address BW	NaN	Free
Carousell: Snap-Sell, Chat-Buy	4.3	Free
Simpli CT	NaN	Free
FH WiFiCam	2.6	Free
Muscle Premium - Human Anatomy, Kinesiology, B...	4.2	Paid
My Teacher - Classroom Play	4.0	Free

Free: 96 (96%)

Paid: 4 (4%)

Total: 100 observations

Dataset: Google Play Store Apps: https://www.kaggle.com/lava18/google-play-store-apps

Random Sampling – Pandas

App	Rating	Type
Duolingo: Learn Languages Free	4.7	Free
Z City	4.3	Free
EF Coach	4.8	Free
Dating Network	4.0	Free
Chess of Blades (BL/Yaoi Game) (No VA)	4.8	Paid
I am rich	3.8	Paid
Gangster Town	4.1	Free
Safest Call Blocker	4.4	Free
BJ's Bingo & Gaming Casino	4.5	Free
Chess School for Beginners	4.3	Free

Free: 10039 (92.62%)

Paid: 800 (7.38%)

Total: 10839 observations

random sampling

```
sample = population.sample(100)
```

different proportions

App	Rating	Type
Badoo - Free Chat & Dating App	4.3	Free
What was I in my Past Life	3.7	Free
Diabetes & Diet Tracker	4.6	Paid
ez Share Android app	3.3	Free
IP address BW	NaN	Free
Carousell: Snap-Sell, Chat-Buy	4.3	Free
Simpli CT	NaN	Free
FH WiFiCam	2.6	Free
Muscle Premium - Human Anatomy, Kinesiology, B...	4.2	Paid
My Teacher - Classroom Play	4.0	Free

Free: 96 (96%)

Paid: 4 (4%)

Total: 100 observations

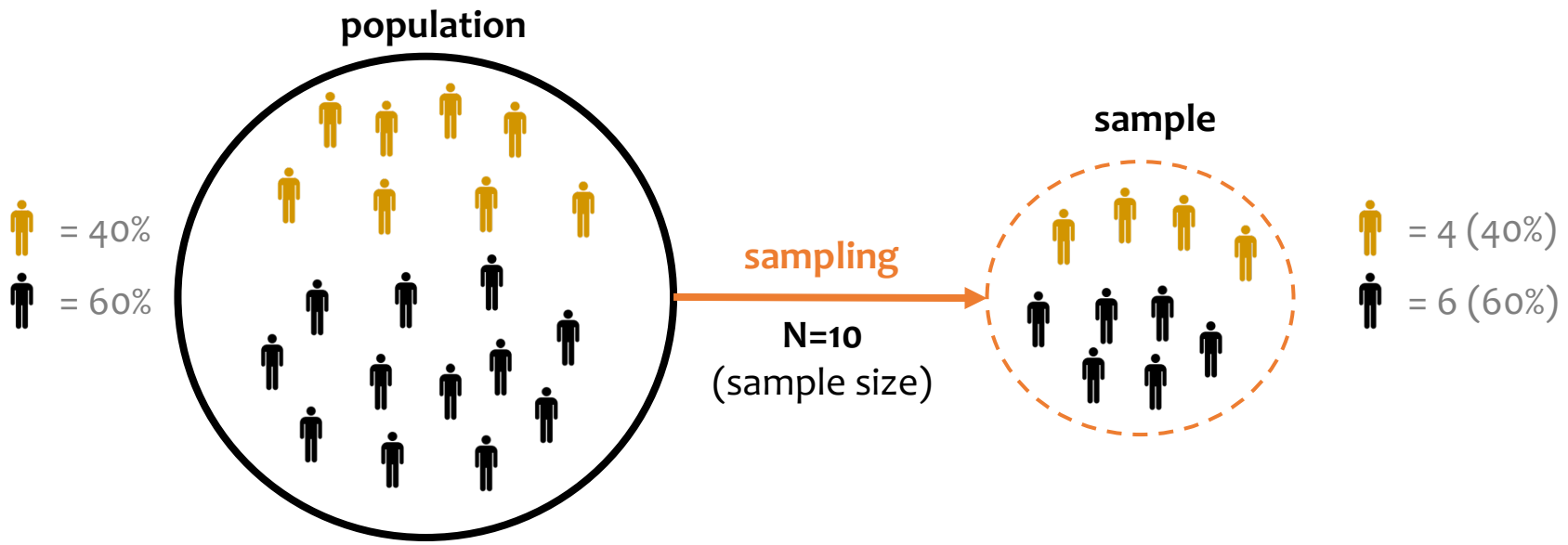
Dataset: Google Play Store Apps: https://www.kaggle.com/lava18/google-play-store-apps

Stratified Sampling

Divide the population into **subgroups (strata)** based on a **relevant characteristic** (e.g., gender, age range, healthy/unhealthy, …)

Perform **random sampling** in each **subgroup**, respecting the **overall population proportion**.

Aim at drawing **more precise conclusions** by ensuring that every **subgroup/class** is properly represented in the **sample**.



Stratified Sampling – Pandas

App	Rating	Type
Duolingo: Learn Languages Free	4.7	Free
Z City	4.3	Free
EF Coach	4.8	Free
Dating Network	4.0	Free
Chess of Blades (BL/Yaoi Game) (No VA)	4.8	Paid
I am rich	3.8	Paid
Gangster Town	4.1	Free
Safest Call Blocker	4.4	Free
BJ's Bingo & Gaming Casino	4.5	Free
Chess School for Beginners	4.3	Free

Free: 10039 (92.62%)

Paid: 800 (7.38%)

Total: 10839 observations

stratified sampling

```
sample = stratified_sampling(population,
                             sample_size=100)
```

N=100 (sample size)

App	Rating	Type
Badoo - Free Chat & Dating App	4.3	Free
What was I in my Past Life	3.7	Free
Diabetes & Diet Tracker	4.6	Paid
ez Share Android app	3.3	Free
IP address BW	NaN	Free
Carousell: Snap-Sell, Chat-Buy	4.3	Free
Simpli CT	NaN	Free
FH WiFiCam	2.6	Free
Muscle Premium - Human Anatomy, Kinesiology, B...	4.2	Paid
My Teacher - Classroom Play	4.0	Free

Free: 93 (93%)

Paid: 7 (7%)

Total: 100 observations

"equal" proportions

```
def stratified_sampling(population, sample_size):
    count = population.groupby('Type').size()
    proportion = count / population.shape[0]
    n_obs = round(proportion * sample_size).astype('int')
    sample = population.groupby('Type', group_keys=False)\
                       .apply(lambda group: group.sample(n_obs.loc[group.name]))

    return sample
```

```
Free    10039
Paid      800
```

```
Free    0.926192
Paid    0.073808
```

```
Free    93
Paid     7
```

