D1EAD – Análise Estatística para Ciência de Dados 2021.1

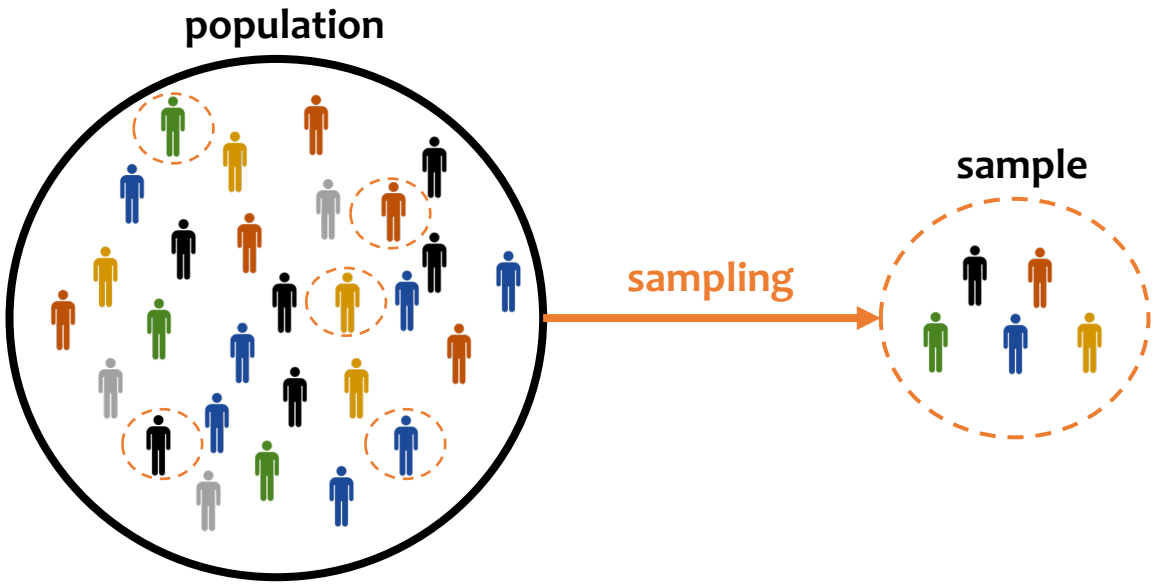
Data Distributions (Part 1)

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

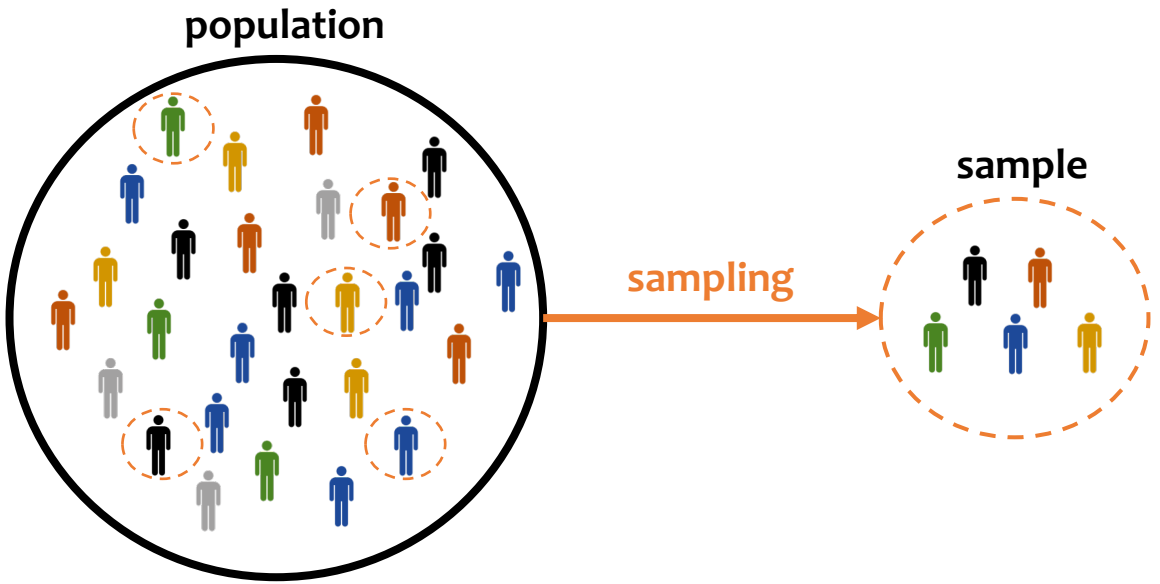
Population vs Sample

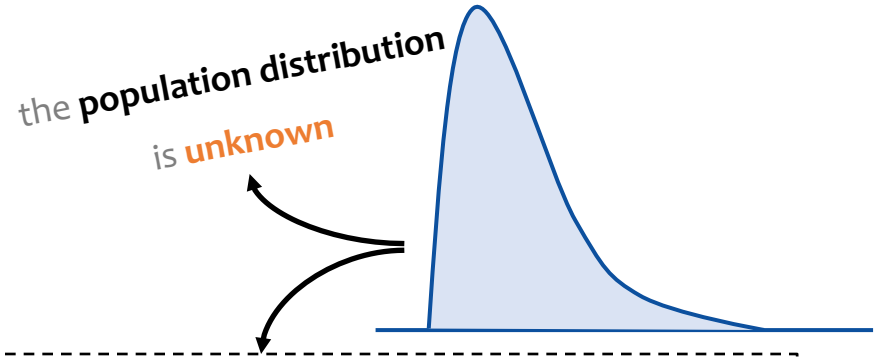
Even in the era of **big data**, **sampling** keeps being important and relevant



