"Data is like garbage! You'd better know what you're going to do with it before you collect it." ~ Mark Twain

DESCRIPTIVE STATISTICS

for Data-driven Decision Making with Python

A GUIDE TO STRAIGHTFORWARD, DATA-DRIVEN DECISION MAKING WITH DESCRIPTIVE STATISTICS

Pratik Shukla & Roberto Iriondo

For all of those who encouraged
me to fly towards my dreams.
Let's take off.
~ Pratik Shukla

For my grandparents, Nelida and Jesus,
my parents, Milder and Jose,
my loving wife, Saray,
my son Dominic,
and those who believe in me.
~ Roberto Iriondo

Published by
Towards AI Co.
15 N Sprague Ave.,
Suite 41086,
Pittsburgh, PA, 15202

Distributed in the United States of America

January 31, 2021

Book design by Roberto Iriondo
Graphics by Pratik Shukla
Illustrations by Canva (canva.com)
Cover Design by Roberto Iriondo
Digital Distribution by Gumroad (gumroad.com)

Additional credits appear on pages 152-153, which constitutes a continuation of this copyright page.

Version 1.0.0

Resources
The accompanying source code for this book is available on:
Google Colab: https://news.towardsai.net/descriptive-statistics-colab
Github: https://news.towardsai.net/descriptive-statistics-github

Thank you for your support!

Contents

Contents (Continued...)

Preface to the First Edition

> “Data is like garbage! You'd better know what you're going to do with it before you collect it.”
>
> ~ Mark Twain

Data science and machine learning are scientific disciplines that are ruled by programming and mathematics. Nowadays, most corporations globally generate immense amounts of data that can be further analyzed and visualized by experts to understand trends and forecast predictions. For instance, we can only perform accurate data visualization if our data is clear and understandable.

However, organizations' data is (frequently) too messy to tinker with—therefore, finding structures and important patterns in data is a crucial task for data science. Statistics provides the methods and tools to find hidden structures and patterns in data so that specialists can make predictions from them — making statistics the most fundamental step in the data science and machine learning scope. We need statistics to transform observations into information. In machine learning, we use a variety of algorithms for prediction, classification, and clustering. Although, there are many useful libraries available to use that will perform mathematical calculations for us.

Nevertheless, we need to know the math behind each of the algorithms and statistical methods we use because knowing these gives us insights into what we are doing and ultimately find our why behind our data-driven decisions.

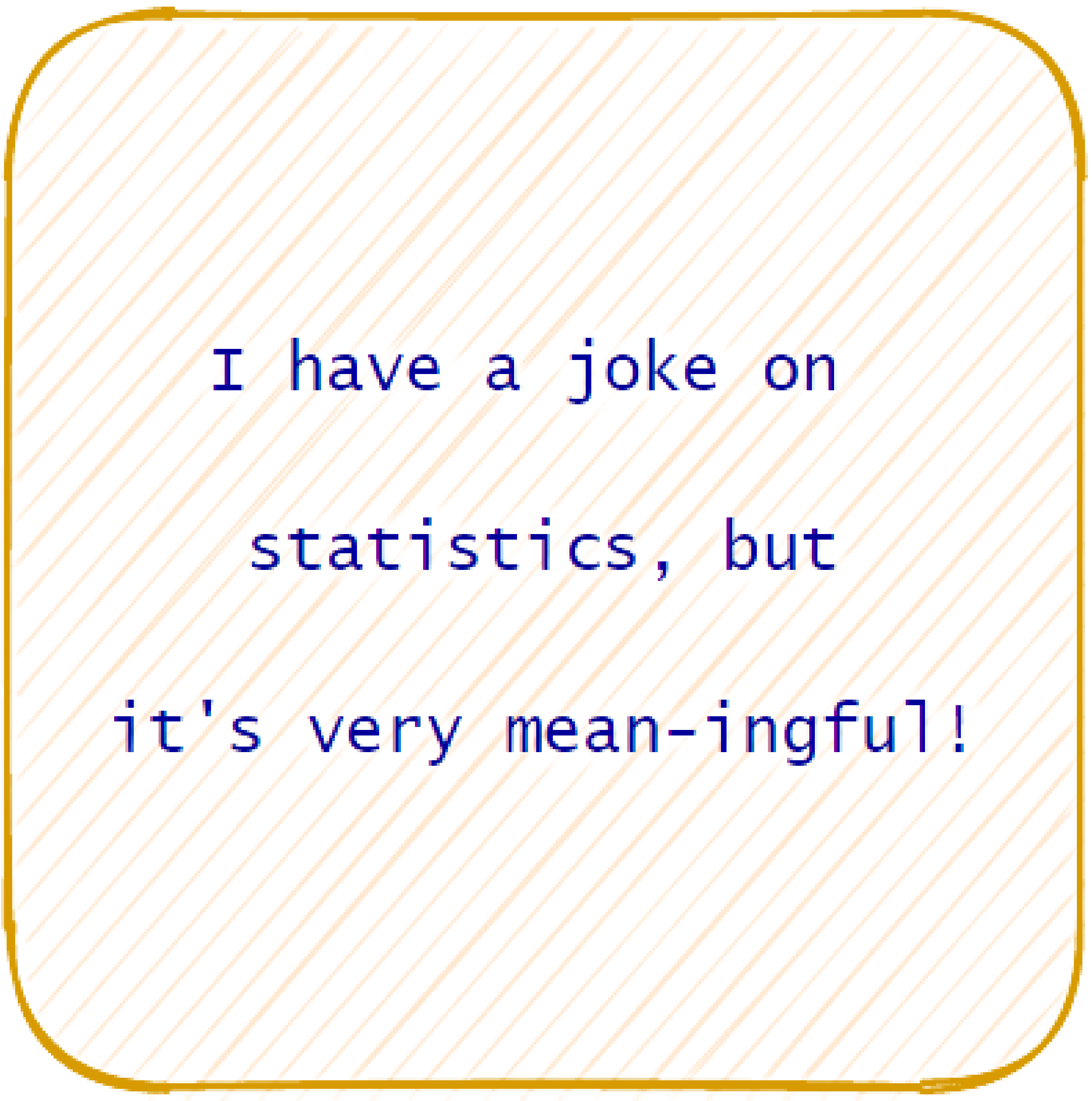
This book aims to understand the core concepts that form the base for data science, machine learning, and related analytical fields. Our primary goal is to show our readers how to perform calculations and why we need such a methodology. In this book, we try our best to showcase a few core statistical methods with their theories and code examples with python.

Please note that in some cases, python programs' output may vary from the outputs we get by applying the theoretical concepts—the reason behind it is that we will be using python libraries to display outputs, and in some cases, the programmers that created such libraries used different logic to create their methods. Consequently, we consider it crucial to understand the core logic of what we explain in the theoretical concepts because once we understand the concept, it is relatively easy to write pseudocode and code for the task at hand.

"The quiet statisticians have changed our world; not by discovering new facts or technical developments, but by changing the
ways that we reason, experiment, and form our opinions..."

–Ian Hacking

Introduction



I have a joke on statistics, but it's very mean-ingful!

In this book, we are going to take a look at the descriptive statistics. However, before going into statistics, it is crucial to know the basic things we will need for it. First of all, statistics works on data. If there is no data with us, then there is no possible way statistics can work. We use the data to perform various operations on them to make some helpful conclusions out of it.

Nevertheless, sometimes it is impossible to gather data of everyone related to the research. For example: If we want to measure all the humans' weight on the earth, then it will be impossible to get data of all the humans. So that is why we take samples of data and then perform operations on them.

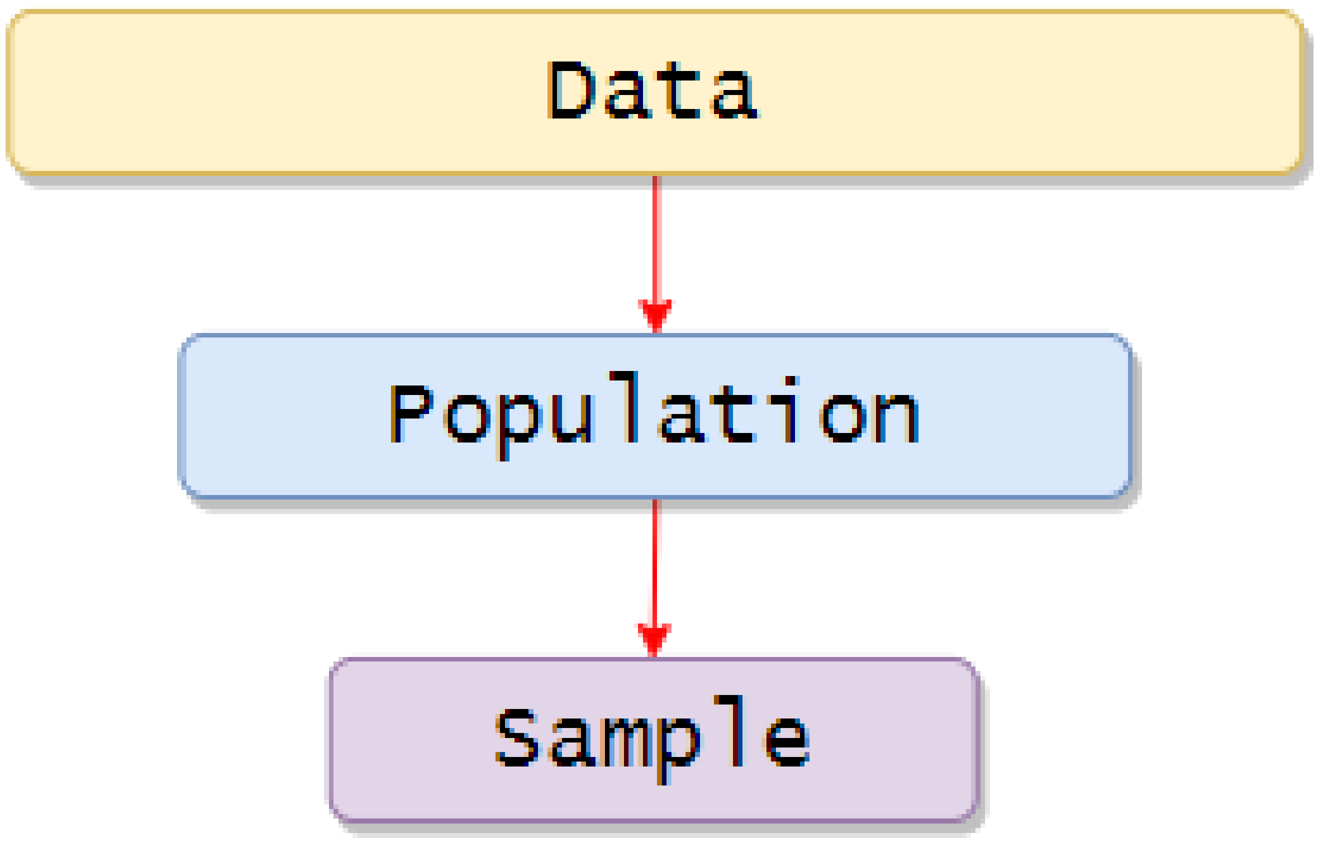
First of all, we will see the population and sample, and then we will discuss a few sampling techniques.

Population and Sample:

In the study of statistics, our primary focus is on data. Let's see two critical types of datasets:

1. Population
2. Sample

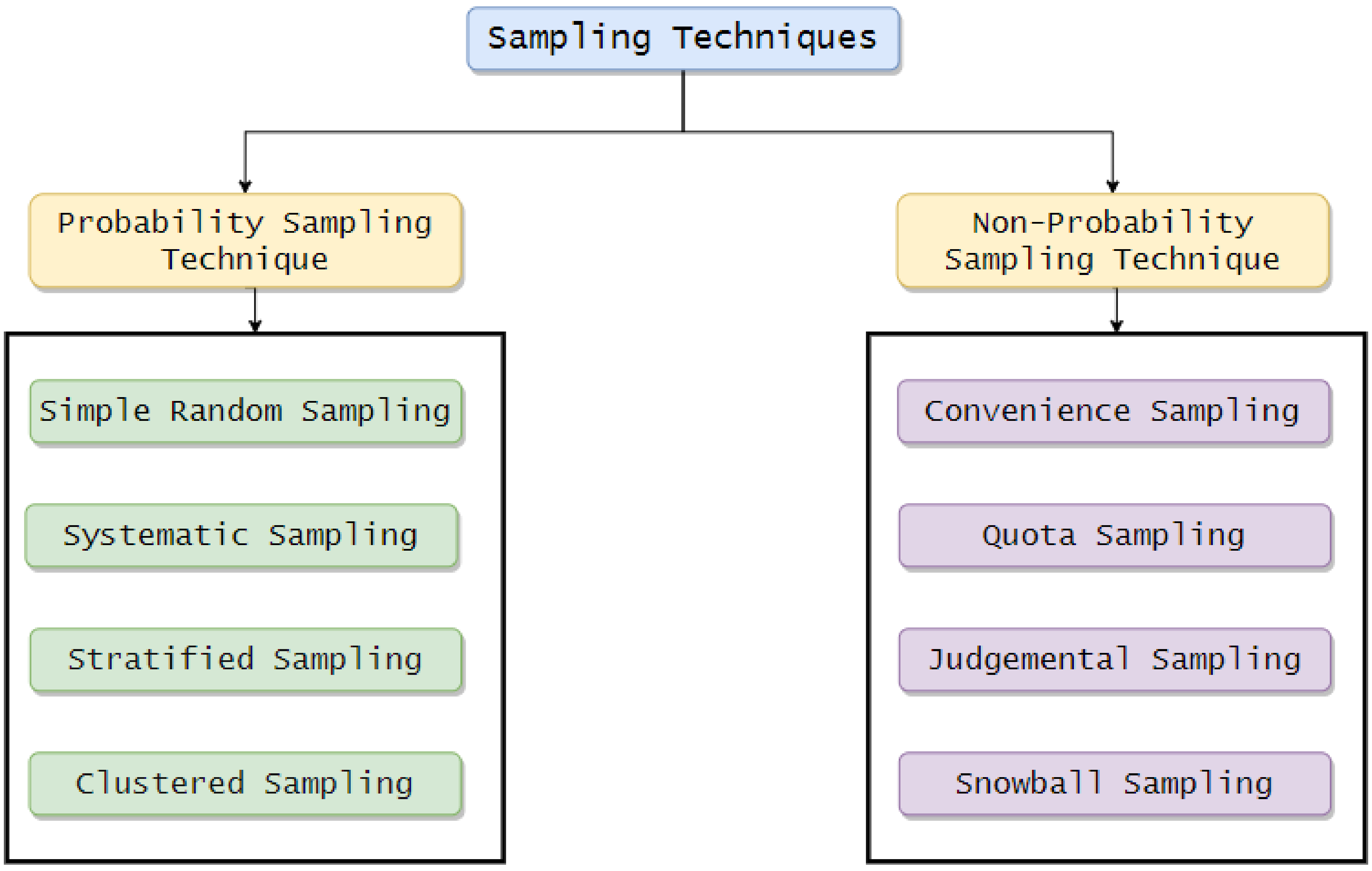
The main difference between population and sample can be viewed by the number of observations in each dataset.