Stratified Sampling – Pandas + Scikit-learn

App	Rating	Type
Duolingo: Learn Languages Free	4.7	Free
Z City	4.3	Free
EF Coach	4.8	Free
Dating Network	4.0	Free
Chess of Blades (BL/Yaoi Game) (No VA)	4.8	Paid
I am rich	3.8	Paid
Gangster Town	4.1	Free
Safest Call Blocker	4.4	Free
BJ's Bingo & Gaming Casino	4.5	Free
Chess School for Beginners	4.3	Free

Free: 10039 (92.62%)

Paid: 800 (7.38%)

Total: 10839 observations

stratified sampling

N=100 (sample size)

App	Rating	Type
Auto Dslr Photo Effect : Auto Focus Effect	3.8	Free
News360: Personalized News	4.4	Free
WiFi Monitor Pro - analyzer of Wi-Fi networks	4.6	Paid
Cut the Rope FULL FREE	4.4	Free
DW Security	5.0	Free
ChangeDA - Cours du DZD sur le marché parallèle	4.4	Free
Evie Launcher	4.7	Free
Counter Terrorist Attack	3.9	Free
Free Book Reader	3.4	Free
PAC-MAN Pop	4.3	Free

Free: 93 (93%)

Paid: 7 (7%)

Total: 100 observations

"equal" proportions

```
from sklearn.model_selection import train_test_split

sample, _ = train_test_split(population, train_size=100,
                             stratify=population['Type'])
```

sample size

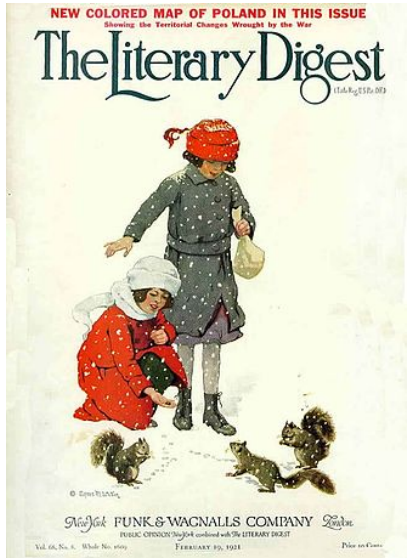
Bias

Bias

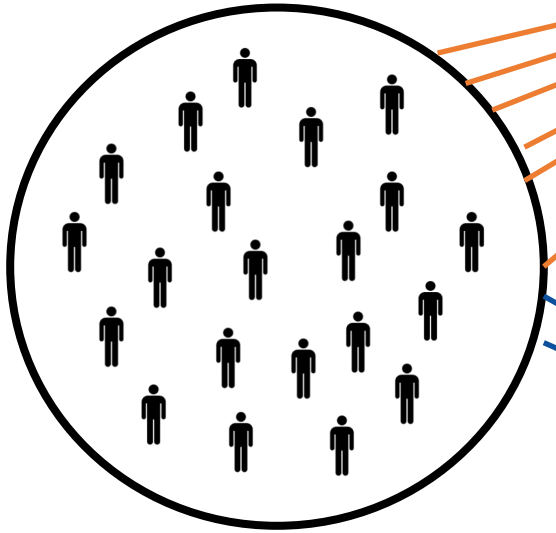
- **Measurement** or **sampling error** that are systemic and produced by the measurement or **sampling process**.
- **Tendency** of a statistic **overstimate** or **underestimate** a parameter.

Classical Example of Selection Bias

The Literay Digest (1936)

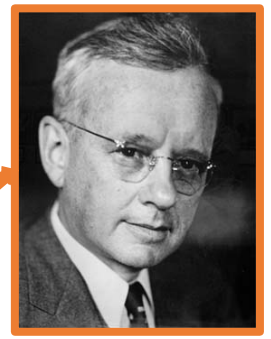


pooled its

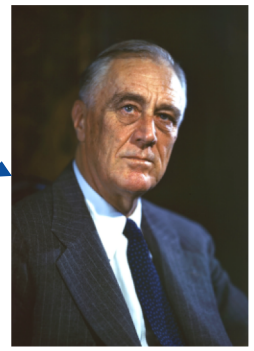


subscribers

with similar:
- profiles
- financial situation
- political tendencies



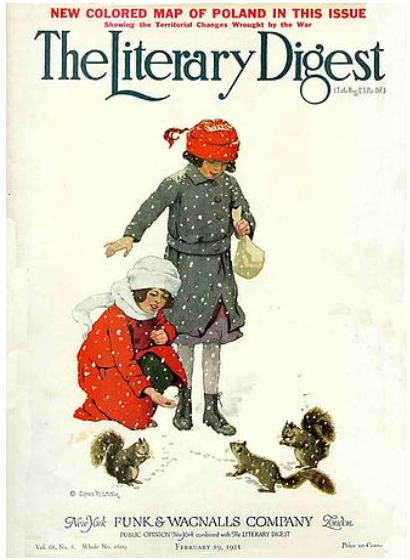
Alfred Landon



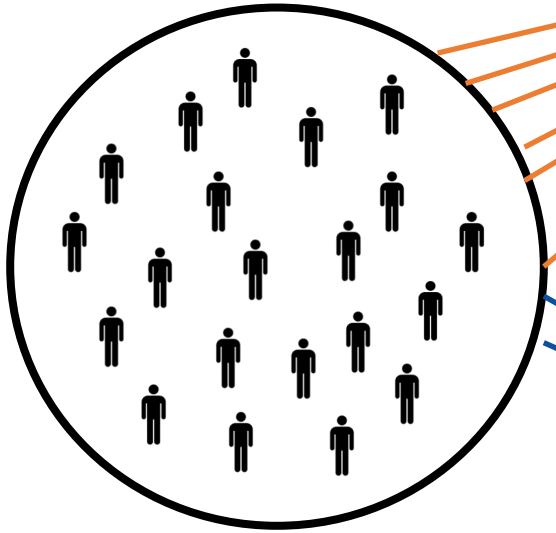
Franklin Roosevelt

Classical Example of Selection Bias

The Literay Digest (1936)