Population vs Sample

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

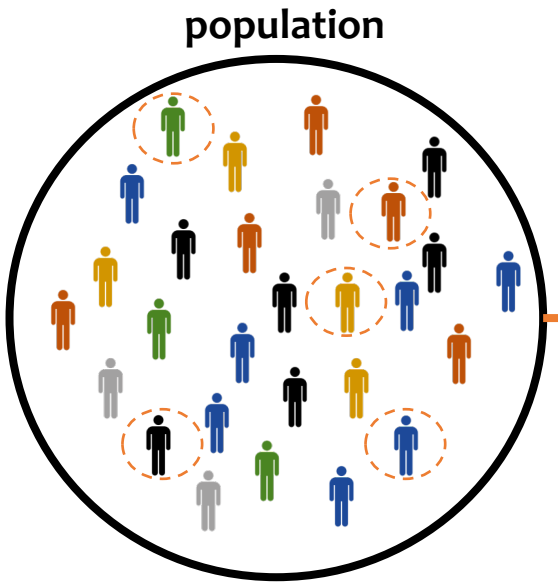


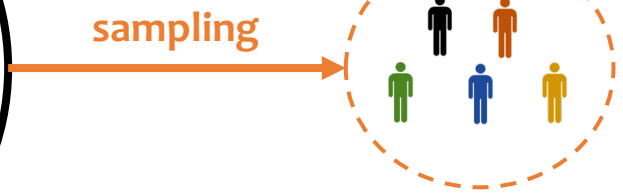


Frequentist stats focus on **asymptotic results**, i.e., as the **sample size** increases it should matches the **population distribution.**