Population includes all the elements or observations that are related to our study. The population is usually denoted with (N), and the numbers we obtain while working with the population are called parameters.

The sample includes one or more observations from the population. Samples are usually denoted with (n), and the numbers obtained by working with samples are called statistics. Now there are several methods to derive samples from the population. In this book, we are going to see a few of them.

Sampling Techniques:



Probability Sampling Techniques:

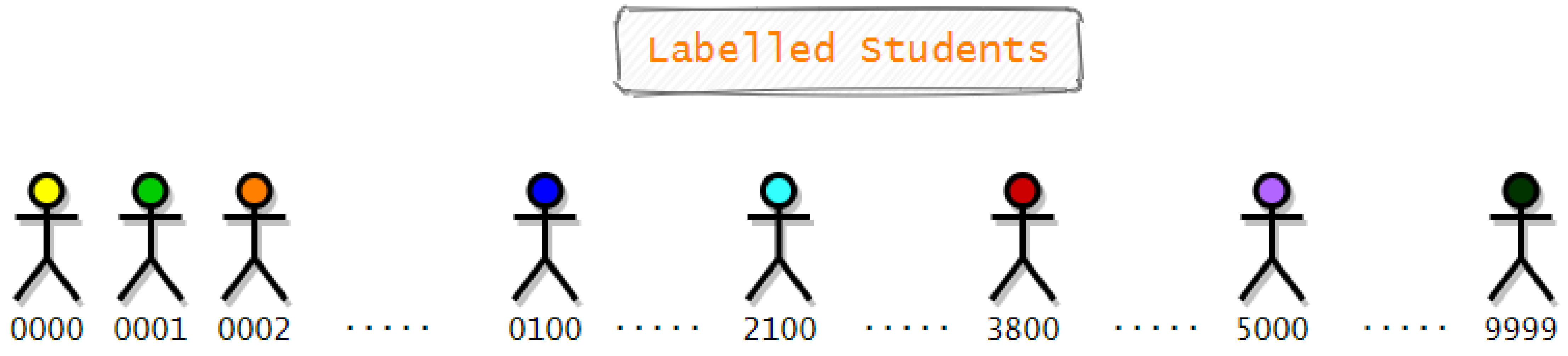
When each subject or entity has an equal, non-zero chance of getting selected for a sample, it is called the probability Sampling Technique. These samples usually represent a larger population. They provide credible results as there are low chances of bias in sampling.

Simple Random Sampling

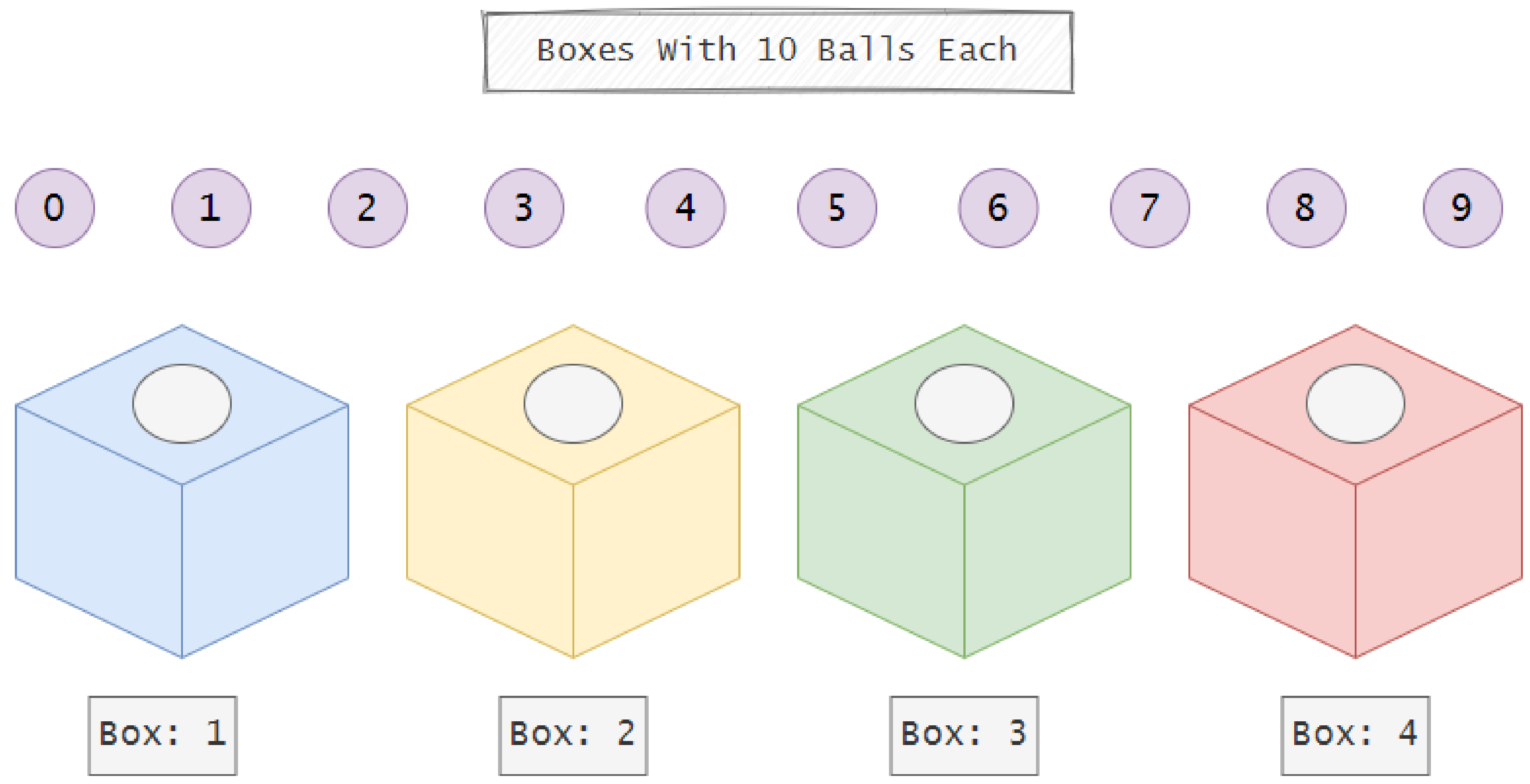
In this method, each individual from the population is chosen pseudorandomly, and each member of the population has an equal probability of being selected. In other ways, we can say each member from the entire population has an equal chance of being selected. A simple random sample will be an unbiased representation of the entire population.

For instance:

Let’s say we want to choose 10 students from a group of 10000 students randomly. First of all, we need to assign labels to each student. Since there are 10000 students, the labels will start from 0 and end with 9999. Here is a visual representation of labeled students.

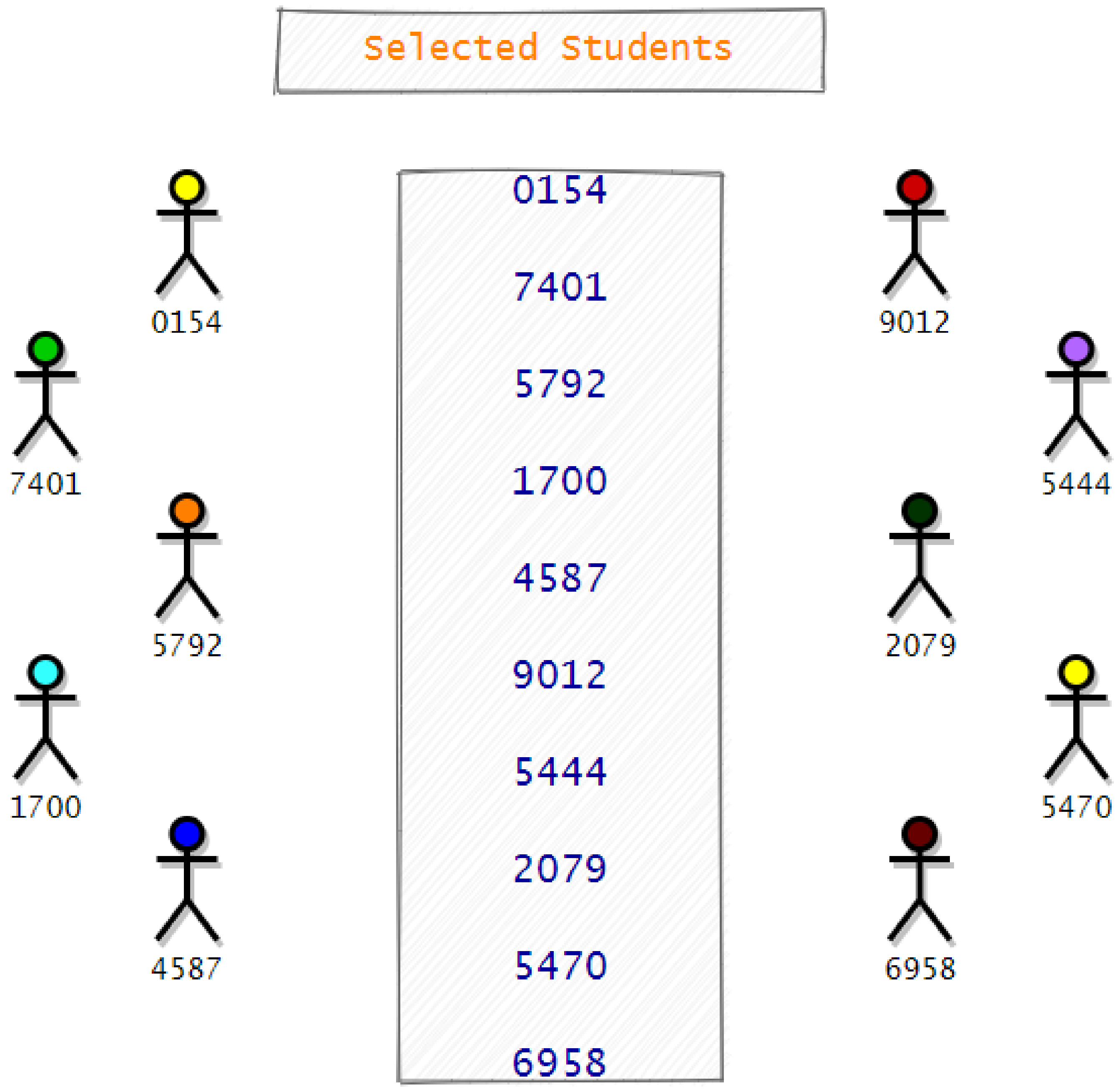


Now we will use simple random sampling to select 10 random students from a total of 10,000 students. To do that, we will use 4 boxes, and each of the boxes will have balls labeled from 0 to 9. Keep in mind that all of the boxes are opaque or non-transparent. So there will be equal chances for each of the students.

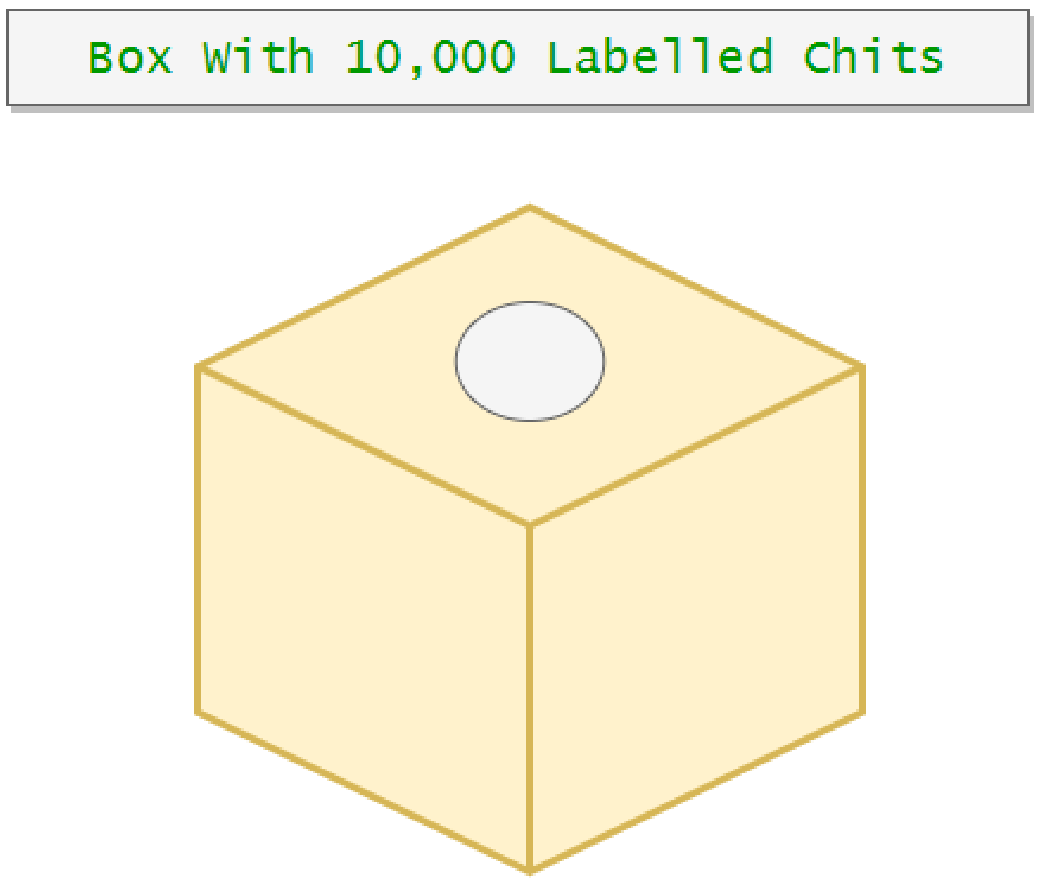


Next, we will call a small child and ask him to draw a ball from each box. Now, we will note down the number printed on the ball. For example, if the child draws a ball labeled as two from the first box, a ball labeled as five from the second box, a ball labeled as nine from the third box, and a ball labeled as 0 from the fourth box, then we will choose the student with the label of 2590.