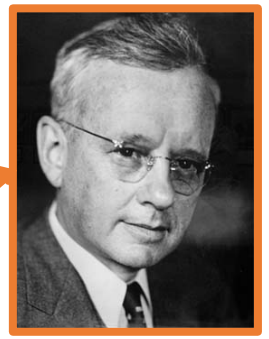


pooled its



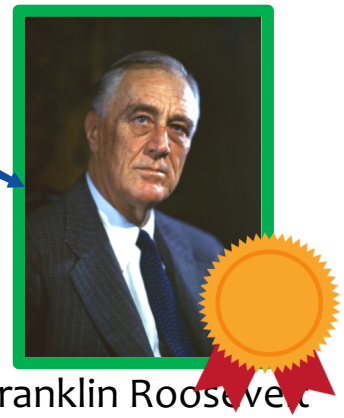
subscribers

with similar:
- profiles
- financial situation
- political tendencies



Alfred Landon

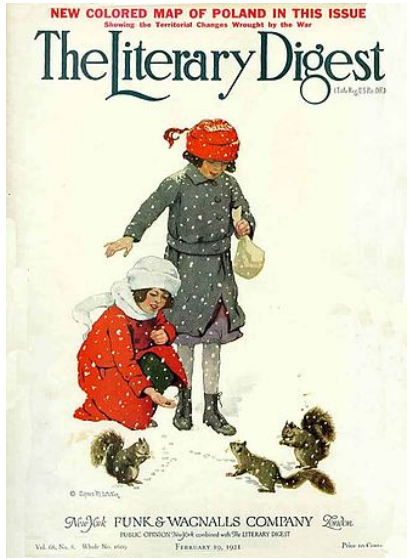
Winner



Franklin Roose[illegible]

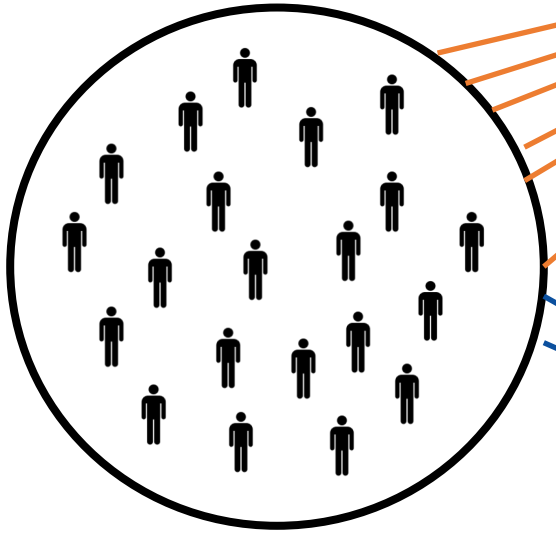
Classical Example of Selection Bias

The Literay Digest (1936)



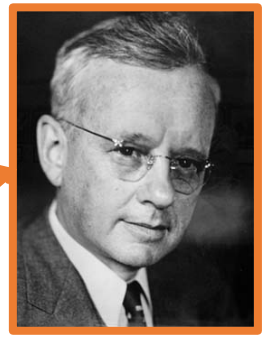
Selection Bias

pooled its



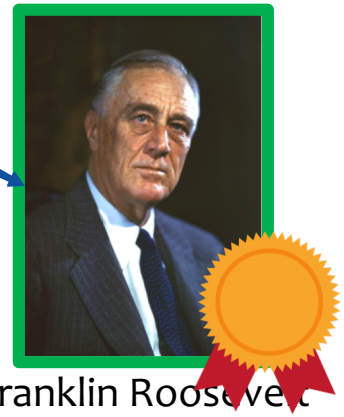
subscribers

with similar:
- profiles
- financial situation
- political tendencies



Alfred Landon

Winner



Franklin Roosevelt

Types of Sampling Biases

- Selection bias
- Self-selection bias
- Publication bias
- Recall bias
- Survivorship bias
- Healthy user bias
- ...

Types of Sampling Biases

- **Selection bias**
- **Self-selection bias**
- Publication bias
- Recall bias
- Survivorship bias
- Healthy user bias
- …

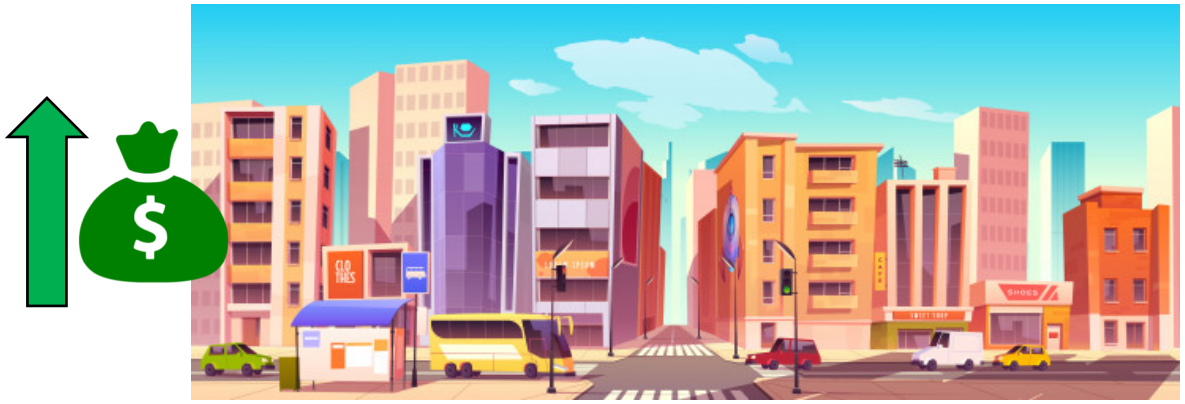
Selection Bias

Observations or **groups** in a study **differ** systematically from the **population** of interest, leading to **errors** in association or outcome.

Selection Bias

Observations or **groups** in a study **differ** systematically from the **population** of interest, leading to **errors** in association or outcome.

Example: Survey at a **specific neighborhood** of a city to **make conclusions** about the **city population**.



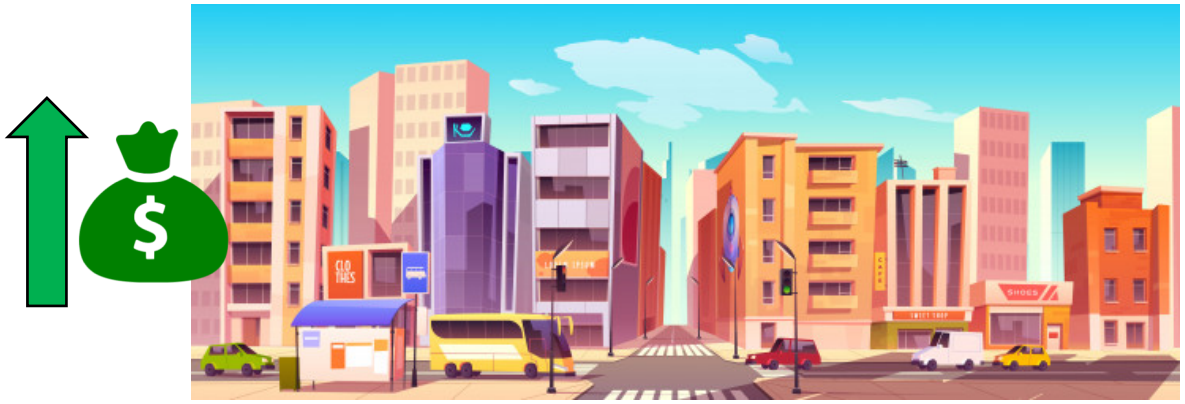
fancy neighborhood

poor neighborhood

Selection Bias

Observations or **groups** in a study **differ** systematically from the **population** of interest, leading to **errors** in association or outcome.