Population vs Sample

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

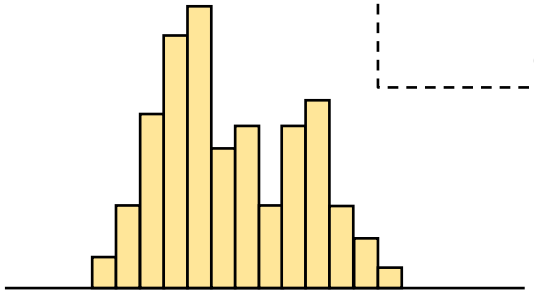




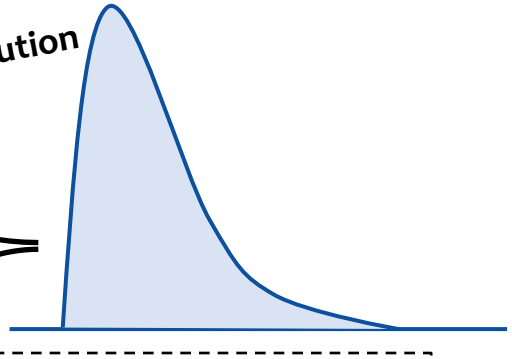
assumptions

are not needed

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

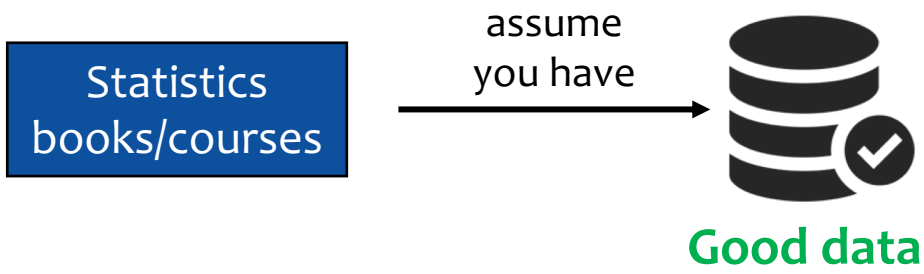


the **population distribution** is **unknown**



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

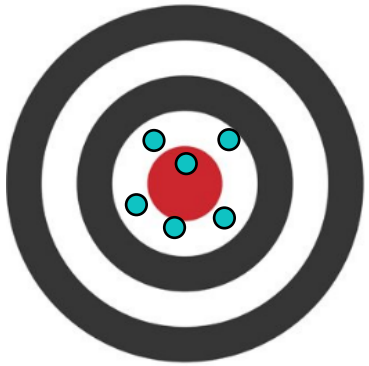