We will repeat the same process until we have labels for ten students. Hereafter noting down the numbers, the balls will be replaced in their original boxes. So at any point in time, there will be ten balls to choose from. Now here is the list of the students selected by the child.



Making this is one way we can choose random students from all of the students. Now notice that in this case, we replaced the balls in their respective boxes. Next, we will see another way we can in which we can implement simple random sampling. However, here we will do it without replacement. So in this method, we will write each number in a chit and put those chits in a single large box, and then we will draw 10 chits from it. Now we do not have to put those chits back in the box. So this is without replacement.



Advantages:

1. Simple Random Sampling reduces sampling bias.
2. Ease of use.

Disadvantages:

1. It requires a little knowledge of the population.
2. It may generate sampling errors.
3. It is not suitable for a large population.

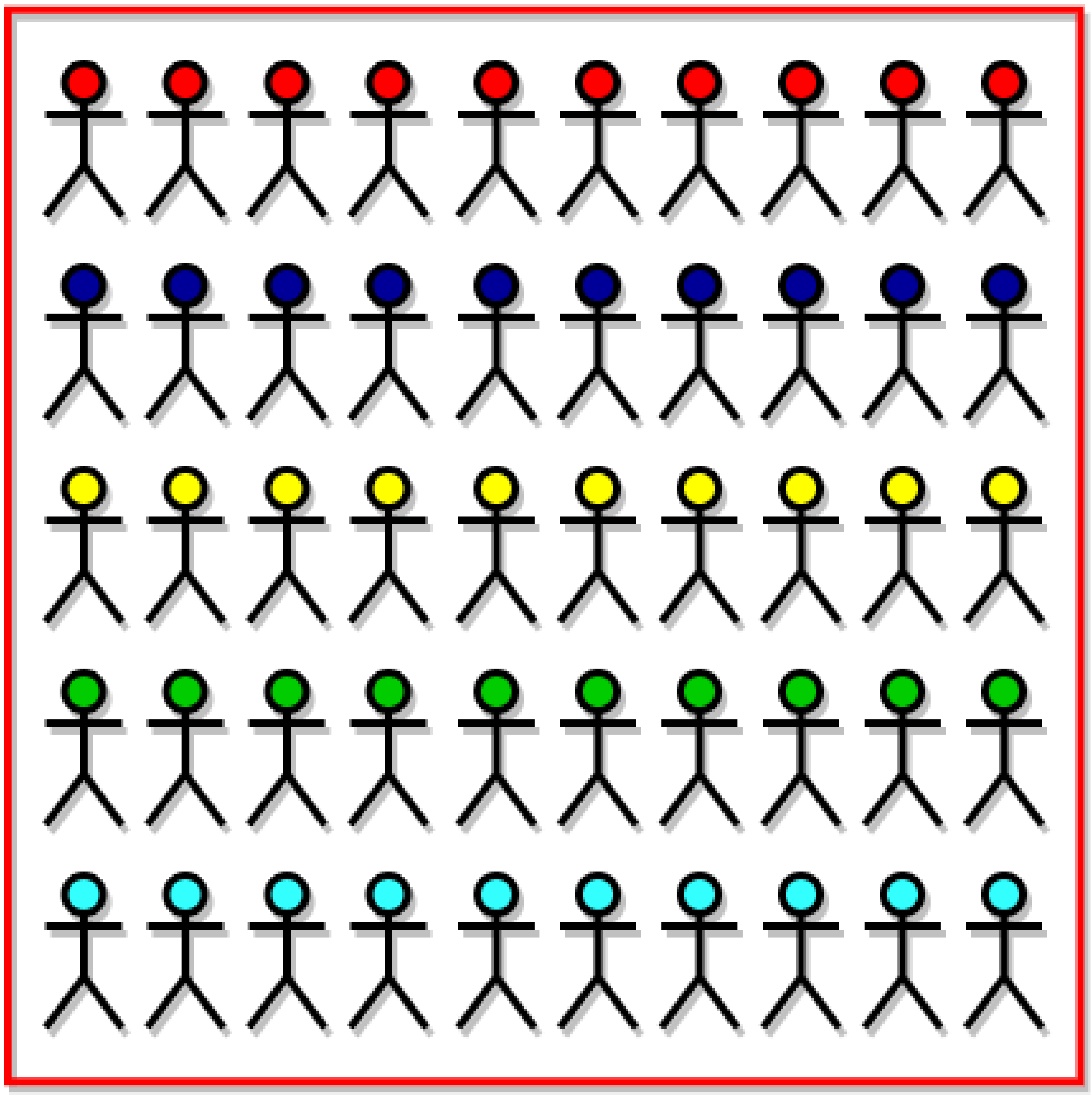
For example, If we are surveying a population of 10,000 students to find out how many students of them are left-handed. If we take a sample size of 50 students, none of the left-handed students may be selected. So this is a significant disadvantage of this method.

Systematic Sampling

In this method, we are going to select individuals at regular intervals. We can create an interval of any size we want, but it should be suitable for the sample size. It is generally used when the opinions of most people are logically homogeneous.

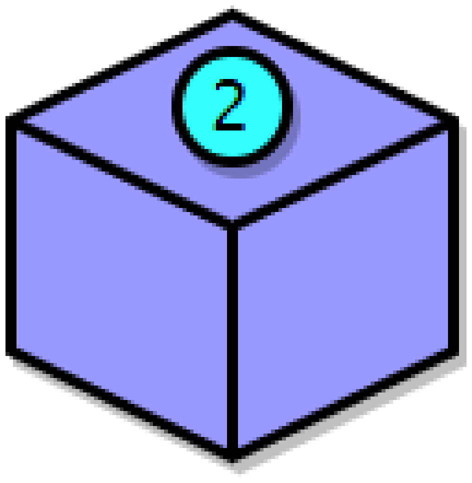
For instance:

Population

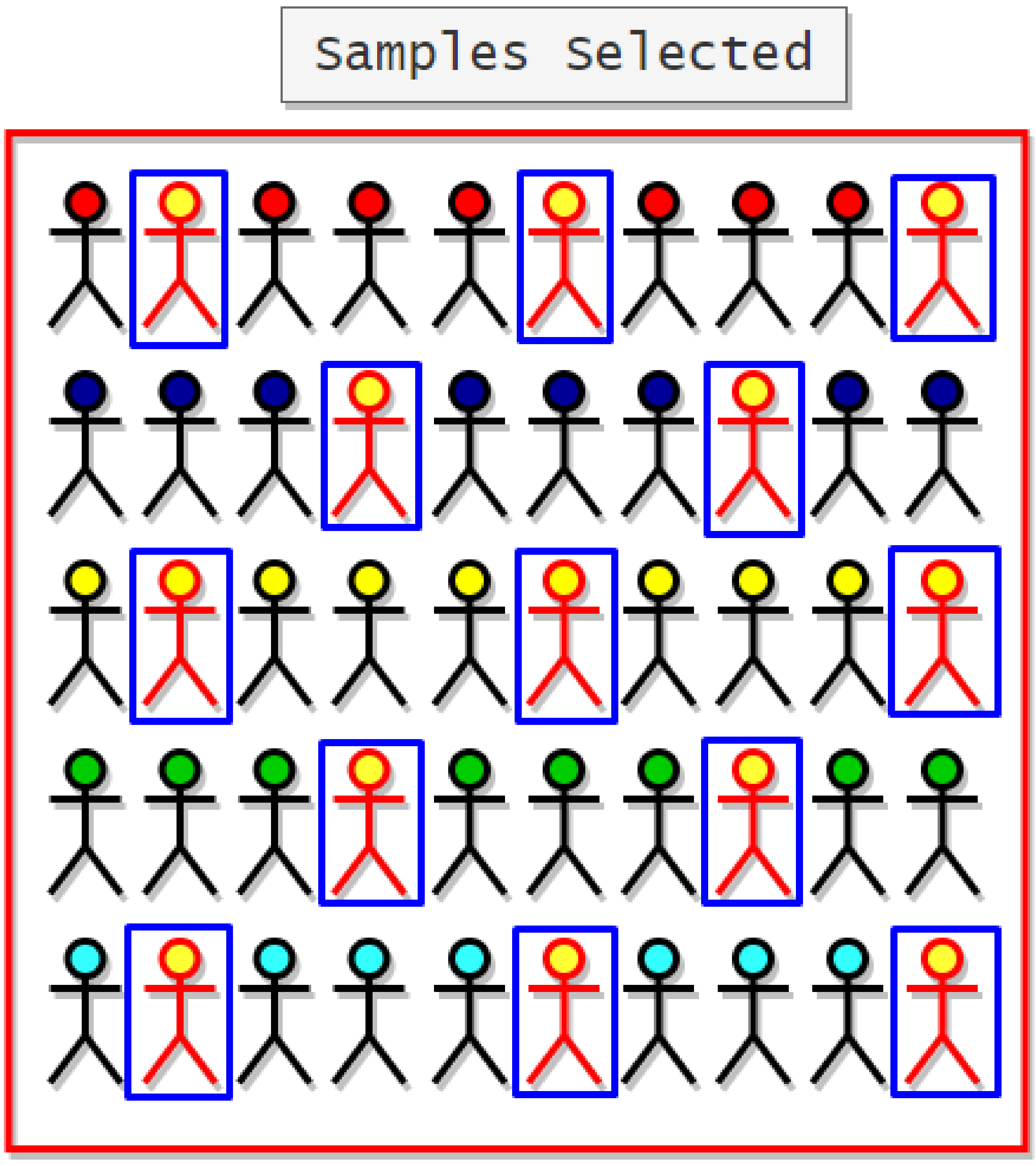


Let's say we have a population size of 50 students, and we want to select some samples from that. To do that, first of all, we need to decide the start point. We can decide it by rolling a die. Now that we have our start point, it's time to decide the interval. For interval size, we can decide to select every fourth person from the students. Notice that we can take the interval size whatever we want, but it should be per our sample size.

Rolling A Die



Here we will start with the second student, and after that, we will choose every fourth student. Here is the sample of students selected.



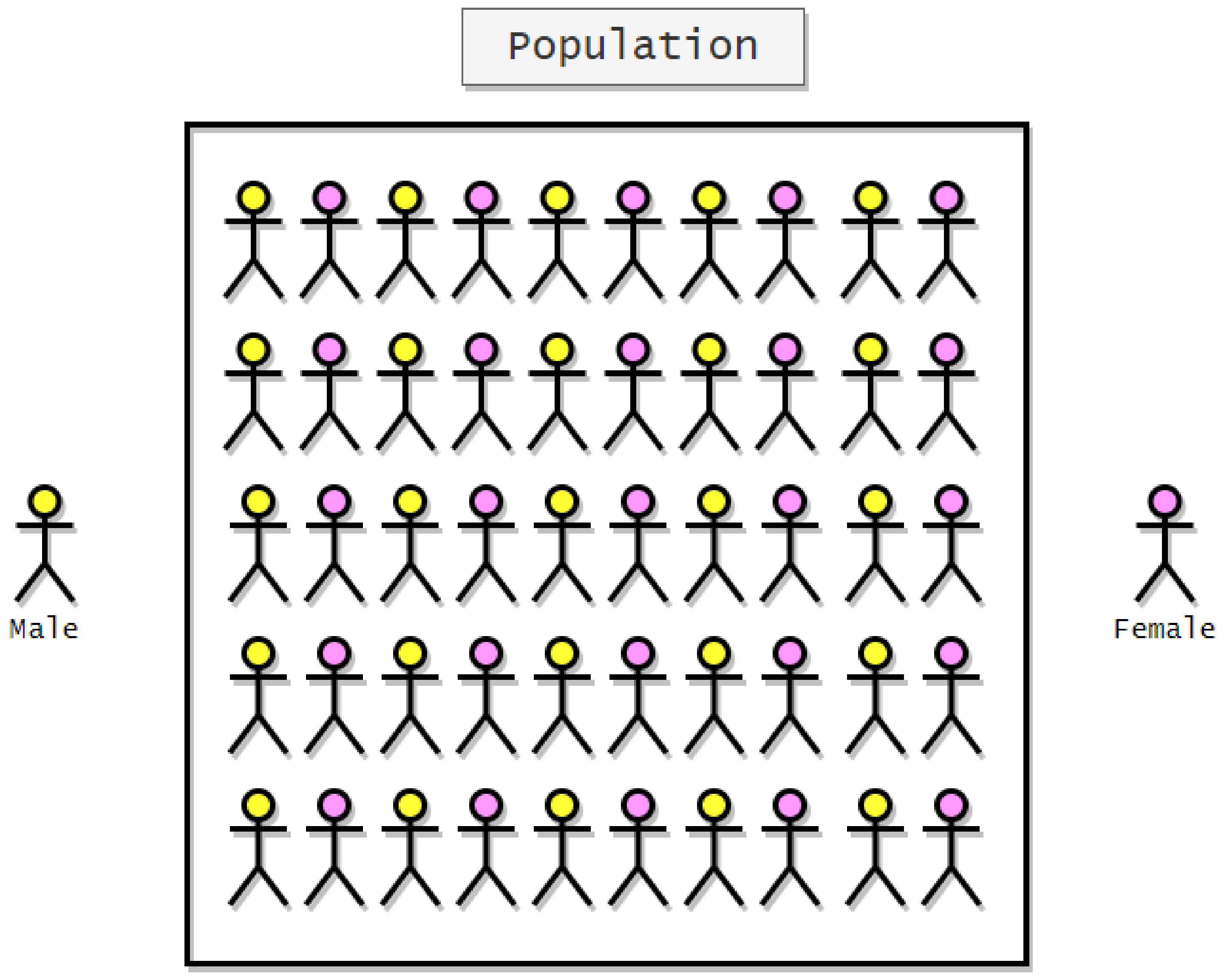
Advantages:

1. It is more convenient than simple random sampling.
2. It is easy to administer.

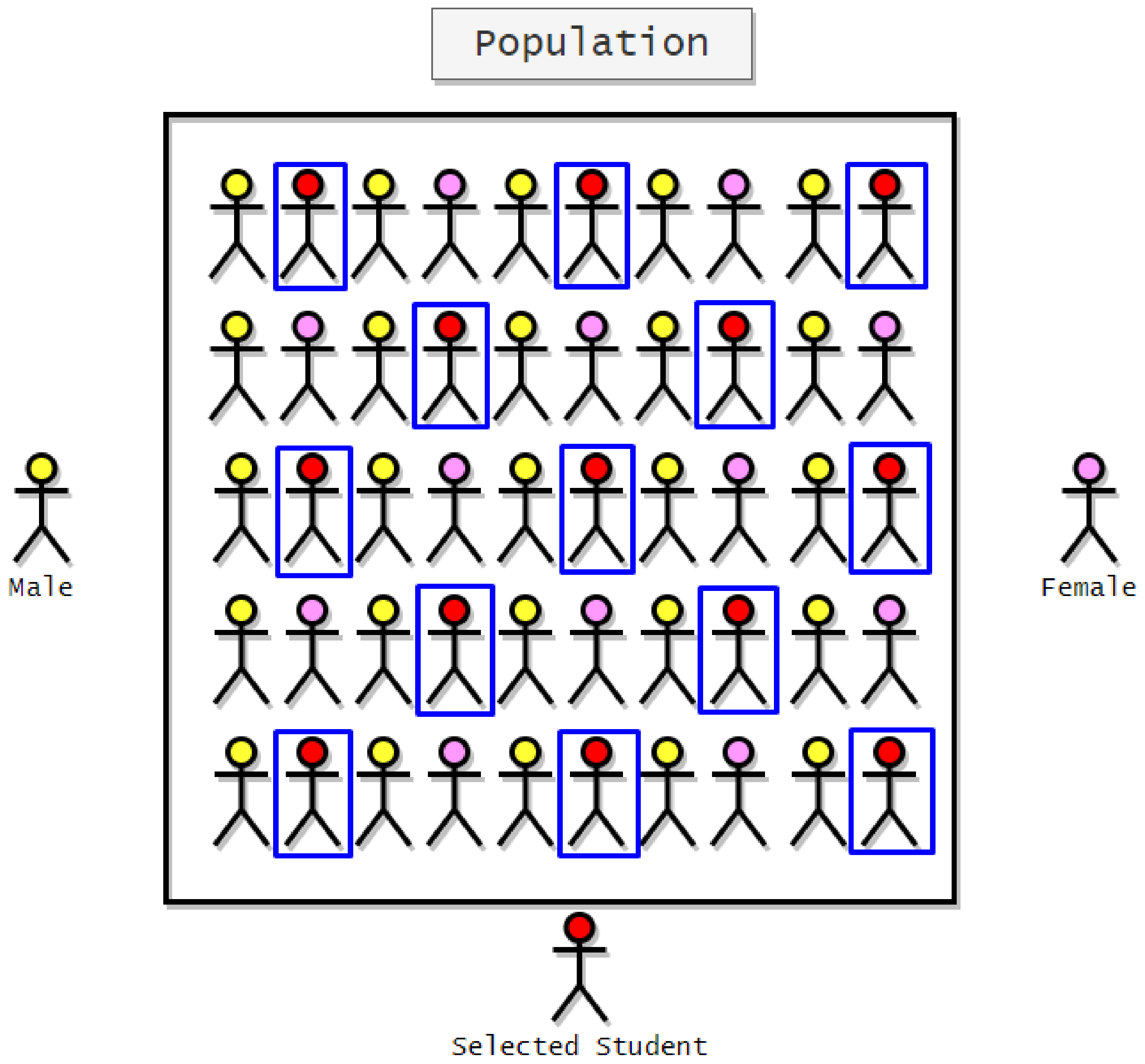
Disadvantages:

1. It gives biased results if there are any underlying patterns in the arrangement.

For example, suppose we want to measure the average weight of students. To do that from a population of 50 students, we are going to select some students randomly. Let’s say we selected the students based on our previous calculations.



Here is the representation of selected students:



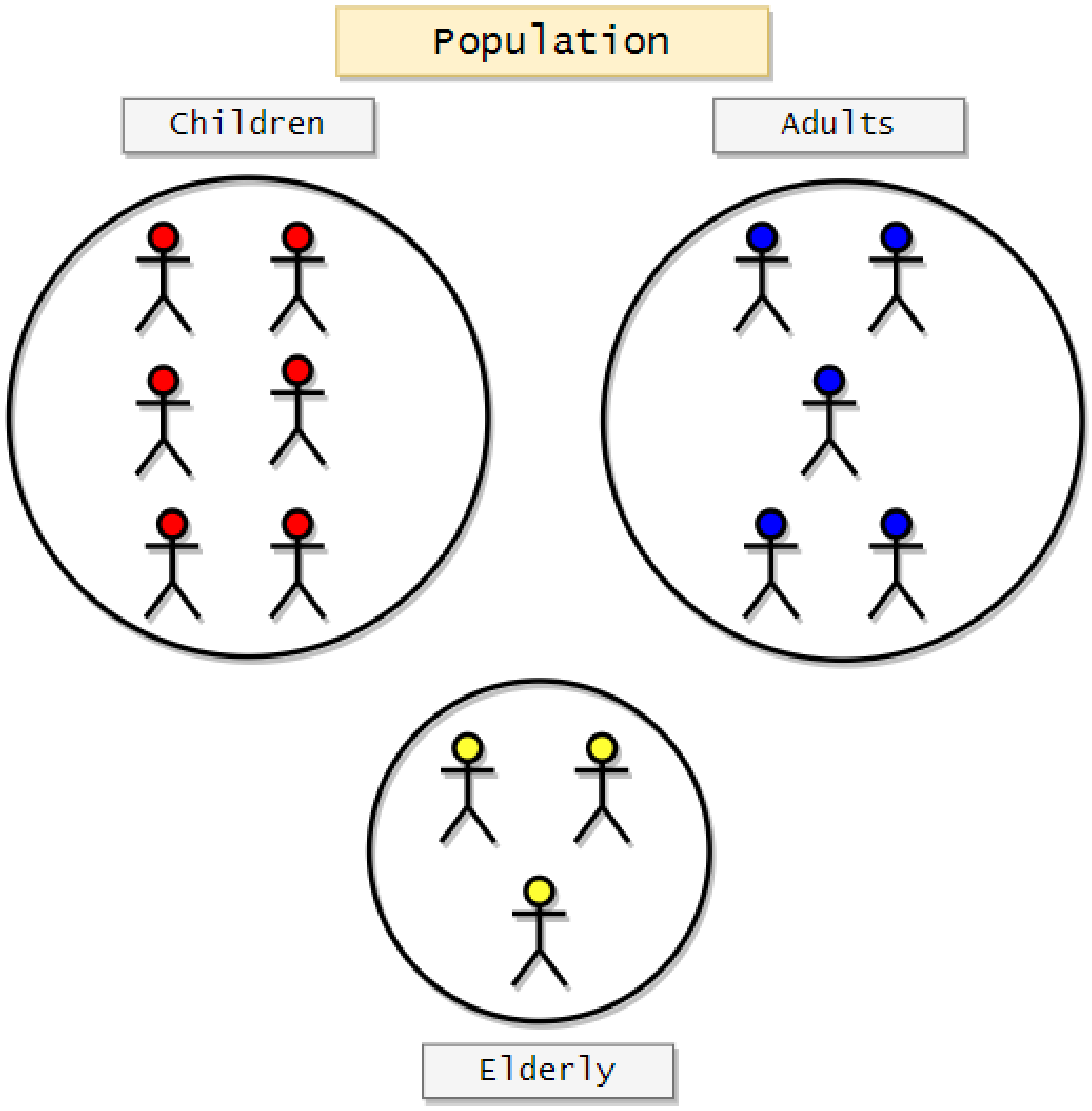
Here we can see that all the selected students are female. So our sample is biased towards female students. Furthermore, we can say that it will not give us the opinion of the entire population. To get rid of this problem, we can shuffle the population internally before taking samples.

Stratified Sampling

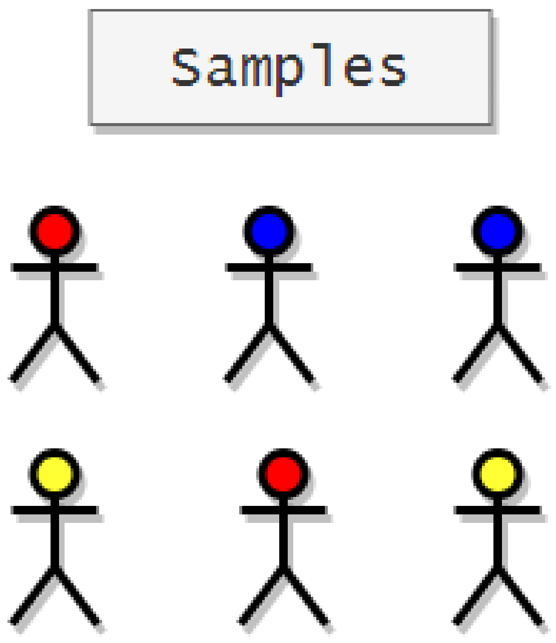
In this method, we will first divide the whole population into groups based on specific characteristics. We generally use this method when we think there will be differences in opinion based on a particular characteristic like age or sex, or race. So what do we do here is we divide the population into subgroups, and then we will use Simple Random Sampling to choose samples from each of these groups. Each of the divided groups is called a Strata. By doing this, we can ensure that we are not overlooking the opinions of a specific group. Let's take an example to understand it better. Suppose that we want to survey a population on the newly launched game. Here we can think that there can be some opinion differences between groups based on age. So we will divide the population into three groups.

1. Children (<18)
2. Adults (18+ and <65)
3. Elderly (65+)

After dividing them into these subgroups, we will use Simple Random Sampling to choose samples from each subgroup.



Now that we have divided the population into subgroups, we can use Simple Random Sampling to choose subjects from each group.



Advantages:

1. It improves the accuracy and overall representativeness of the entire population.

Disadvantages:

1. It required knowledge of dividing the population-based on specific characteristics.
2. It is complex and time-consuming.

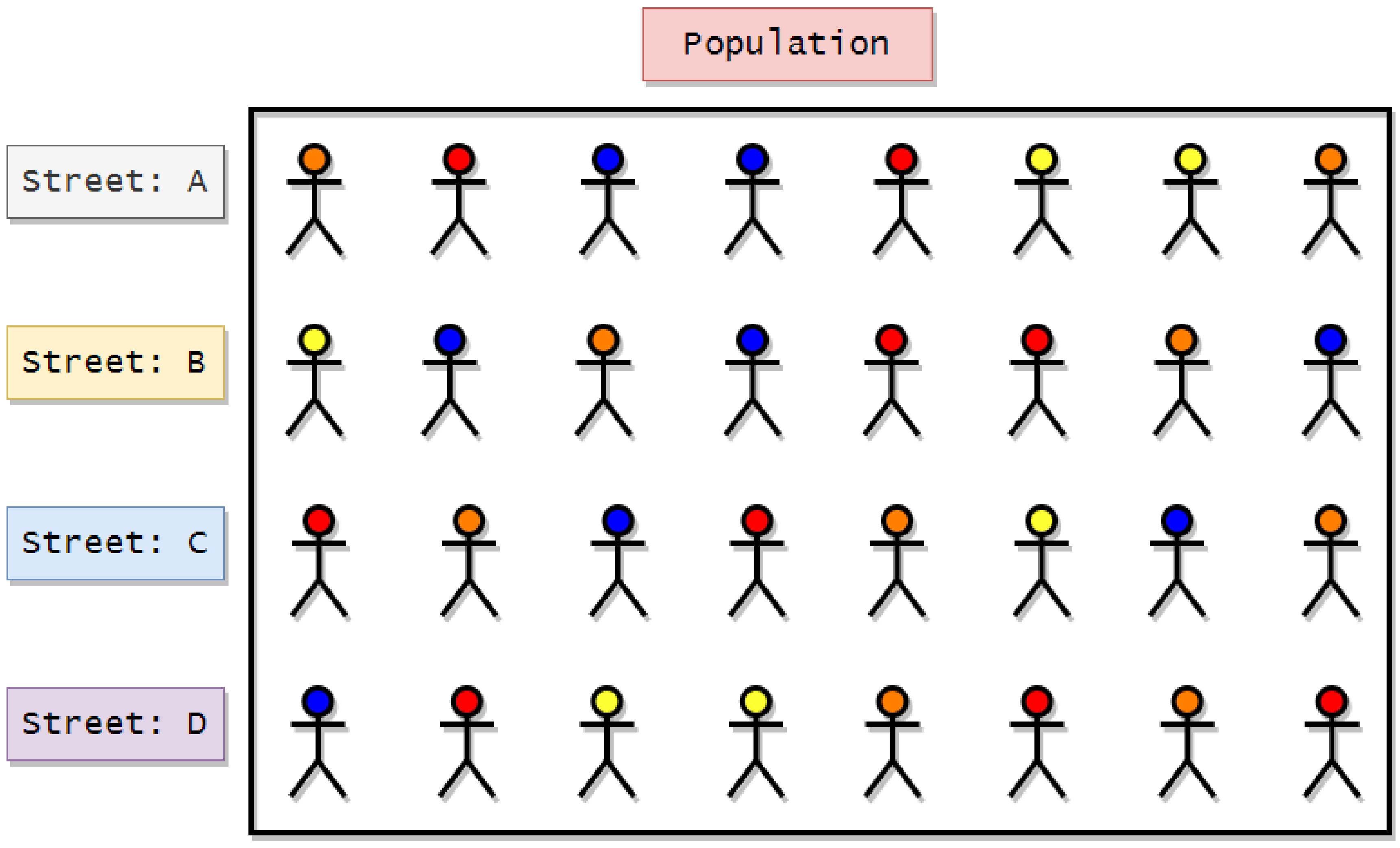
Clustered Sampling

In clustered sampling, we use clusters or subgroups to choose samples. Here the clusters are formed naturally. We do not have to create clusters based on specific characteristics.