Example: Survey at a **specific neighborhood** of a city to **make conclusions** about the **city population**.



fancy neighborhood

poor neighborhood

Both surveys are likely to be **biased** by the fact that people have **different characteristics**.

We can use them to investigate their **specific people/neighborhood**.

Self-selection Bias

Individuals **not randomly selected and motivated** to be part of a sample.

Data Distribution

Motivation

- Some Machine Learning models are designed to work best under some **distribution** **assumptions**.
- Knowing with which **distributions** we are working with can help us to:
    - **identify** which machine-learning models are best to use.
    - Make analysis and inference easier during the **exploratory data analysis**.

- But, let's take a look at some **basic concepts** first.

Notations

fixed number from population

change from sample to sample

Population Parameter	Sample Statistic	Description
N	n	Number of elements.
μ	$\bar{x}$	Mean
σ	s	Standard deviation
ρ	r	Correlation coefficient.

Basic Concepts

Random experiment

- Process by which we observe something **uncertain**.
- Experiment, trial, or observation that **can be repeated** numerous times under the **same conditions**.
- Ex: toss a coin, roll a die, perc. of calls dropped due to errors over a particular time period, …

Outcome

- A result of a **random experiment**.
- The **outcome** of an individual random experiment **must not be affected by any previous outcome** and **cannot be predicted with certainty**.

Sample space

- The set of all possible outcomes of a random experiment.

Basic Concepts

Random experiment

- Process by which we observe something **uncertain**.
- Experiment, trial, or observation that **can be repeated** numerous times under the **same conditions**.
- Ex: toss a coin, roll a die, perc. of calls dropped due to errors over a particular time period, …

Outcome

- A result of a **random experiment**.
- The **outcome** of an individual random experiment **must not be affected by any previous outcome** and **cannot be predicted with certainty**.

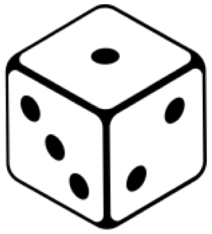
Sample space

- The set of all possible outcomes

Example 1:

Random experiment: roll a die

Sample space: S = {1, 2, 3, 4, 5, 6}.



Example 2:

Random experiment: toss a coin.

Sample space: S = {head, tail}.

Example 3:

Random experiment: number of iPhones sold in Brazil in 2020.

Sample space: S = {0, 1, 2, 3, … }.

Basic Concepts

Event

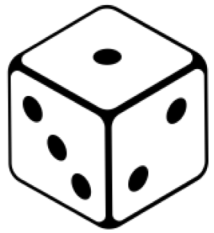
- An **outcome** or a **collection of outcomes** of a random experiment.
- **Any subset** of a sample space.

Ex:

Random experiment: roll a die

Sample space: S = {1, 2, 3, 4, 5, 6}.

Event: Getting an even number -> E = {2, 4, 6}.



Basic Concepts

Random variable

- Variable whose **values** depend on **outcomes** of a random phenomenon (e.g., random experiment).
- Think of it as a rule to decide what number you should record in your dataset after a real-world event happens.
- It can be **discrete** (takes countable number of distinct values) or **continuous** (the values between the range/interval and take infinite numbers).

Basic Concepts

Random variable

- Variable whose **values** depend on **outcomes** of a random phenomenon (e.g., random experiment).
- Think of it as a rule to decide what number you should record in your dataset after a real-world event happens.
- It can be **discrete** (takes countable number of distinct values) or **continuous** (the values between the range/interval and take infinite numbers).

Ex 1:

Random experiment: toss a coin.

Random variable: X = 0 (Head), 1 (Tail)

Basic Concepts

Random variable

- Variable whose **values** depend on **outcomes** of a random phenomenon (e.g., random experiment).
- Think of it as a rule to decide what number you should record in your dataset after a real-world event happens.
- It can be **discrete** (takes countable number of distinct values) or **continuous** (the values between the range/interval and take infinite numbers).

Ex 1:

Random experiment: toss a coin.

Random variable: X = 0 (Head), 1 (Tail)

Ex 2:

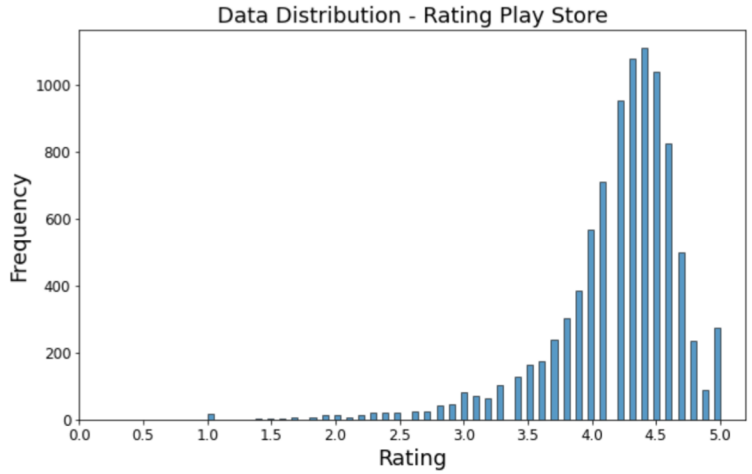
Random experiment: a soccer match.

Basic Concepts

Data Distribution

- Distribution of **individual data points** from a dataset.
- It is a function or a listing which shows **all the possible values (or intervals)** of the data.
- It also tells you how often each value occurs **(frequency)**.
- Often referred to as **probability distributions**.