Garbage In → Garbage Out

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

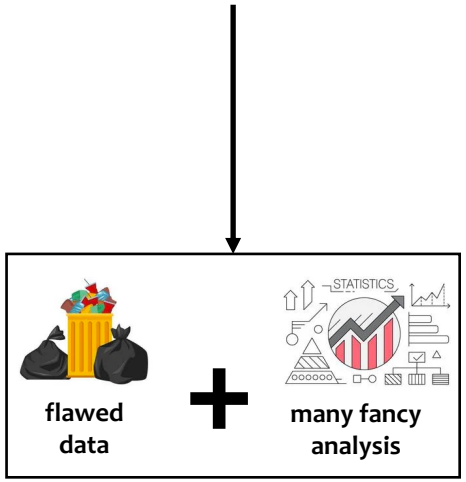
✔ Accuracy

✘ Precision

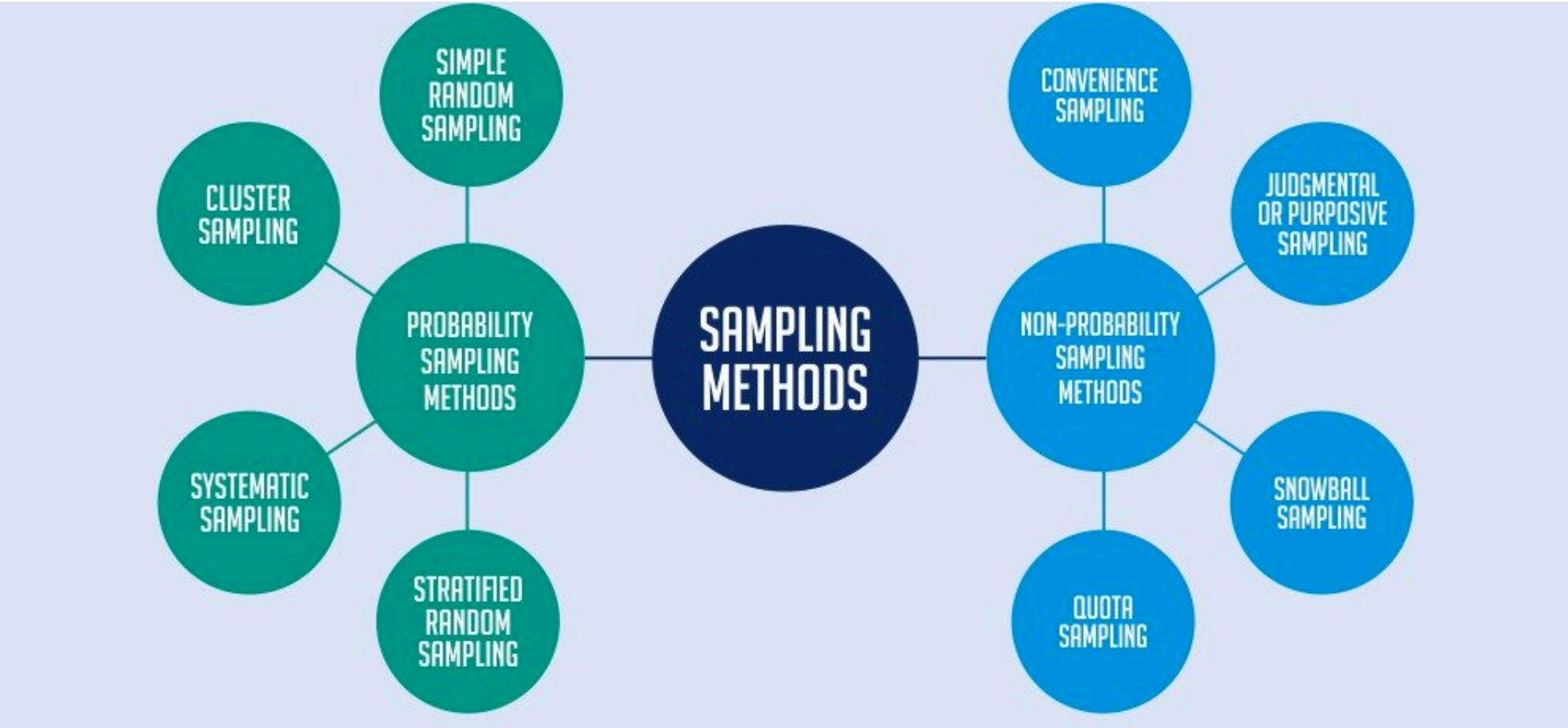
✘ Accuracy

✔ Precision

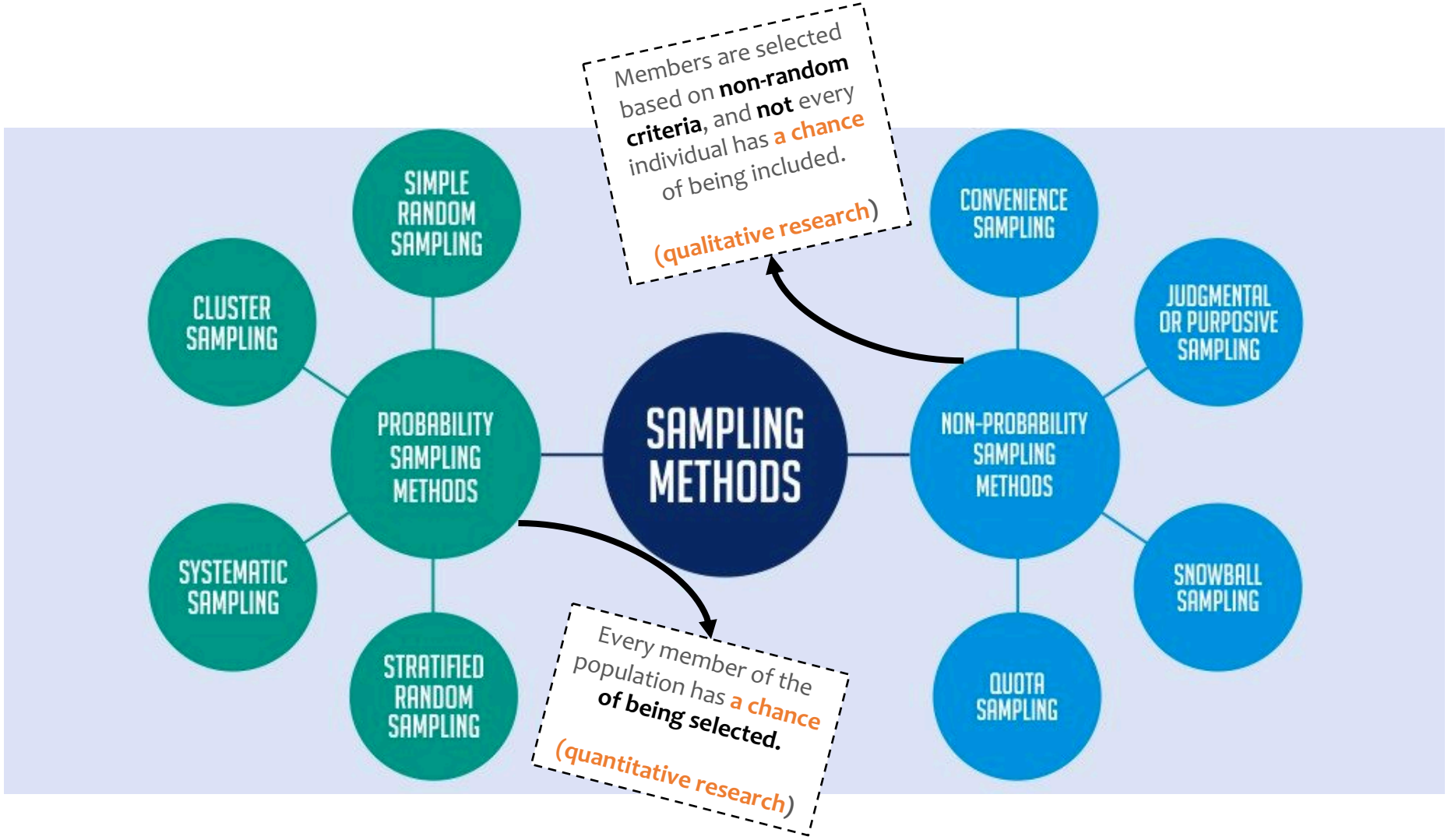
✔ Accuracy



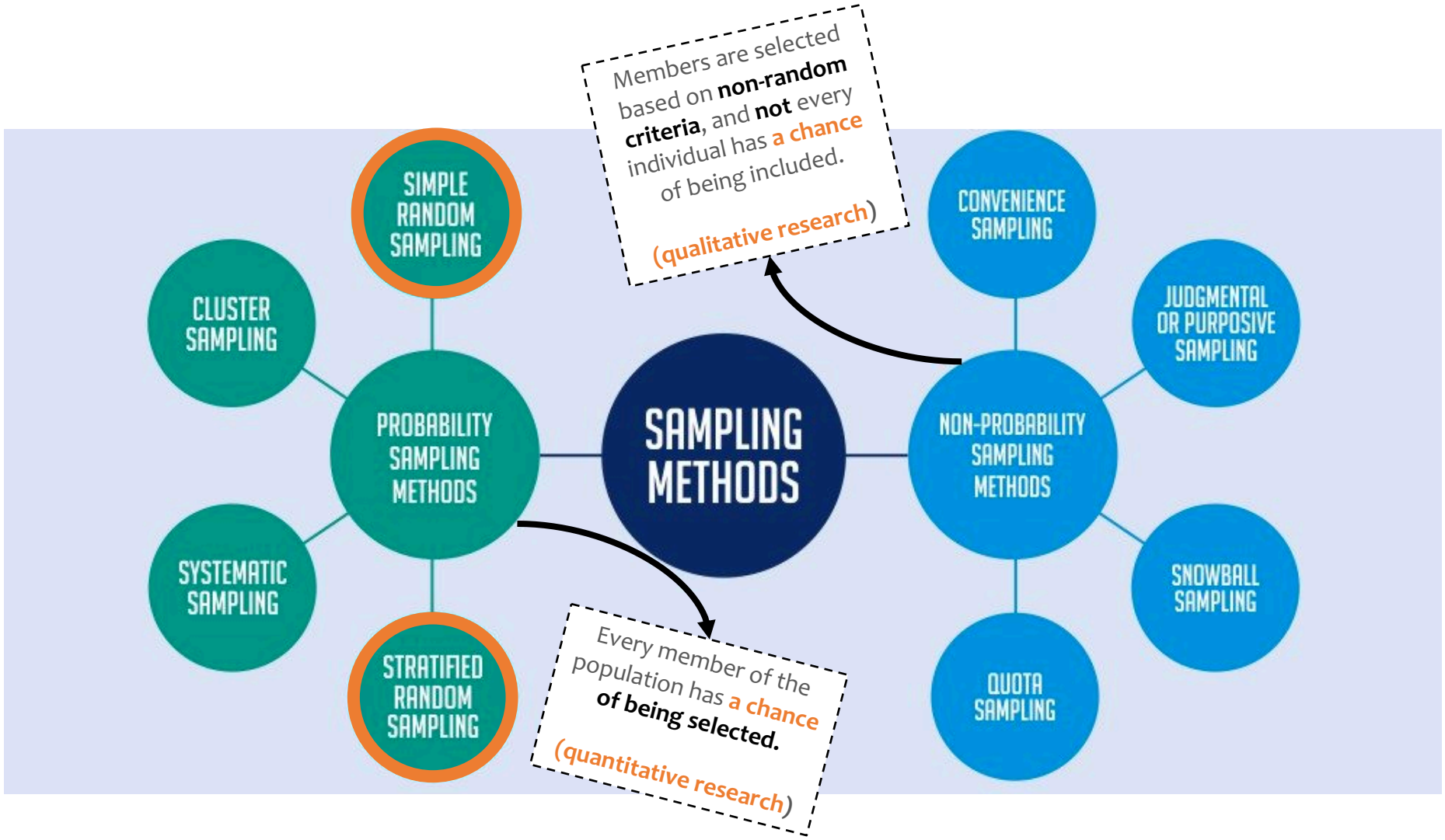
Types of Sampling



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