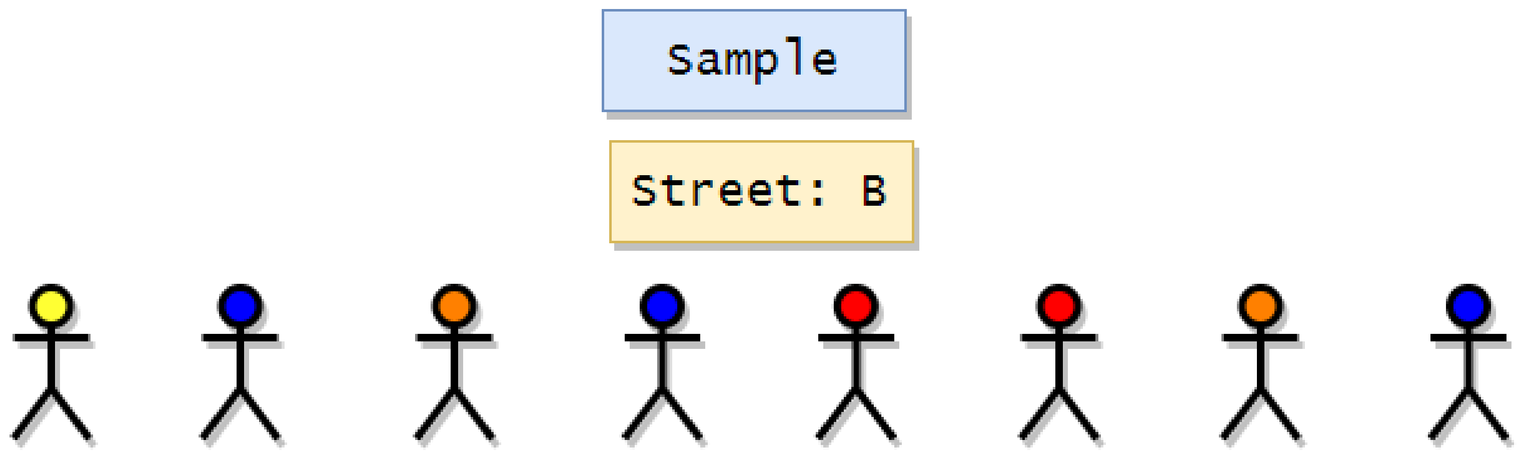
Let’s take an example to understand it better. Suppose we want to get information on the weights of people in a particular town. Now consider that town has four streets. So, in this case, our population will be the town or the four streets. Here we can say that we have four subgroups or clusters based on the streets. Now we will choose one cluster from the 4 clusters to gather the required information. So in this method, we know that our population is already divided into four streets. So here, what we can do is we can apply random sampling here to choose one of the four streets from the town, and then we can get information from that street. So we can say that this method runs on top of Simple Random Sampling. Suppose in a town there are four streets.

- Street A
- Street B
- Street C
- Street D

Now we just have to apply Simple Random Sampling to choose 1 out of 4 streets. Here is the visual representation of the technique:



Apply simple random sampling to choose the four streets. Suppose we get street B in our random sampling. Then we can take the necessary information for all people from street B that will represent the entire population.



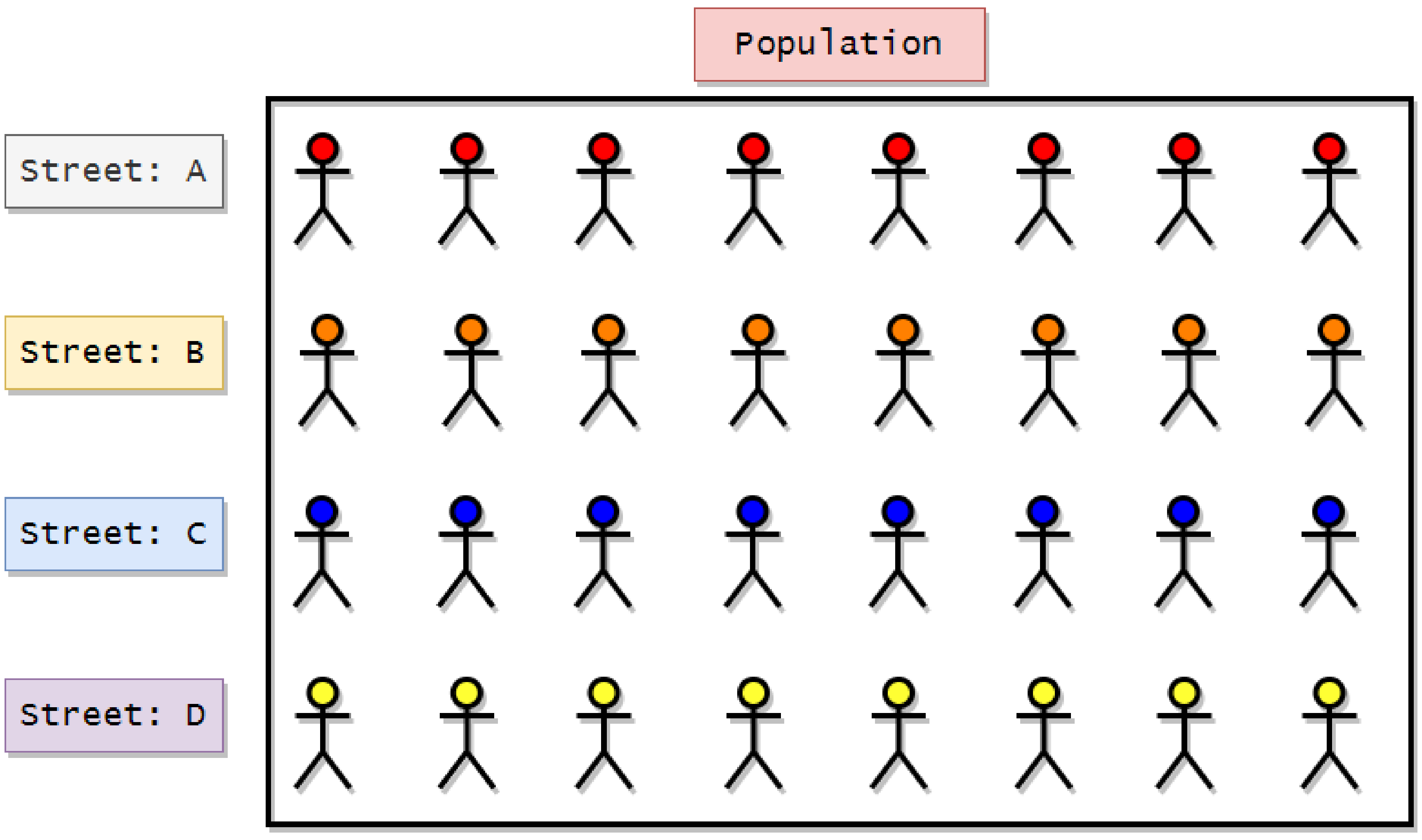
Advantages:

1. If we have a large geographical area, this technique helps a lot because it is easier to get many data from a single cluster than to get a little data from many clusters.

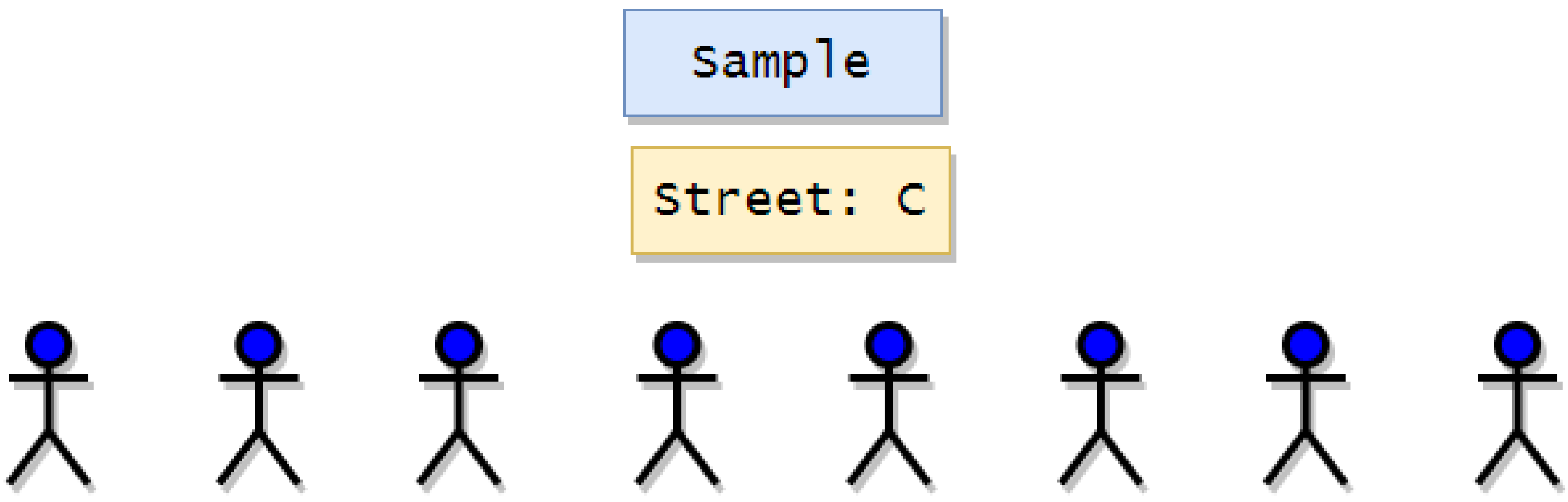
Disadvantages:

1. Increased risk of bias.

For example, in the above example, if people in each street are clustered with specific characteristics, this method will give us biased results. If their zodiac signs or race naturally clusters people in this town, this technique will give us biased results.



If we apply Simple Random Sampling and if we get street C, then here we can see that the result will be biased and will only represent blue characters.



Non-Probability Sampling Techniques:

In these techniques, the samples or subjects are collected with no specific probability structure in mind. Here we can say that the sampling is not entirely random. The samples are not indeed the representation of the entire population.

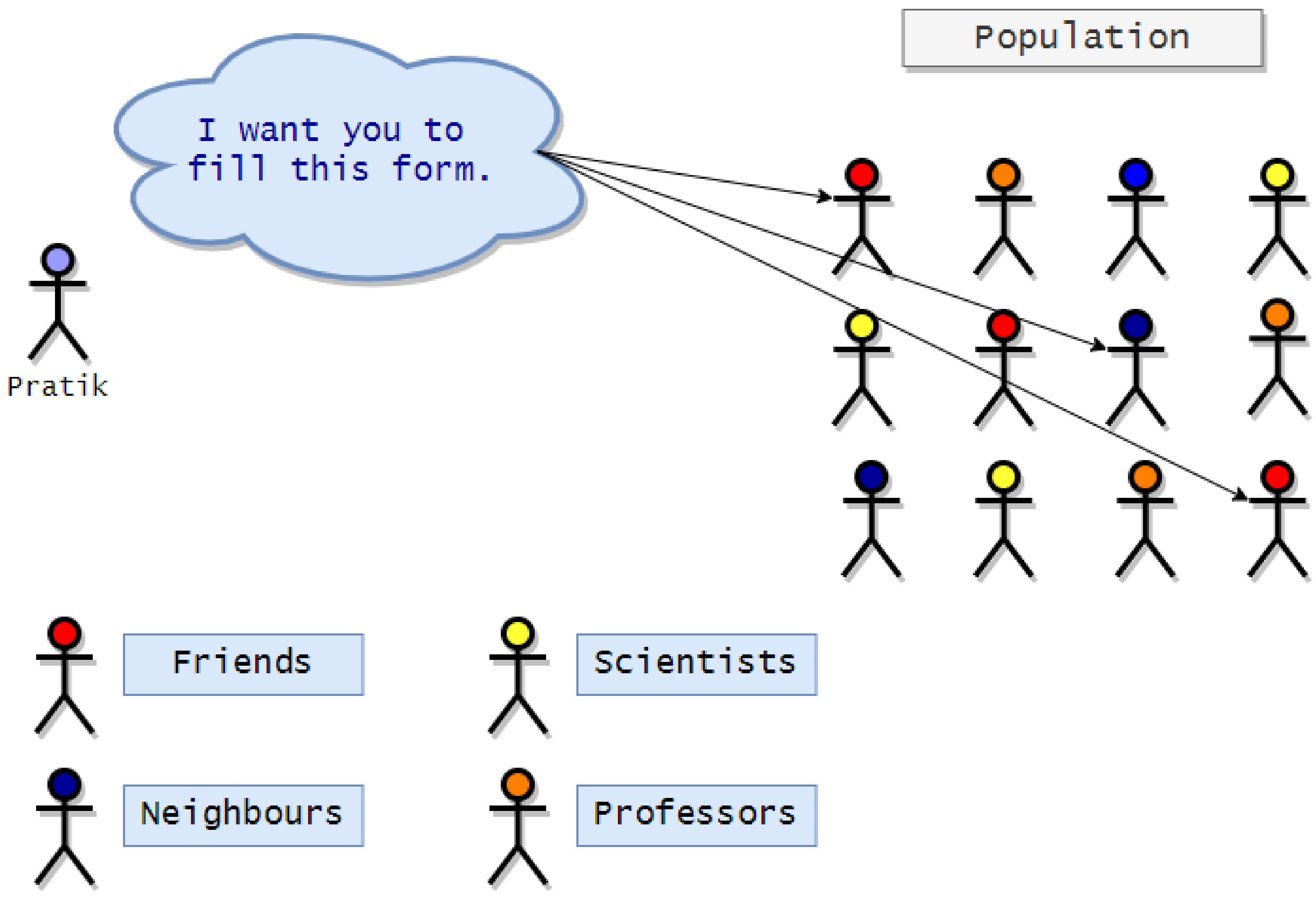
Convenience Sampling

Convenience sampling is probably the easiest sampling method because the samples or the participants are selected based on their availability and willingness to participate. In this method, the results are prone to significant bias, and it may not represent the views of the entire population. One more thing to notice here is that the sample may not represent characteristics like sex or age. It is generally used for preliminary research.