Ex: Ratings Play Store

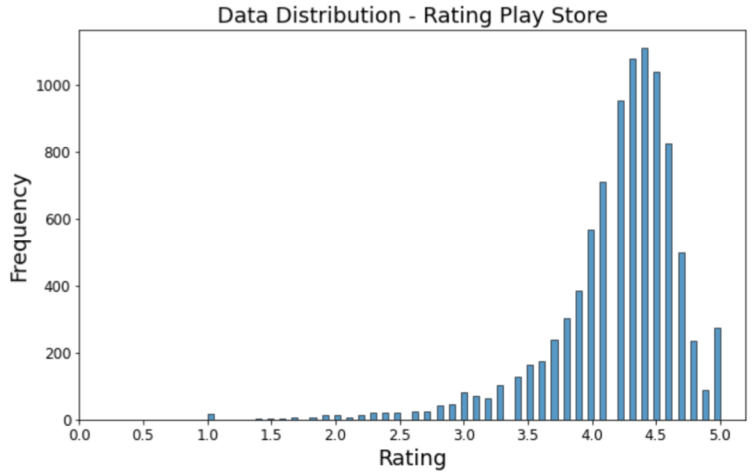


Basic Concepts

Data Distribution

- Distribution of **individual data points** from a dataset.
- It is a function or a listing which shows **all the possible values (or intervals)** of the data.
- It also tells you how often each value occurs **(frequency)**.
- Often referred to as **probability distributions**.

Ex: Ratings Play Store



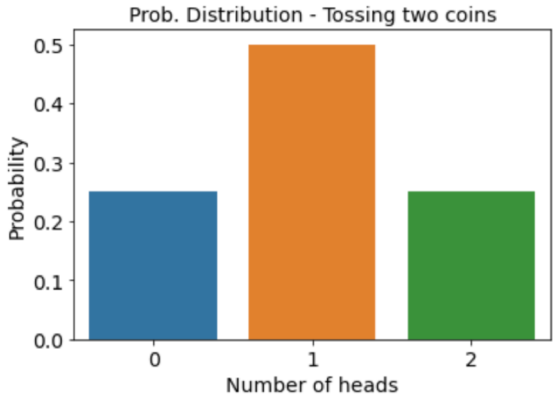
Probability Distribution

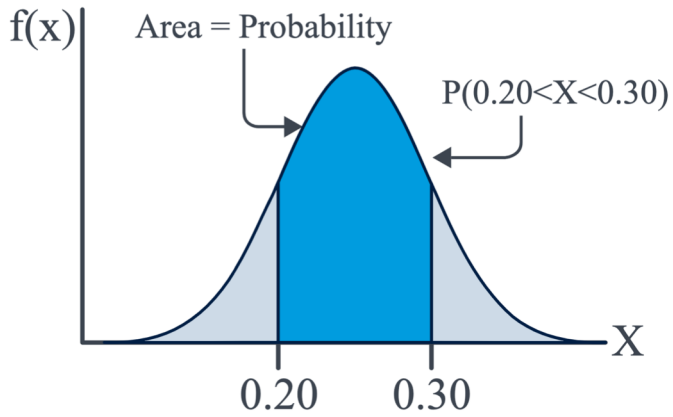
- Mathematical function that gives the **probabilities of occurrence** of different possible **outcomes** for an **experiment**.

Ex: Toss a coin twice

Sample Space: S = {HH, HT, TH, TT}

Event: Prob. of getting heads





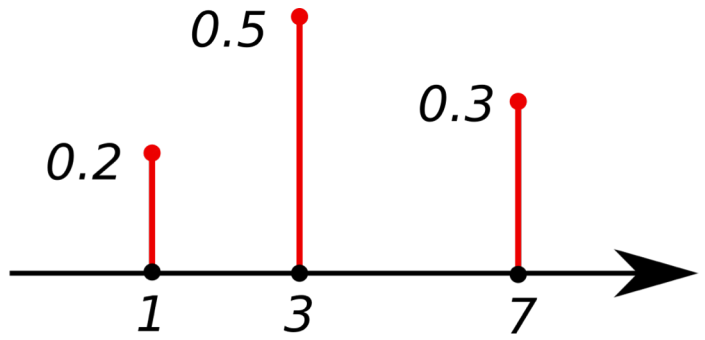
Basic Concepts

Probability Mass Function (PMF)

- The **probability distribution** of a **discrete random variable.**

Properties:

- $\boldsymbol{P(X = x) = f(x)}$
  - Prob. of the random variable X at a **specific x**
- All probabilities are positive: $\boldsymbol{P(x) \geq 0}$
- Any event in the distribution has: $\boldsymbol{0 \leq P(x) \leq 1}$
- The sum of all probabilities is 1. So $\boldsymbol{\sum P(x) = 1}$



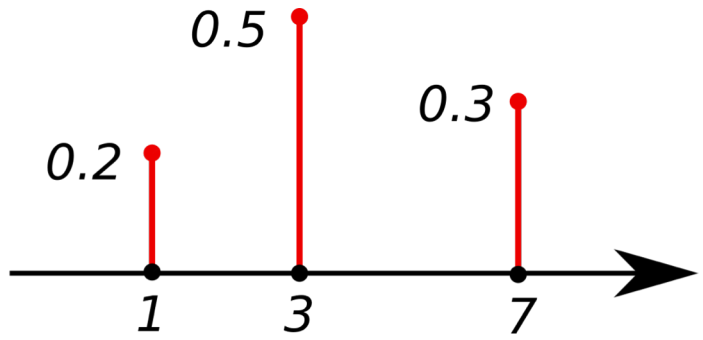
Basic Concepts

Probability Mass Function (PMF)

- The **probability distribution** of a **discrete random variable.**

Properties:

- $\boldsymbol{P(X = x) = f(x)}$
  - Prob. of the random variable X at a **specific x**
- All probabilities are positive: $\boldsymbol{P(x) \geq 0}$
- Any event in the distribution has: $0 \leq \boldsymbol{P(x) \leq 1}$
- The sum of all probabilities is 1. So $\boldsymbol{\sum P(x) = 1}$

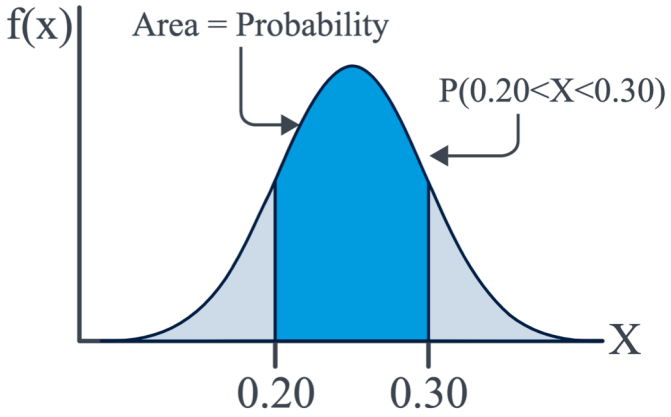


Probability Density Function (PDF)