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

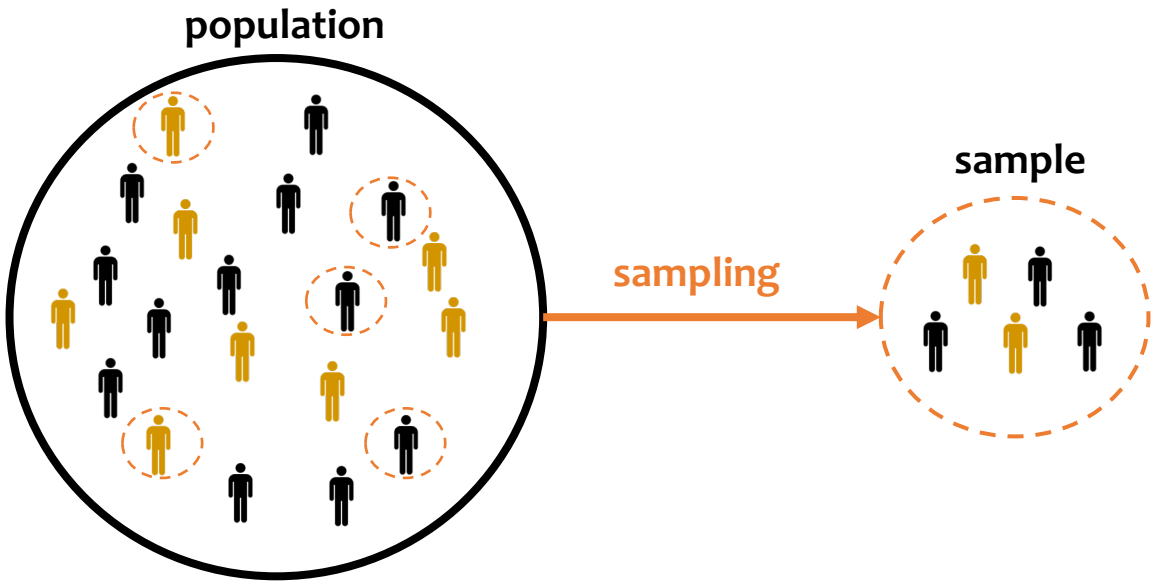


https://www.questionpro.com/blog/types-of-sampling-for-social-research/

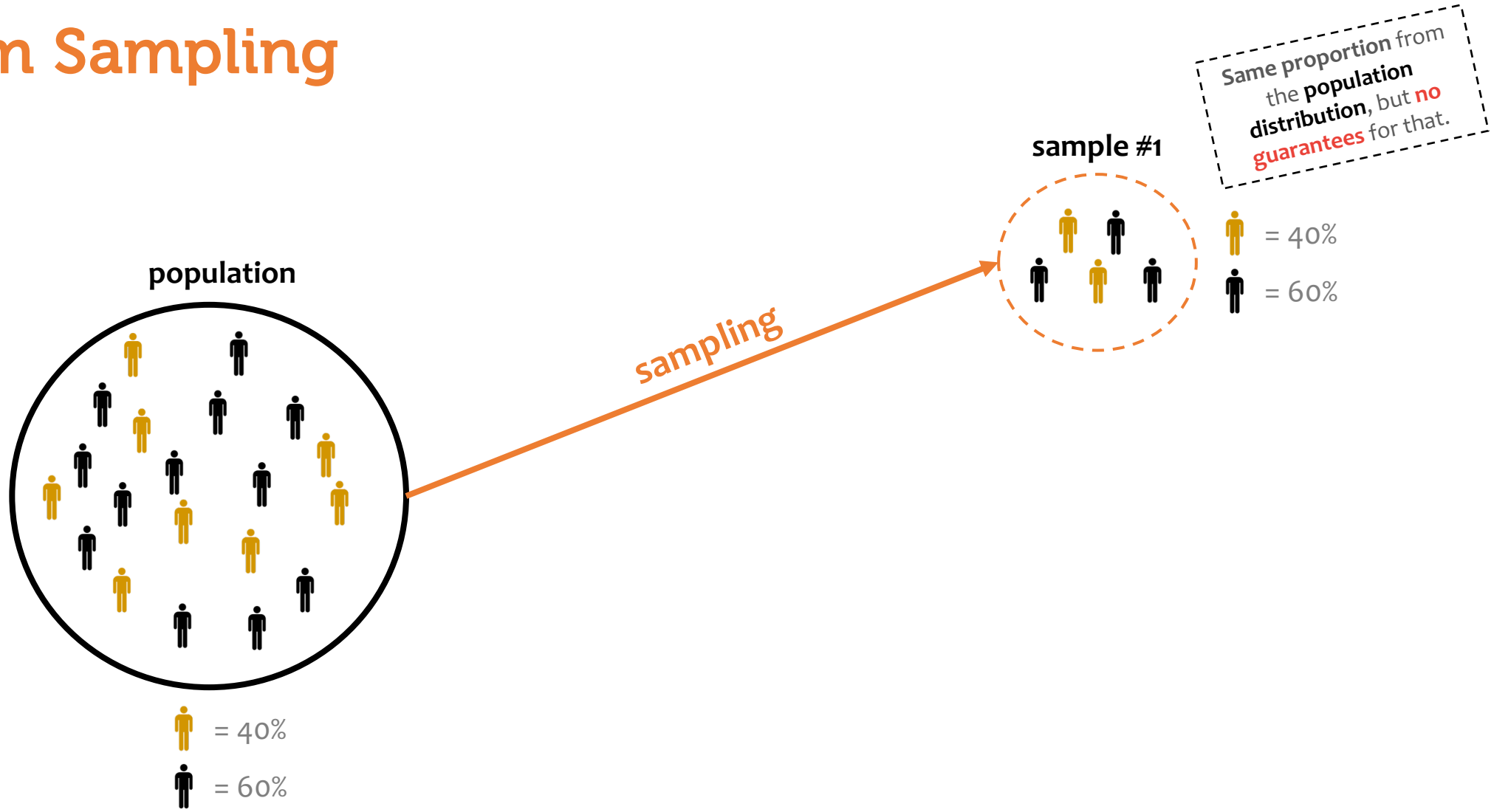
Random Sampling

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

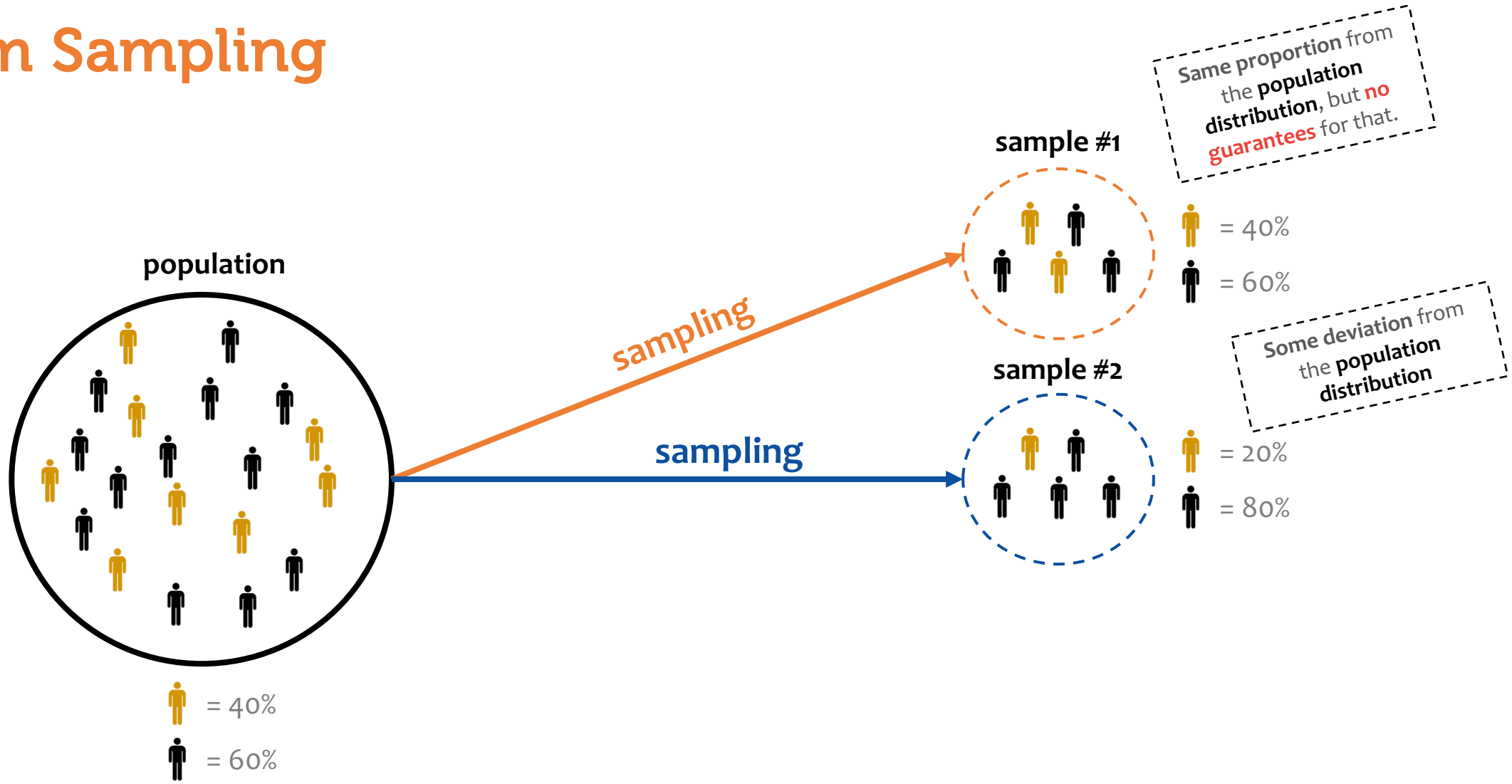
Each **member/observation** has **an equal chance** of being chosen for the **sample** at each draw.