For example:

1. Surveying friends.
2. Surveying people at the mall.
3. Online polls.

Visual representation of convenience sampling:



In the representation above, we can see that the researcher (Pratik) chooses his friends and neighbors as subjects for his research as they are easy to convince or reach.

Advantages:

- It is easy and can be generated quickly.

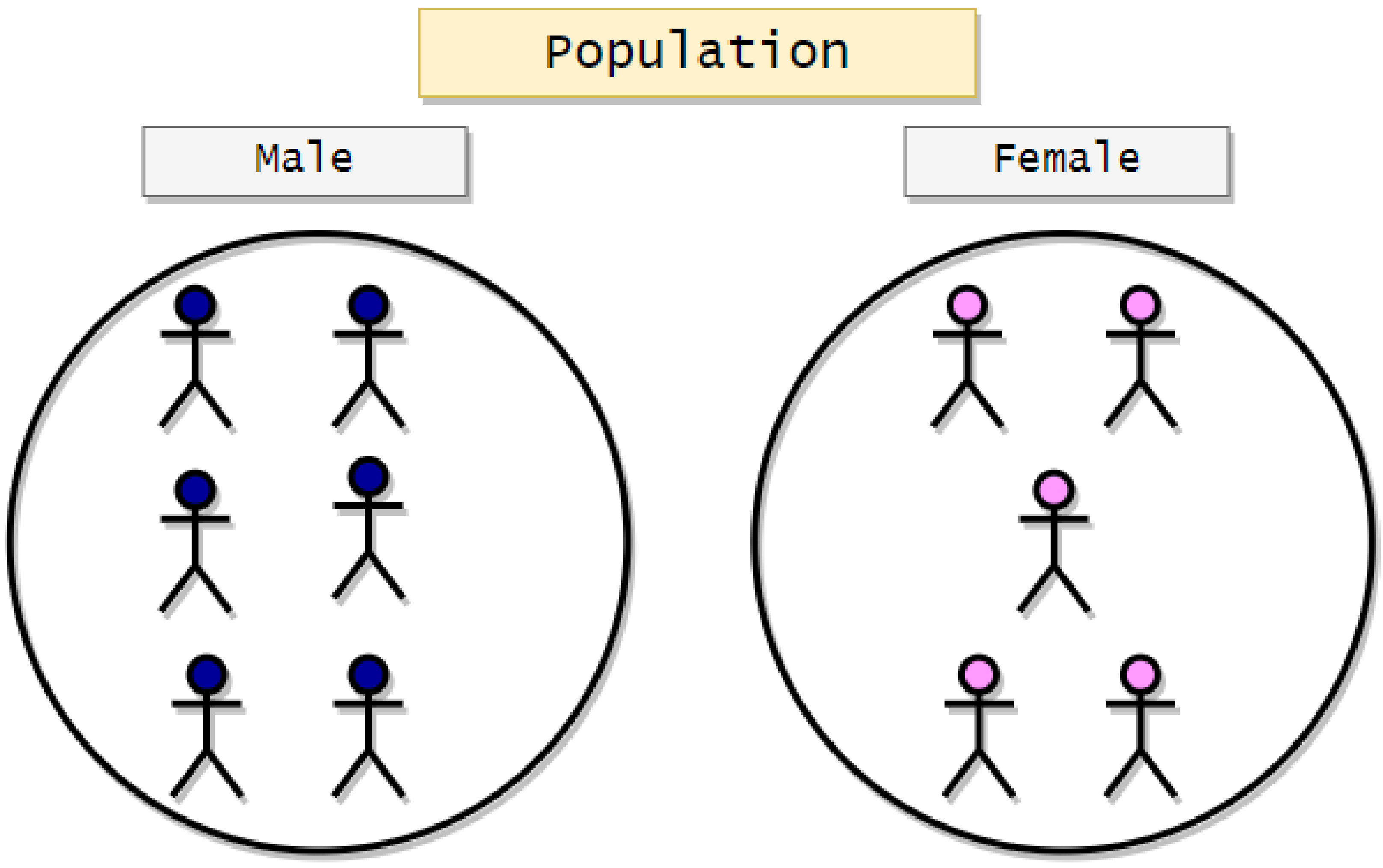
Disadvantages:

- It is generally a poor representation of the entire population.

Quota Sampling

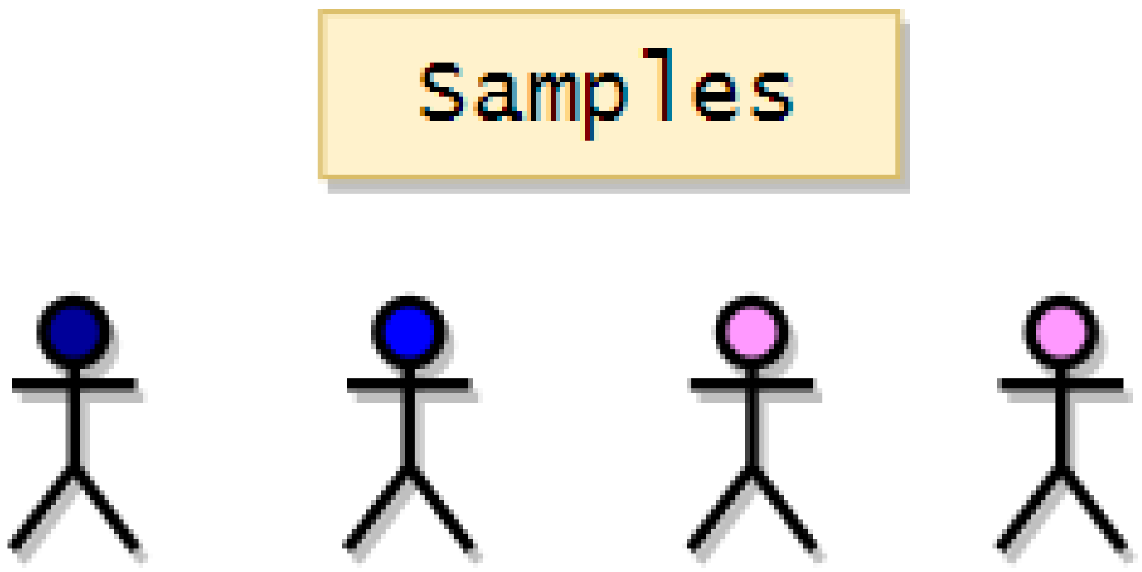
This method is a non-probabilistic approach to stratified sampling. Here we have to divide the whole population into clusters based on some characteristics. After dividing the population into clusters, we can take any person from the groups at our convenience. The subjects will not be chosen randomly. Here we can also initially fix a quota value to choose the desired number of subjects from each group. For example, let's say we want to select some candidates for an interview, and we have to select 10 males and 10 females from the population. So what we can do here is we can cluster the population-based on sex, and then we can choose any 10 male and 10 female at our convenience. There is no randomization in choosing the subjects after clustering.

The following is a visualization of the sampling technique discussed:



Samples:

Here we are going to choose two candidates from both groups.



Advantages:

- It is straightforward to administer.
- It is fast and inexpensive.
- We can take into account population proportions if required.

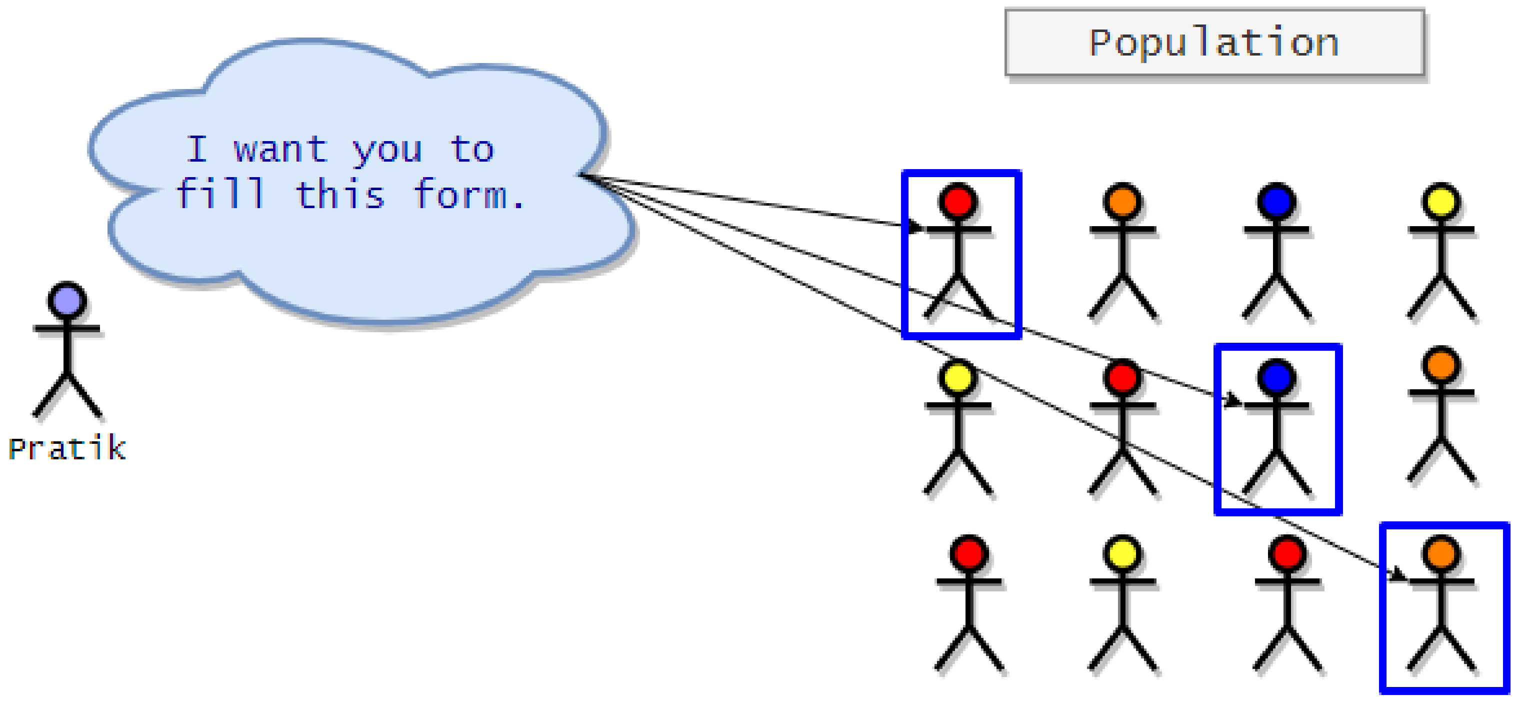
Disadvantages:

- Selection is not random.

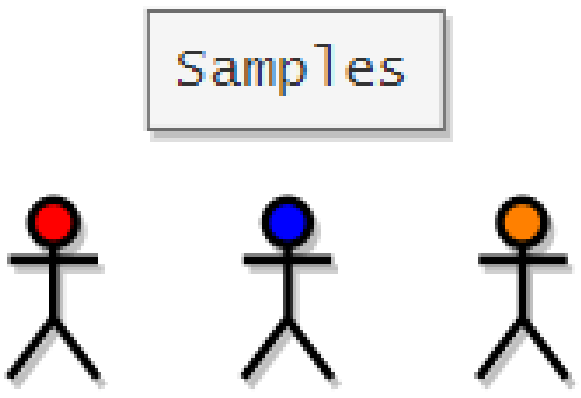
Judgemental Sampling

In this method, the researcher chooses his subjects or participants based on his judgment or guts. The researcher can specifically choose a group of people with specific characteristics. The researcher may only choose female participants. So we can say that this method can be biased based on the researcher's judgment. Using this method, the researchers can choose only those subjects that he/she deems to be a perfect fit for his/her research. It may not represent the opinion of the entire population.

Here is a visualization of the technique discussed.



Samples:



Advantages:

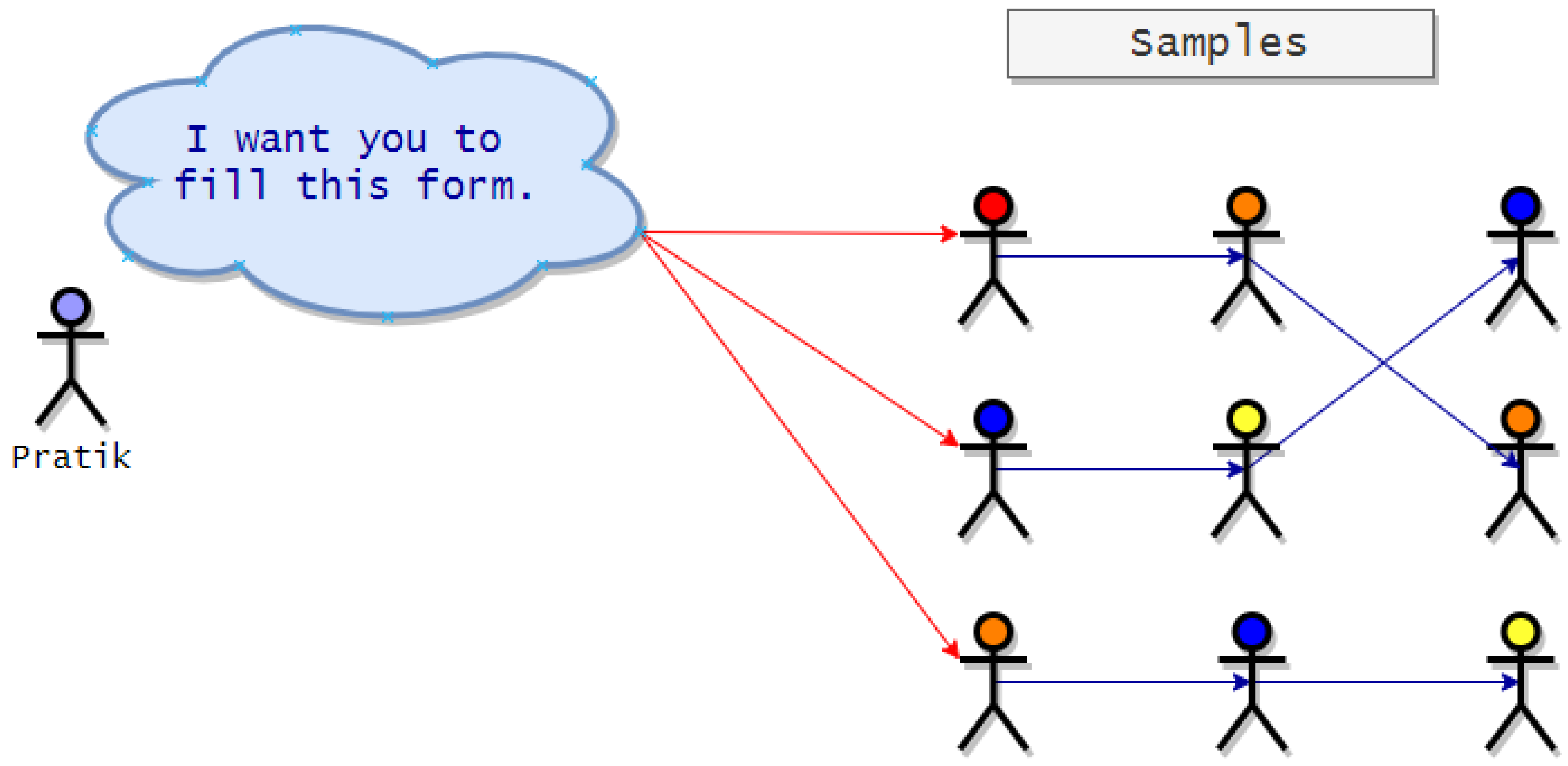
- It's time and cost-effective.