- The **probability distribution** of a **continuous random variable.**

Properties:

- $\boldsymbol{P(X = x) = 0}$ (it is always zero)
- $\boldsymbol{P(a \leq X \leq b) = \int_a^b f(x)}$
- $\boldsymbol{f(x) \geq 0}$, for all $x \in \mathbb{R}$
- $\boldsymbol{\int_{-\infty}^{+\infty} f(x) = 1}$

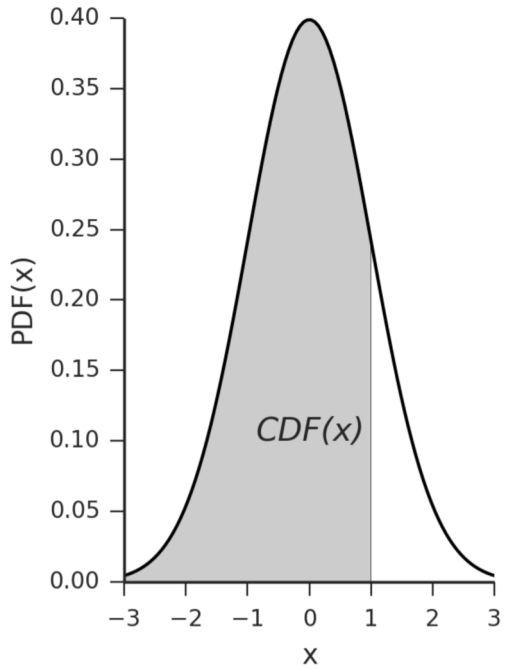


Basic Concepts

Cumulative Distribution Function (CDF)

- Gives the **cumulative value** from $-\infty$ up to a value **x** for a **random variable X (discrete or continuous)**
- It is the **probability function** that X will take a value **less than or equal to x**.

$$\boldsymbol{P(X \leq x)}, \text{ for all } x \in \mathbb{R}$$



$$\boldsymbol{P(X \leq 1)}$$

Basic Concepts

Expected Value

- A practical approach results in a **data/frequency distribution** and a **mean value**
- A theoretical approach results in a **probability distribution** and an **expected value**.

$$\boldsymbol{E(X) = \sum_{x \in S} xP(X = x)}$$

S is the **sample space**

Ex:

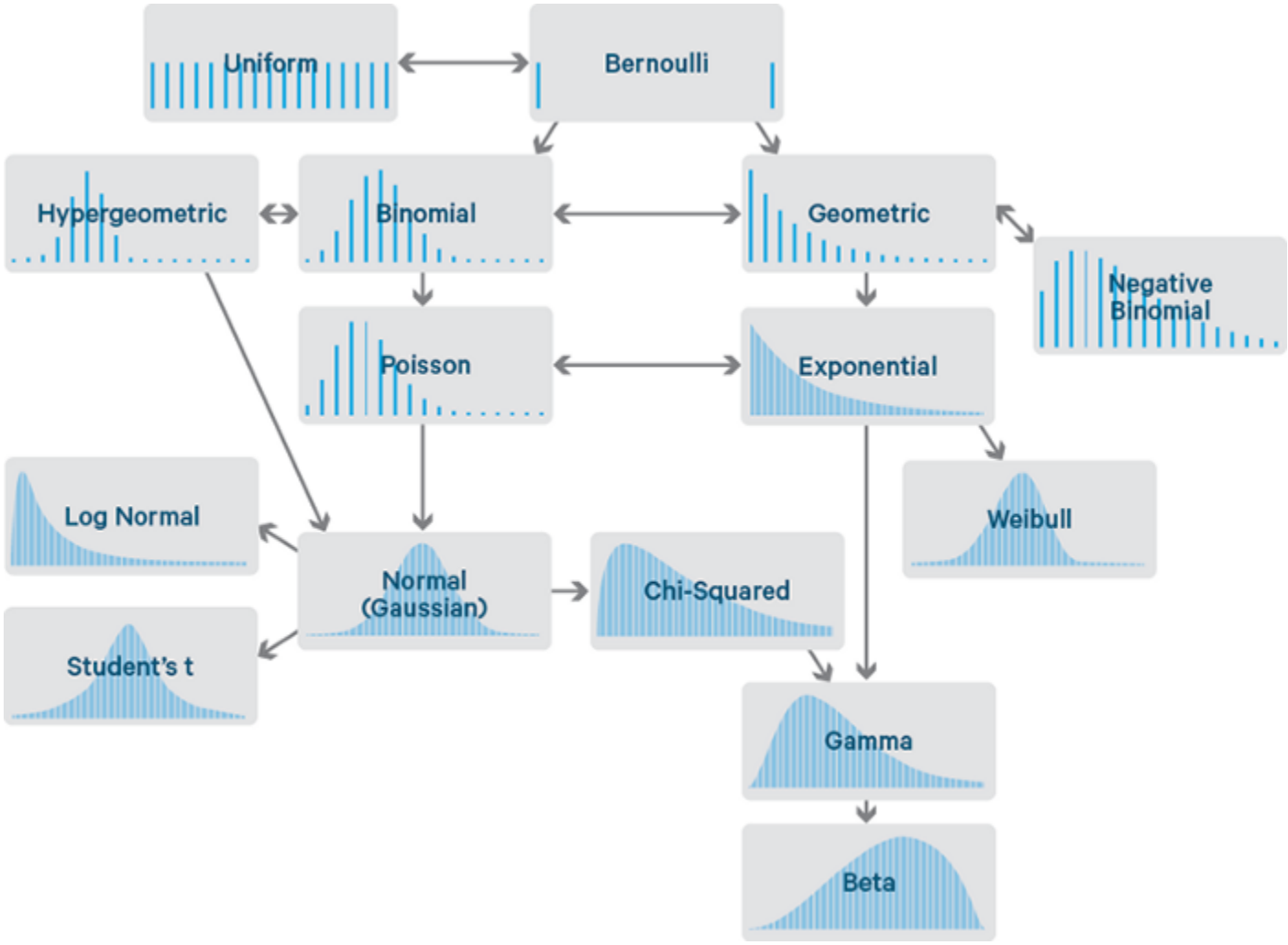
Suppose a **discrete random variable X** with the following sample space and PMF:

$$X = \begin{cases} 1\ with\ probability\ 1/8 \\ 2\ withprobability\ 3/8 \\ 3\ withprobability\ 1/2 \end{cases}$$

The **expected value** for X is:

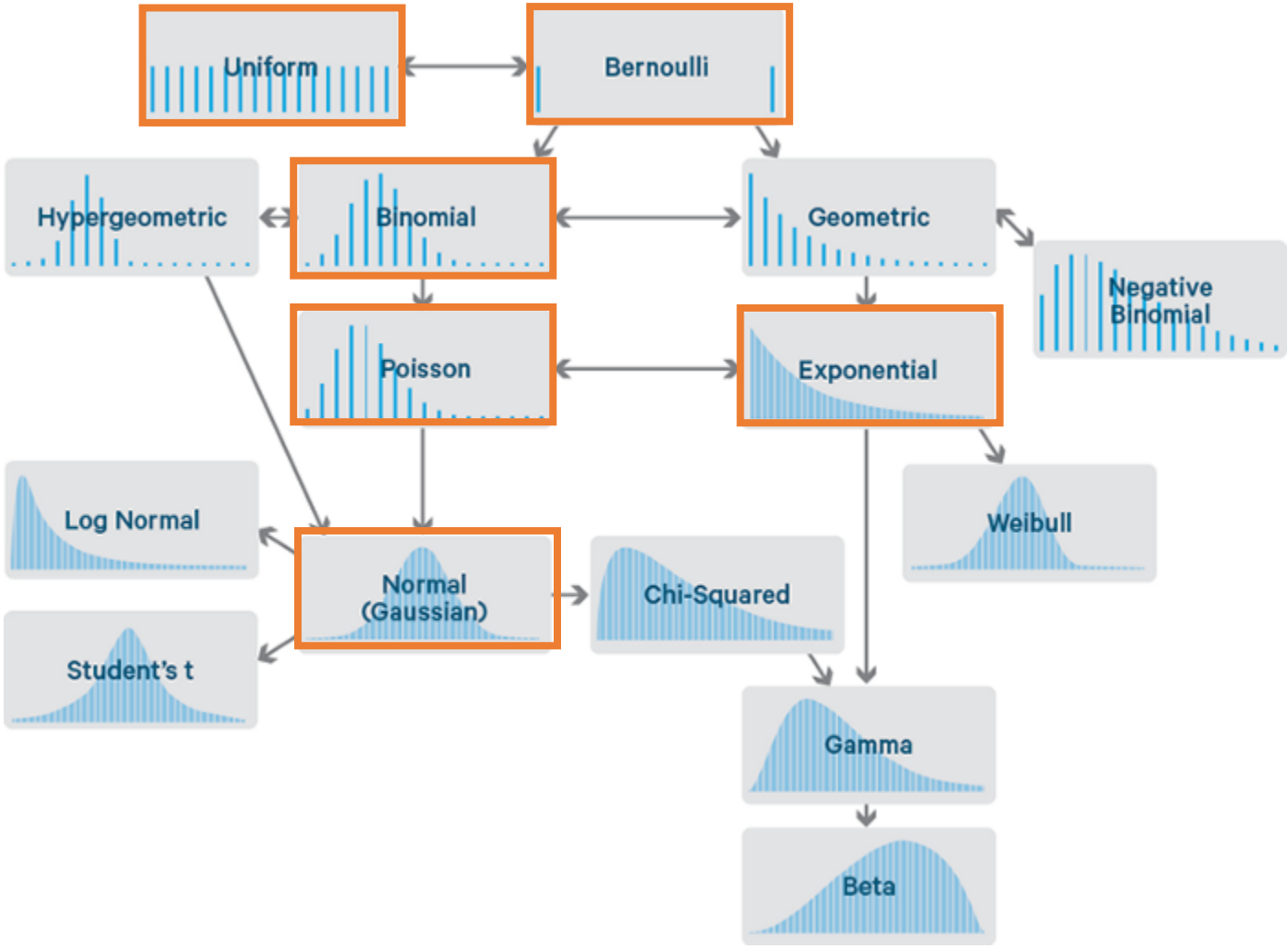
$$\boldsymbol{E(X) = 1 \cdot \left(\frac{1}{8}\right) + 2 \cdot \left(\frac{3}{8}\right) + 3 \cdot \left(\frac{1}{2}\right) = 2.375}$$

Probability Distributions



https://medium.com/diogo-menezes-borges/introduction-to-statistics-for-data-science-7bf596237ac6

Probability Distributions



https://medium.com/diogo-menezes-borges/introduction-to-statistics-for-data-science-7bf596237ac6

1. Bernoulli Distribution

- The simplest distribution.
- Only **two possible outcomes**:
  - 1 (success)
  - 0 (failure)
- A **single trial**.

$$P(X = x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

$$\mu = np$$

$$\sigma = \sqrt{pq}$$

p: probability of success.

1. Bernoulli Distribution

$$P(X = x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

$$\mu = np$$

$$\sigma = \sqrt{pq}$$

p: probability of success.

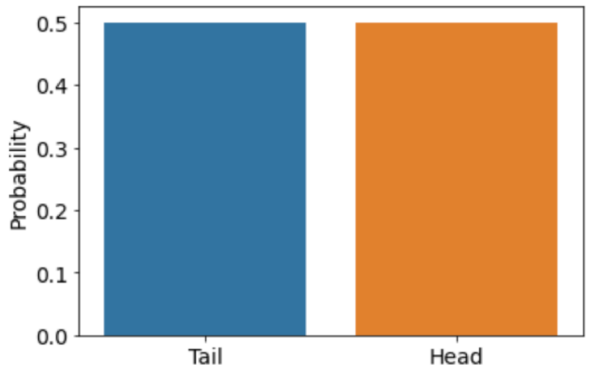
Ex 1: Tossing a coin.

X = {1 (head), 0 (tail)}

$$P(X = x) = \begin{cases} 0.5, & x = 0 \\ 0.5, & x = 1 \end{cases}$$

$$\mu = 0.5$$

$$\sigma = 0.25$$



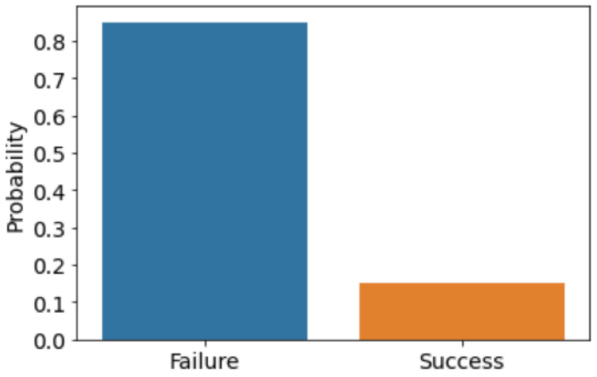
Ex 2: Samuka scoring a goal.