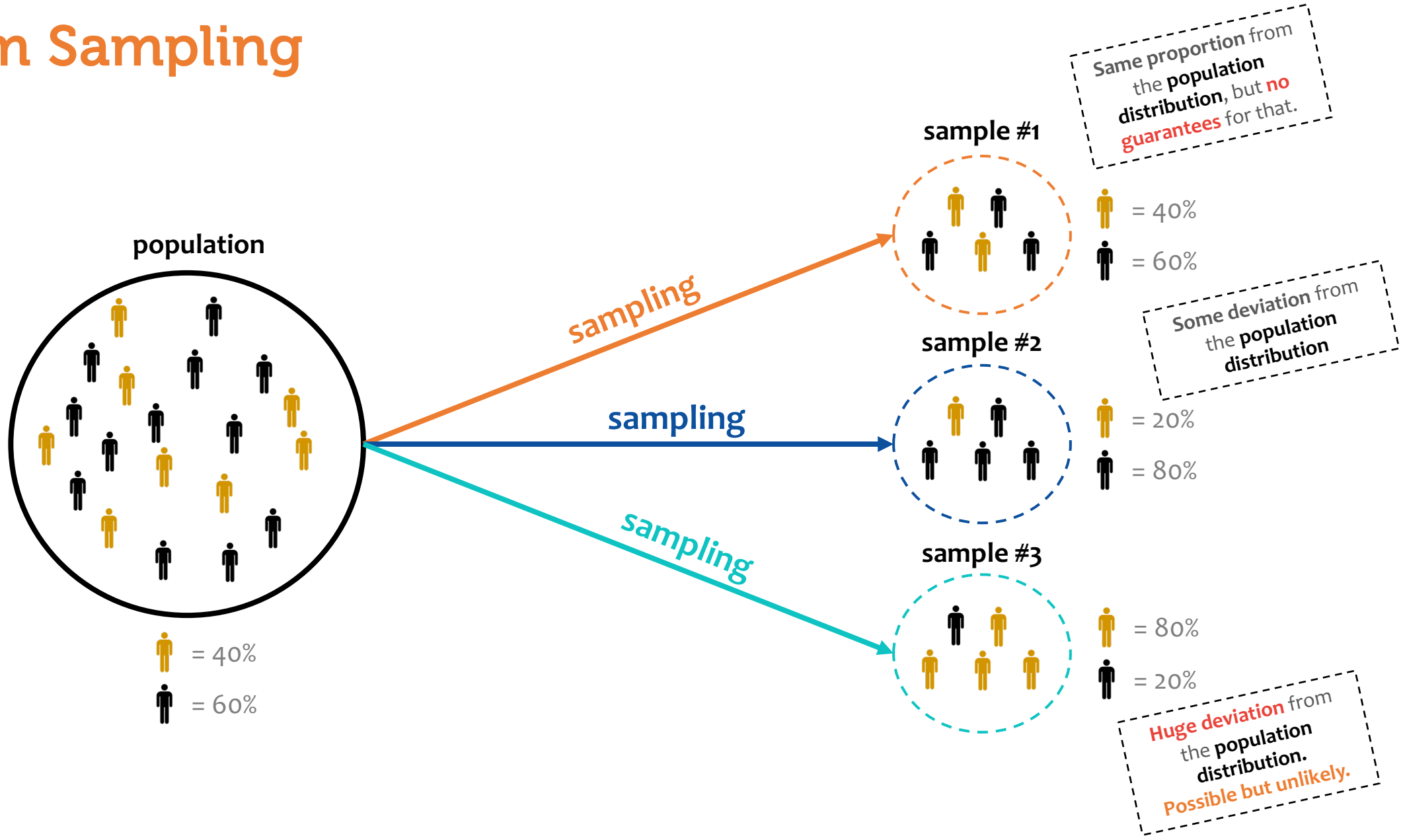
Random Sampling



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

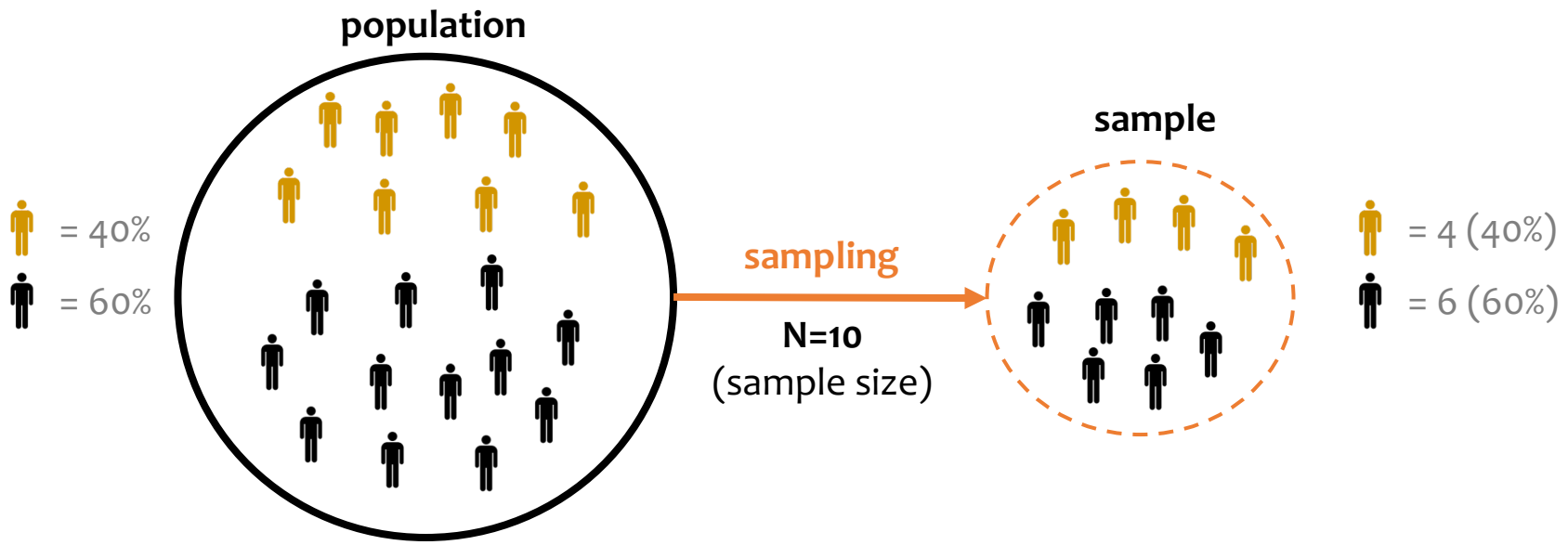
Dataset: Google Play Store Apps: https://www.kaggle.com/lava18/google-play-store-apps

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

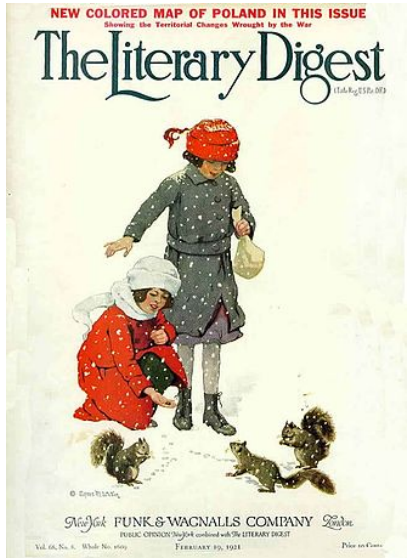
Bias

Bias

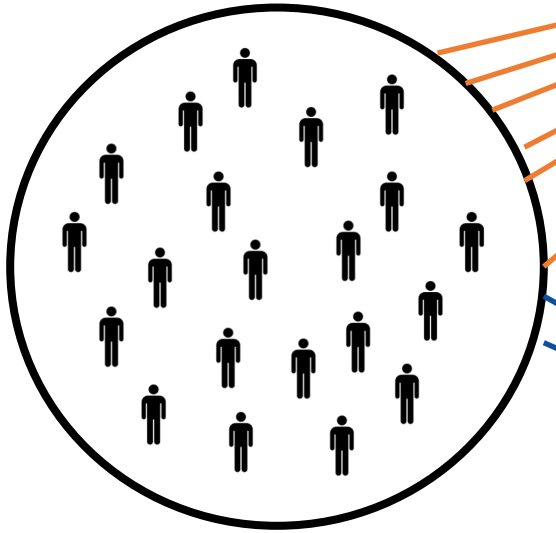
- **Measurement** or **sampling error** that are systemic and produced by the measurement or **sampling process**.
- **Tendency** of a statistic **overstimate** or **underestimate** a parameter.

Classical Example of Selection Bias

The Literay Digest (1936)

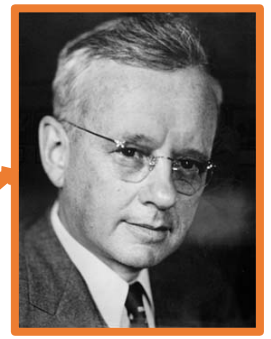


pooled its

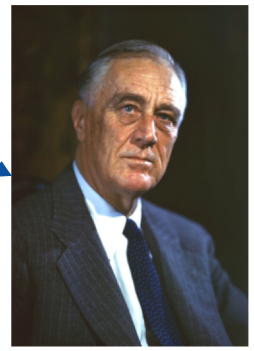


subscribers

with similar:
- profiles
- financial situation
- political tendencies



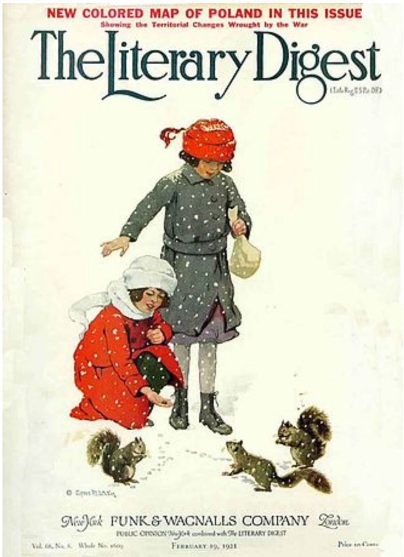
Alfred Landon



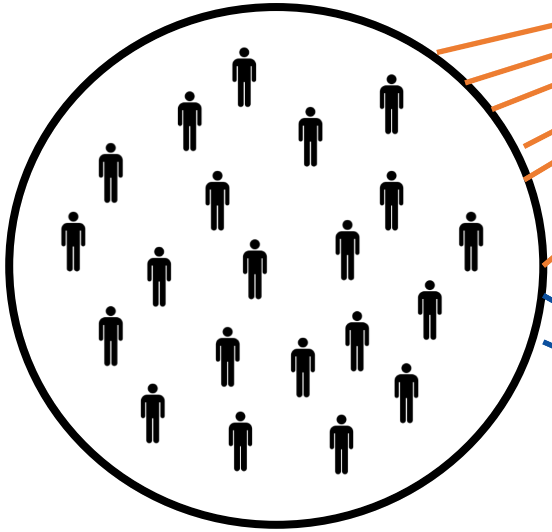
Franklin Roosevelt

Classical Example of Selection Bias

The Literay Digest (1936)

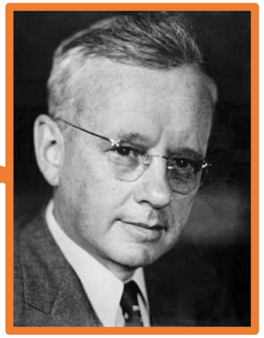


pooled its



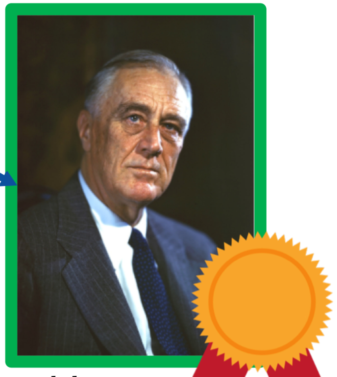
subscribers

with similar:
- profiles
- financial situation
- political tendencies



Alfred Landon

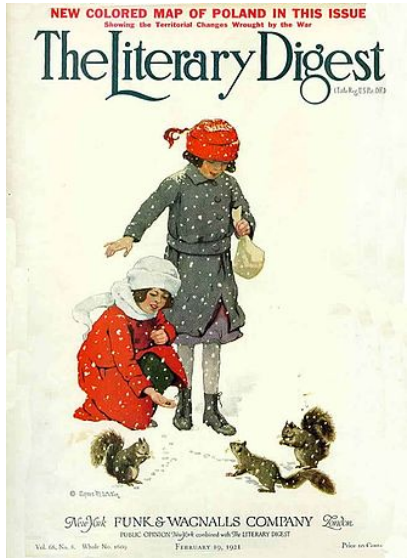
Winner



Franklin Roosevelt

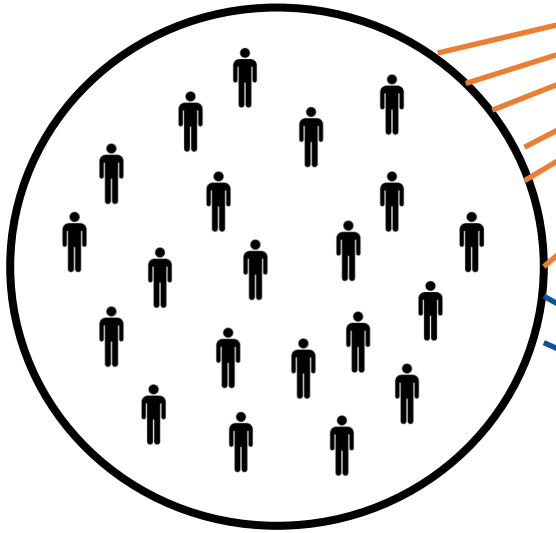
Classical Example of Selection Bias

The Literay Digest (1936)



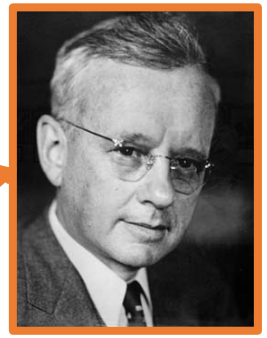
Selection Bias

pooled its



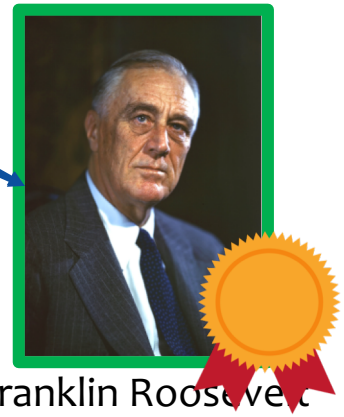
subscribers

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
- …

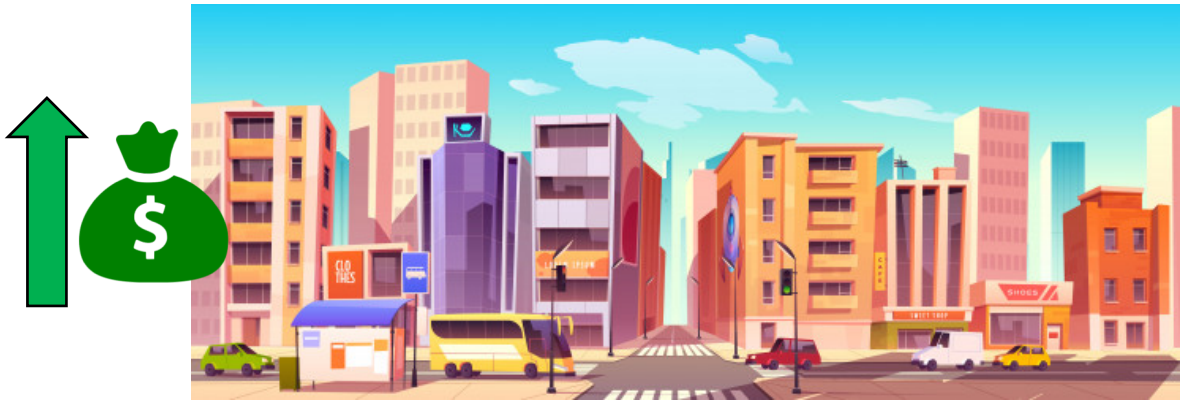
Selection Bias

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**.



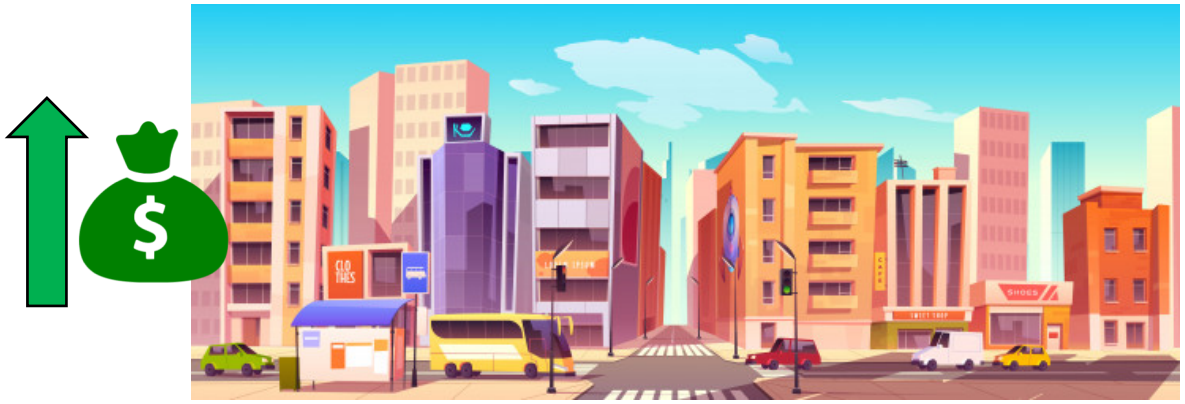
fancy neighborhood

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.

D1EAD – Análise Estatística para Ciência de Dados 2021.1

Data Distributions (Part 1)

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