Disadvantages:

- The results can be biased.

Snowball Sampling

It is a non-probabilistic technique where existing subjects are asked to nominate other subjects best known to them. Here the sample size will grow like a rolling snowball. That is why it is called snowball sampling. Here the existing subjects will recruit other subjects, and the cycle will go on. It's generally used in social-science surveys where subjects with particular characteristics are hard to find. In the following visualization, we can see that initially, Pratik has three subjects, and then the subjects recruit others and so on.

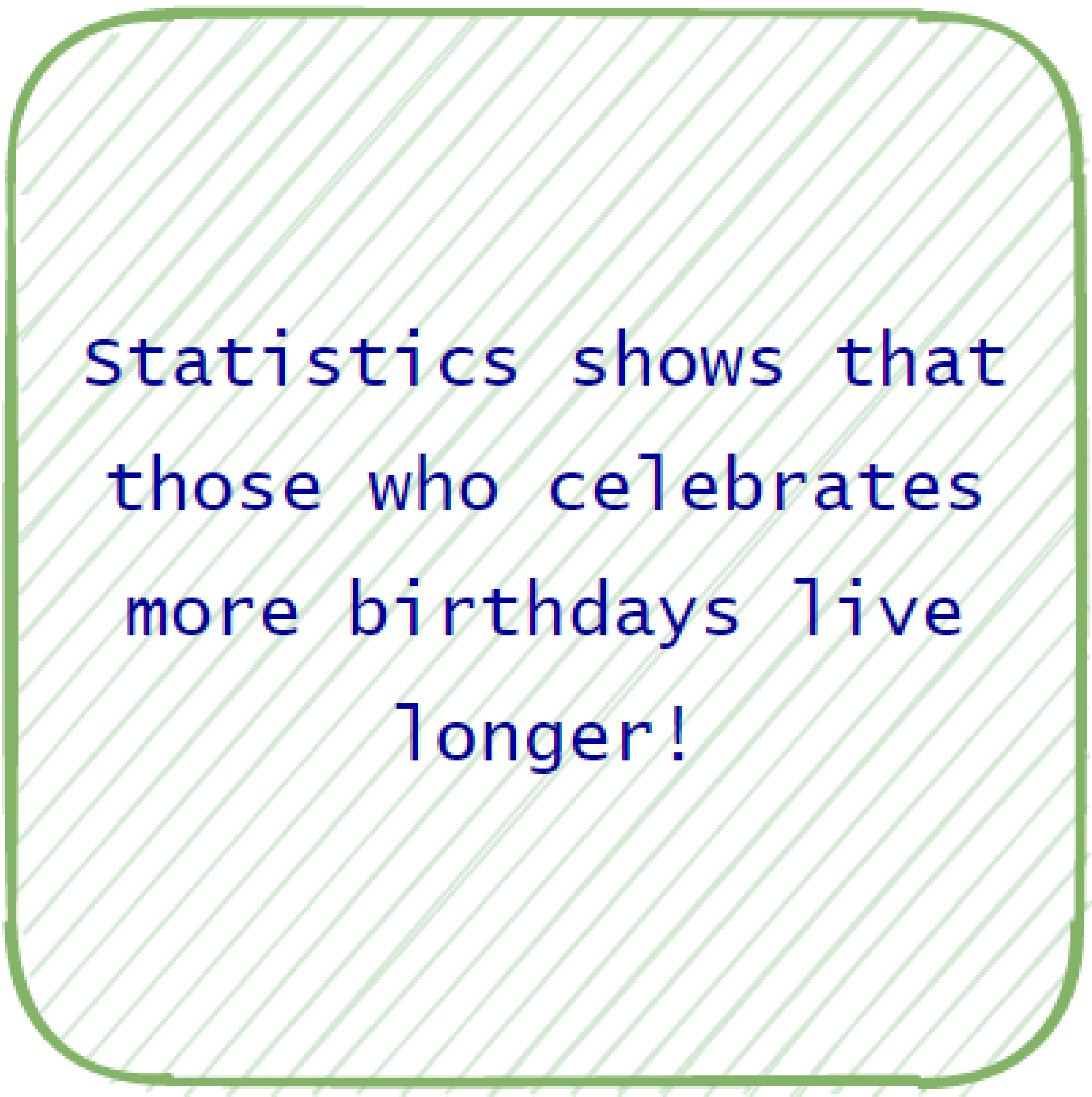


Advantages:

- It is instrumental when the samples are hard to find.
- Low cost.
- Very relevant samples to our study.

Disadvantages:

- This method depends on subjects recruiting other subjects, so there is a high chance of selection bias.
- It only works if subjects have other relevant connections.



Statistics shows that those who celebrates more birthdays live longer!

What is statistics?

According to Wikipedia, statistics is a discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. We can also say that statistics is a science of collecting and analyzing numerical data in large quantities. We can also say that statistics is a matter of science and logic.

Importance of statistics in Data Science and Machine Learning

Data science is a scientific discipline that is dominated by programming and mathematics. Every various data corporation in the world generates tonnes of data that can be further analyzed and visualized by experts to understand the trends. Data visualization can only be performed if the data is clear and understandable. However, the data generated by organizations are too messy to handle. So we can say that finding structures and basic patterns in data is an essential task for data science. Statistics provides the methods and tools to find the hidden structures and patterns in the data so that experts can make predictions from them. Statistics is the fundamental step in the Data Science world. We can say we need statistics to transform observations into information. In machine learning, we use various kinds of algorithms for prediction, classification, and clustering. However, there are many useful libraries available for use that will perform mathematical calculations for us. Nevertheless, it is vital to know the math behind each of the algorithms because it provides insights into what we are doing and why?

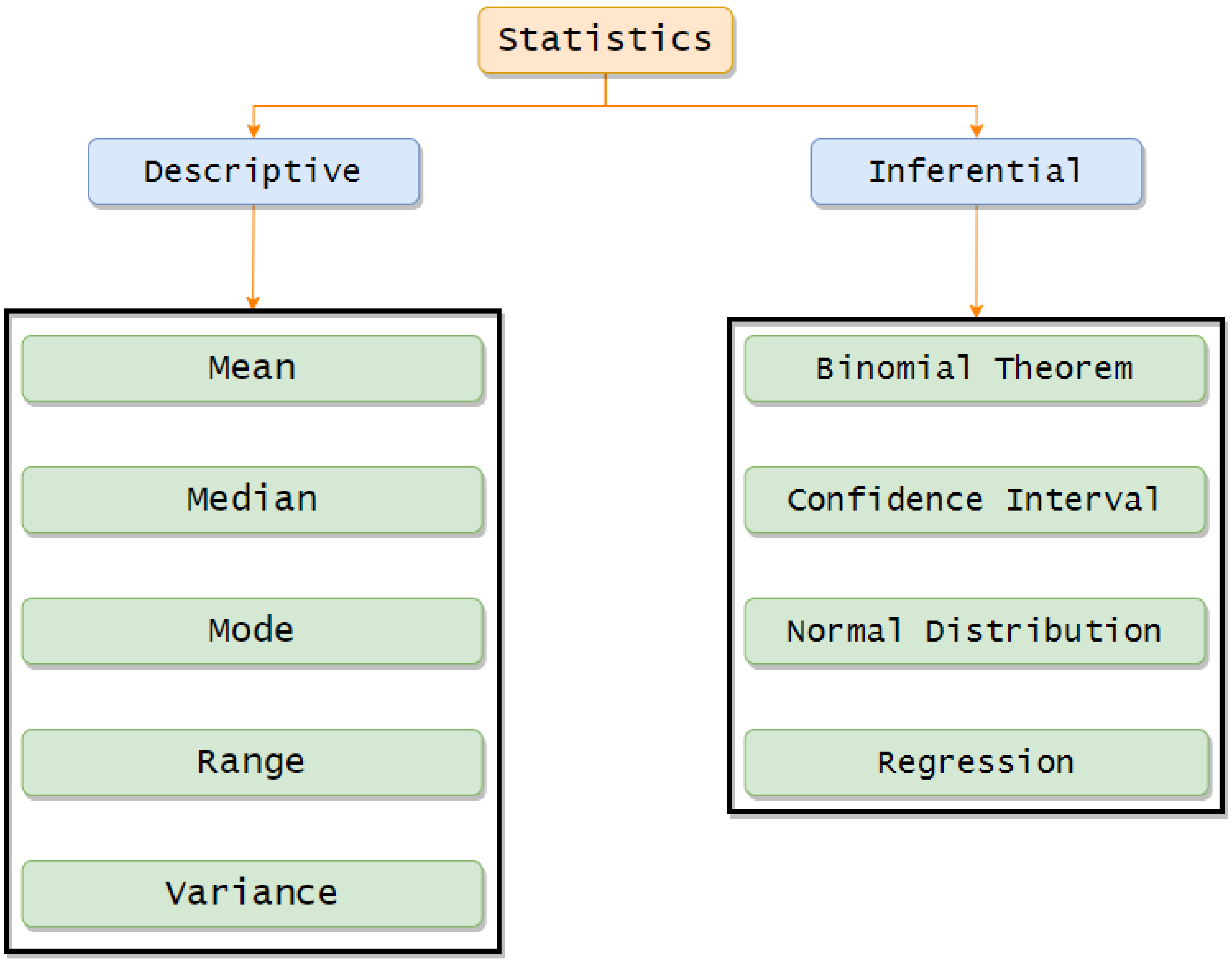
Instead of applying cool-sounding machine learning algorithms to our data to make predictions, it is imperative to understand the pattern to know the distribution of the data. Now how will the distribution of data help us? After knowing the distribution of data, we can look at the limitations of machine learning algorithms and apply them to give us the best results. In our projects, we use a part of data to train the algorithm and make predictions from it. To train the model for our algorithm, we generally use the Python programming language.
Moreover, we are well aware that python is relatively slower than other programming languages. However, the simple

syntax and well-developed libraries give programmers a reason to lean towards it. In the real world, the data will be in large quantities, so here we cannot take the risk of training our model based on an algorithm that provides no useful insights. That is why it is essential to understand the distribution of data. In future work, we will show the different types of distributions of data.

Types of Statistics:

Statistics can be divided into two main categories.

- Descriptive Statistics.
- Inferential Statistics.



Descriptive Statistics:

Descriptive statistics fundamentally works on organizing and summarizing data using graphs. We can summarize the data and visualize it using Bar graphs, Histograms, and Pie charts.

We can also view the shape and skewness of the graphs. Descriptive statistics include measures to find central tendency values like mean, median, and mode. Other than that, we can also find the measure of variability or the spread of data with the range, variance, and standard deviation values.

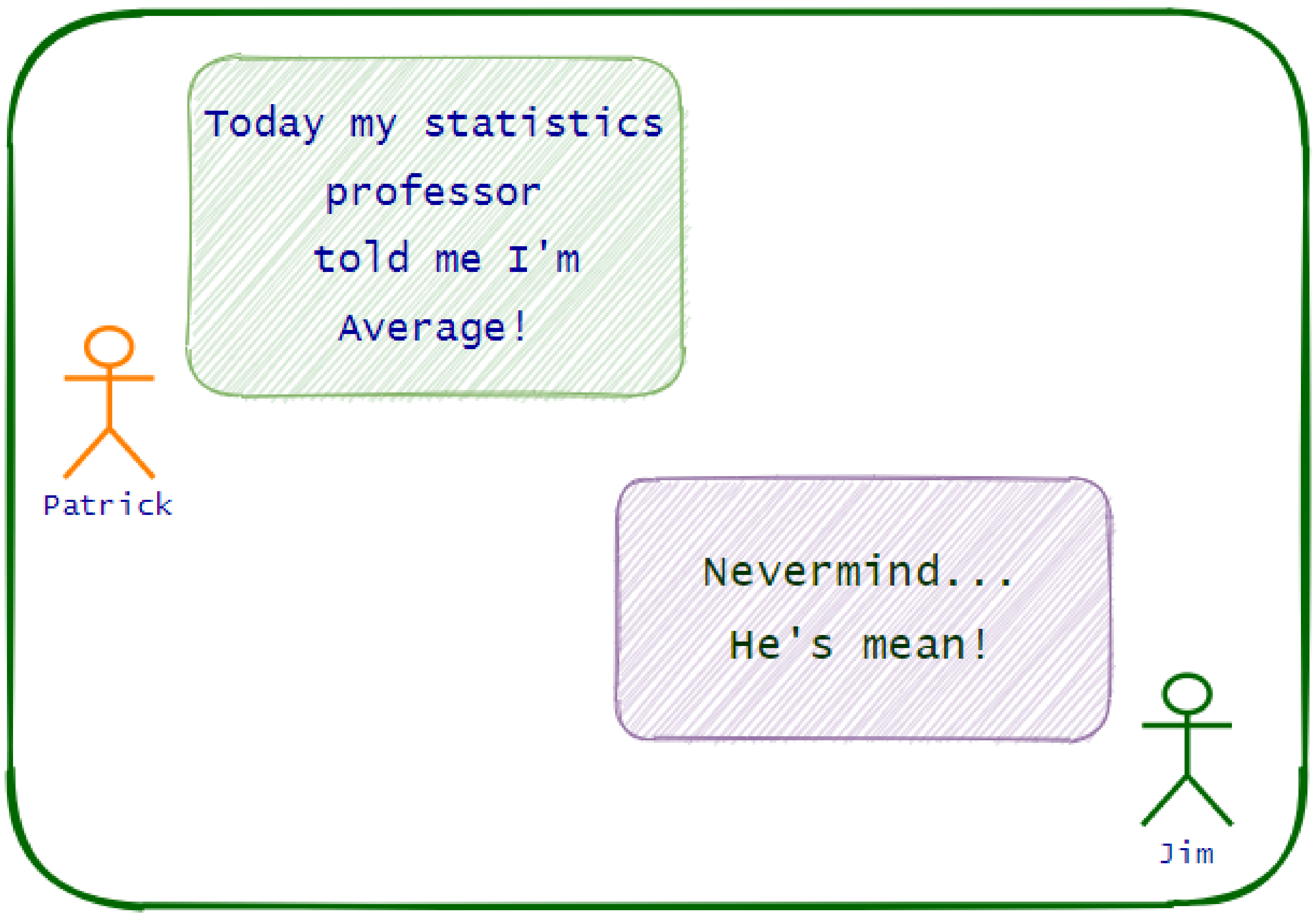
Inferential Statistics:

In inferential statistics, we use the sample data to make an inference or draw the population's conclusion. It uses probability to find out the confidence of the predictions we make.

In this book, we will mainly focus on Descriptive Statistics.

Measure of central tendency:

Central tendency refers to an idea that suggests one number that best summarizes our whole dataset. It may also be called the center of the distribution.



Arithmetic Mean:

Mean can be referred to as a central tendency of the data. It is a single number around which our data is spread around. In short, we can say that it is a single number that best represents the whole dataset.

The mean or average of a dataset is found by adding all the numbers in it and then dividing the sum by the dataset's length.

Formula of the mean for population:

$$Mean(Population) = \mu = \frac{\sum x}{N}$$

Formula of the mean for sample:

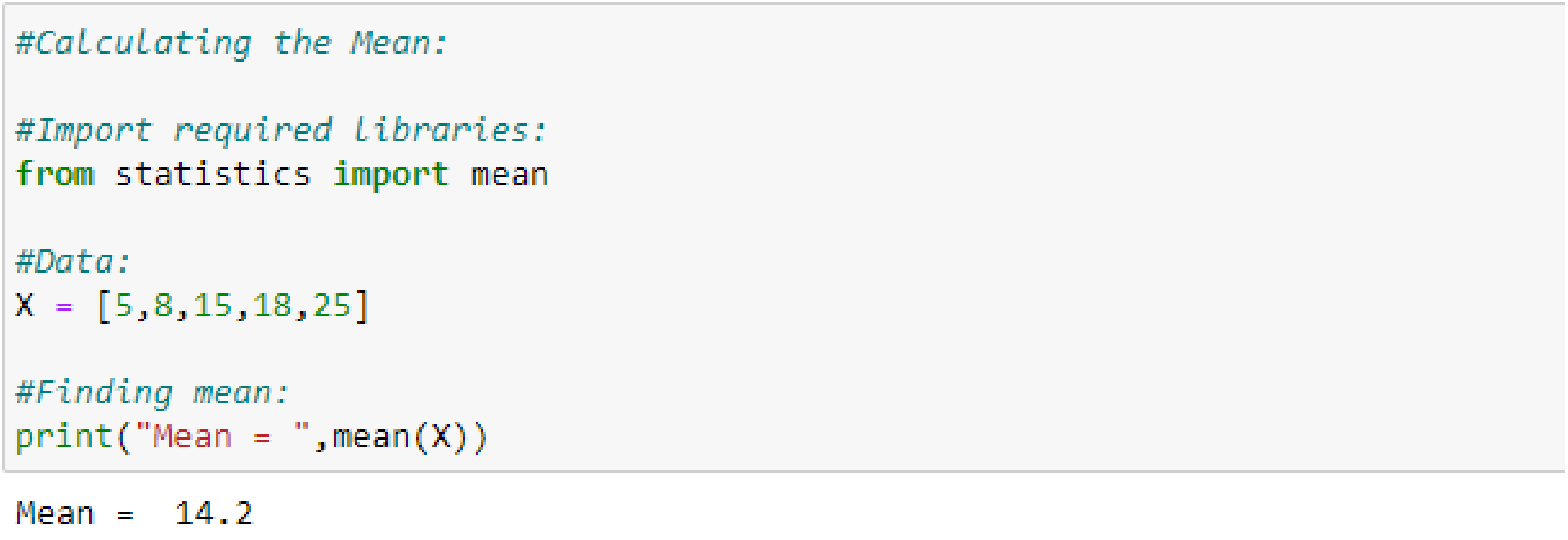
$$Mean(Sample) = \bar{x} = \frac{\sum x}{n}$$

For example:

Find mean of the following dataset: [5,8,15,18,25]

$$\bar{x} = \frac{(5 + 8 + 15 + 18 + 25)}{5} = \frac{71}{5} = 14.2$$

Python implementation:



```
#Calculating the Mean:

#Import required libraries:
from statistics import mean

#Data:
X = [5,8,15,18,25]

#Finding mean:
print("Mean = ",mean(X))
```

```
Mean =  14.2
```

Weighted Mean:

When we have the same numbers many times in some data, then to find the mean instead of simply adding it up and then dividing by its length, we will find the weighted frequency for each number so that our process will become faster.

$$\overline{x}_w = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + \ldots\ldots + w_n \cdot x_n}{w_1 + w_2 + w_3 + \ldots\ldots + w_n}$$

For example:

Find the weighted mean for the following data:

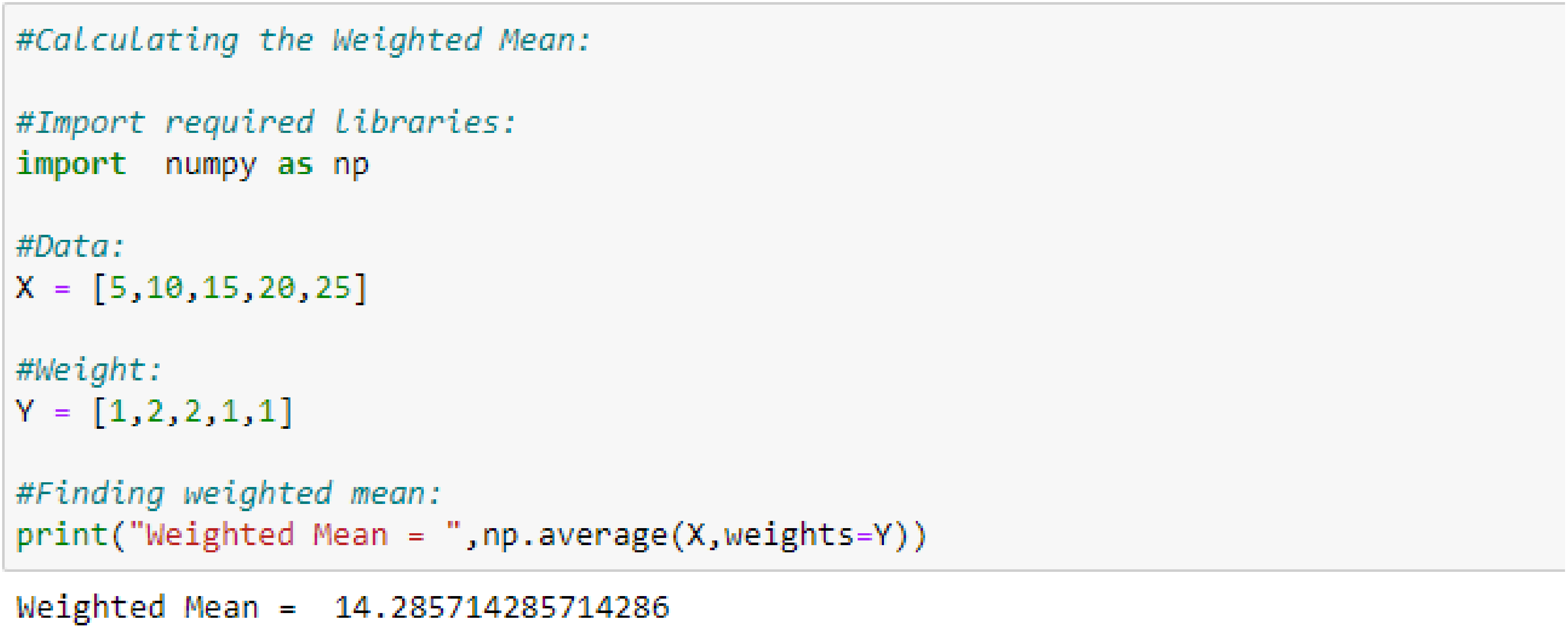
X	W
5	1
10	2
15	2
20	1
25	1

$$\overline{X_w} = \frac{5*1+10*2+15*2+20*1+25*1}{1+2+2+1+1} = 14.2$$

Launch the Python implementations in Google Colab:

https://news.towardsai.net/descriptive-statistics-colab

Python implementation:



```
#Calculating the Weighted Mean:

#Import required libraries:
import  numpy as np

#Data:
X = [5,10,15,20,25]

#Weight:
Y = [1,2,2,1,1]

#Finding weighted mean:
print("Weighted Mean = ",np.average(X,weights=Y))
```

```
Weighted Mean =  14.285714285714286
```

Mean for Categorical Dataset:

Let's say we went to a pet show that only has dogs and cats. As we move forward, we are noting whether the pet is a dog or cat. The following is the final observation or data.

[Dog, Cat, Cat, Dog, Cat, Cat, Dog, Cat, Cat, Dog]

Now we want to find the mean of this categorical dataset. To do that, we have to convert the categorical dataset into a numerical dataset. Here we are denoting cat as 0 and Dog as 1. Hence, here is our numerical dataset.

[1,0,0,1,0,0,1,0,0,1]

Now we can apply our regular mean formula:

$$\bar{x} = \frac{1+0+0+1+0+0+1+0+0+1}{10} = 0.4$$

Now here we can see that the mean is centered towards 0. So here we can say that in our dataset the number of cats is higher than the number of dogs.

Launch the Python implementations in Google Colab:

https://news.towardsai.net/descriptive-statistics-colab

Geometric Mean:

The geometric mean is the "nth" root when we multiply n numbers.

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

In a simplified way:

$$\bar{x}_{geom} = \sqrt[n]{x_1 * x_2 * x_3 * \dots * x_n}$$

For example:

Find the geometric mean of [3, 4, 5, 6, 7].

$$\bar{x}_{geom} = \sqrt[5]{3 * 4 * 5 * 6 * 7} = 4.7$$

Use-case:

It is generally used when we are trying to calculate the average growth rate where the growth is determined by multiplication, not addition.

For example:

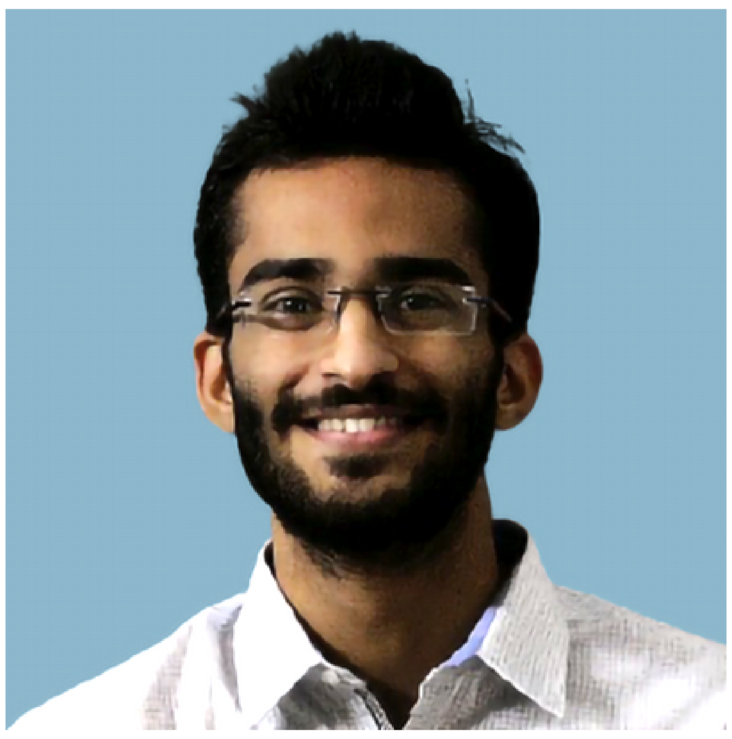
"My Apple stock gained 20% in year 1, 30% in year 2, and 40% in year 3."

Then what is the average yearly rate of return?

Launch the Python implementations in Google Colab:

https://news.towardsai.net/descriptive-statistics-colab

About the Authors:



Pratik Shukla is a machine learning engineer with Towards AI. He is pursuing his master's degree in computer science in the US starting in 2021. His current goals are motivated by the purpose of learning something new and remarkable every day. His research interests lie in machine learning and its applications, especially in astronomy and astrophysics. Previously, he received his B.Tech. from Gujarat Technological University, and his work has been featured in many places, from KDNuggets, Nightingale, and others.

Roberto Iriondo is the founder of Towards AI, a globally recognized publication and software company, and a front-end engineer at Carnegie Mellon University. As a builder and strategist by heart, his work has helped several companies to achieve their business goals and needs, from Anyscale, Superb AI, Determined AI, Lambda, Udacity, and many others.

DISCLAIMER: The views expressed in this book are those of the author(s) and do not represent the views of any company (directly or indirectly) associated with the author(s). This book does not intend to be a final product, yet rather a reflection of current thinking along with being a catalyst for discussion and improvement.