X = {1 (success), 0 (failue)}

$$P(X = x) = \begin{cases} 0.85, & x = 0 \\ 0.15, & x = 1 \end{cases}$$

$$\mu = 0.15$$

$$\sigma = 0.1275$$



2. Binomial Distribution

- It is the **frequency distribution** of the **number of successes (X)** in a given **number of trials (n)** with specified **probability (p) of success** in each trial.
- Ex: getting two heads when tossing three coins, nº of defective PCs in a shipment, nº of girls in a family, etc.

Binomial experiment

1. Fixed number of **identical trials**;
2. Trials are **independent** of each other;
3. Only **two** outcomes are possible (e.g., success and failure, head and tail, true and false, etc);
4. Fixed probability of **success**: **p** (consequently, the probability of **failure** is **q = 1 - p**)

2. Binomial Distribution

$$C_x^n \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

n: number of trials

p: probability of success.

$q = (1 - p)$: probability of failure.

Exercise 1

In an admission test for the Data Science specialization, **10 questions** with **3 possible choices** in each question.

Each question scores equally. Suppose that a candidate have not been prepared for the test. She decided to guess all answers.

Let the test has the **maximum score of 10** and **cut-off score of 5** for being approved for the next stage.

Provide the probability that this candidate will **get 5 questions right**, and the probability that she will **advance to the next stage of the test**.

Exercise 2

In the last World Chess Championship, **the proportion of female participants was 60%.**

The total of teams, with 12 members, in this year's championship is 30.

According to these information, **how many teams should be formed by 8 women?**

3. Poisson Distribution

- Used to describe the **number of occurrences** within a **specific period of time** or **space**.
- Some more examples are:
    - The number of emergency calls recorded at a hospital in a day.
    - The number of thefts reported in an area on a day.
    - The number of customers arriving at a salon in an hour.
    - The number of suicides reported in a particular city.
    - The number of printing errors at each page of the book.

Poisson experiment

A distribution is called **Poisson distribution** when the following assumptions are valid:

1. The **probability of success** is the sample over at the whole interval.
2. Any **successful event should not influence** the outcome of another successful event.
    - The nº of occurrences at a given interval is **independent** from the nº of occurrences at other intervals.
3. The **probability of success** (a given occurrence) is the same at intervals with **equal length**.
4. The **probability of success** in an interval **approaches zero** as the **interval becomes smaller**.

3. Poisson Distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, 2, \ldots$$
$$\mu = \lambda$$
$$\sigma = \sqrt{\lambda}$$

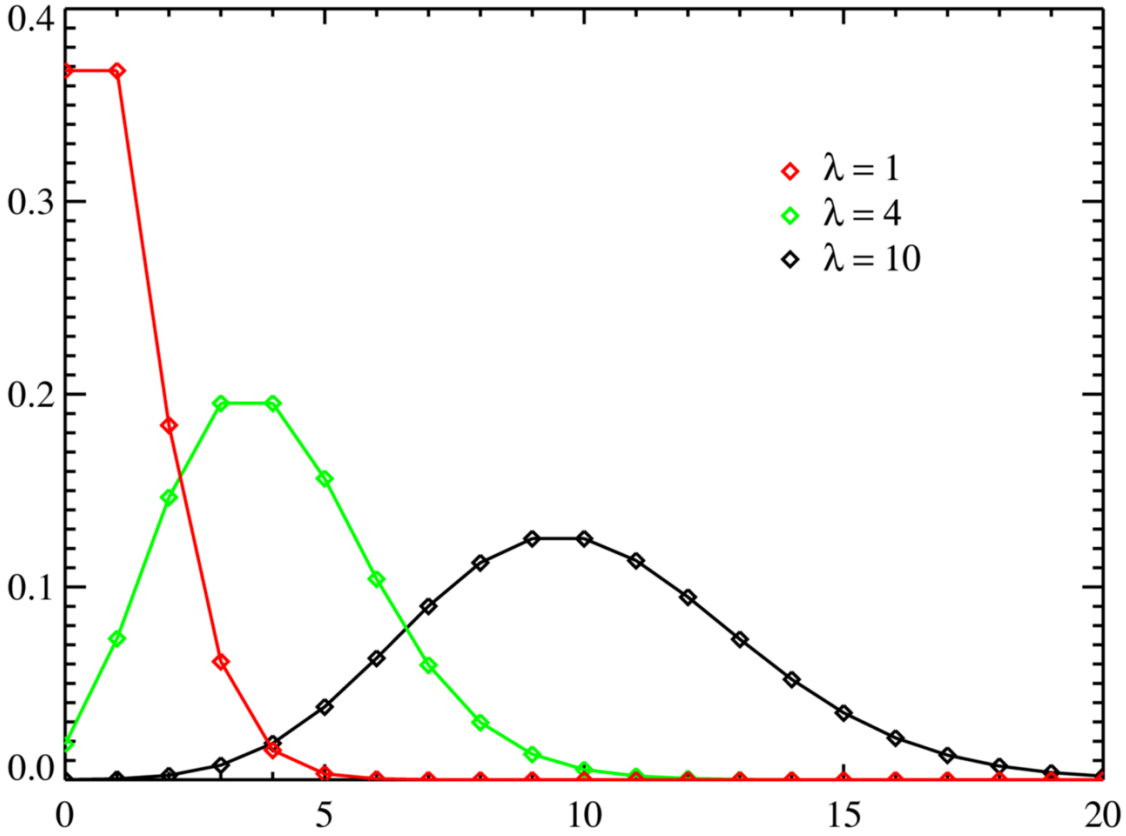
e: 2.71828…

λ: mean nº of occurrences/events (frequency) within a period of time.

x: nº of successes within the period of time.

3. Poisson Distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0, 1, 2, \ldots$$
$$\mu = \lambda$$
$$\sigma = \sqrt{\lambda}$$



f occurrences/events (frequency) within a period of time.
esses within the period of time.

Exercise 1

A restaurant receives **20 orders per hour**. What is the chance that, at a given hour chosen at random, the restaurant will receive **15 orders**?

Exercise 2

Vehicles pass through a junction on a busy road at an average rate of 300 per hour.

Find the probability that none passes in a given minute.

What is the expected number passing in two minutes?

Find the probability that this expected number actually pass through in a given two-minute period.

Exercise 3

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

Uniform Distribution

Normal Distribution

Central Limit Theorem

Confidence Interval

Bootstrapping

D1EAD – Análise Estatística para Ciência de Dados 2021.1

Data and Sampling Distributions

Prof. Ricardo Sovat
sovat@ifsp.edu.br
Prof. Samuel Martins (Samuka)
samuel.martins@ifsp.edu.br

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas