Sampling Theory
Théorie de l’échantillonnage

Advanced Statistical Analysis

Feryal WINDAL

Definitions

L'inférence L'inférence est le processus par lequel on tire des conclusions à partir de propositions supposées vraies, en suivant des principes logiques établis.

Exemple : Propositions présumées vraies (principes établis) :

- Tous les hommes sont mortels.
- Socrate est un homme.

Conclusion déduite :
Socrate est mortel.
Dans cet exemple, à partir des deux propositions données
comme vraies, on infère logiquement que Socrate est mortel.

Inference is the process of drawing conclusions from presumed true propositions, based on established logical principles.

Example Presumed true propositions (established principles):

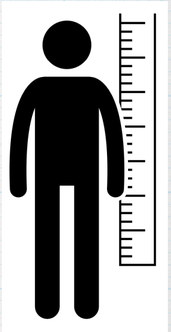
- All men are mortal.
- Socrates is a man.

Deduced conclusion:
Socrates is mortal.
In this example, based on the two given propositions considered true, we logically infer that Socrates is mortal.

Definitions

L'inférence statistique est définie comme le processus d'utilisation des données d'un échantillon pour estimer ou tester des hypothèses sur les caractéristiques numériques (paramètres) d'une population

Exemple Vous souhaitez déterminer la taille moyenne des habitants d'une ville. Plutôt que de mesurer chaque personne, vous prenez un échantillon aléatoire de 100 individus et mesurez leur taille. Les données recueillies montrent une taille moyenne de 173 cm. Grâce à l'inférence statistique, vous pourriez estimer que la taille moyenne de l'ensemble des habitants est proche de cette valeur, sans avoir mesuré tout le monde.



Statistical inference involves using sample data to make educated guesses or test hypotheses about the characteristics (parameters) of a larger population.

Example Suppose you want to determine the average height of adults in a city. Instead of measuring everyone, you take a random sample of 100 people and measure their heights. The sample reveals an average height of 173 cm. Using statistical inference, you estimate that the average height of the entire population is close to this value, without having to measure every individual.

Definitions

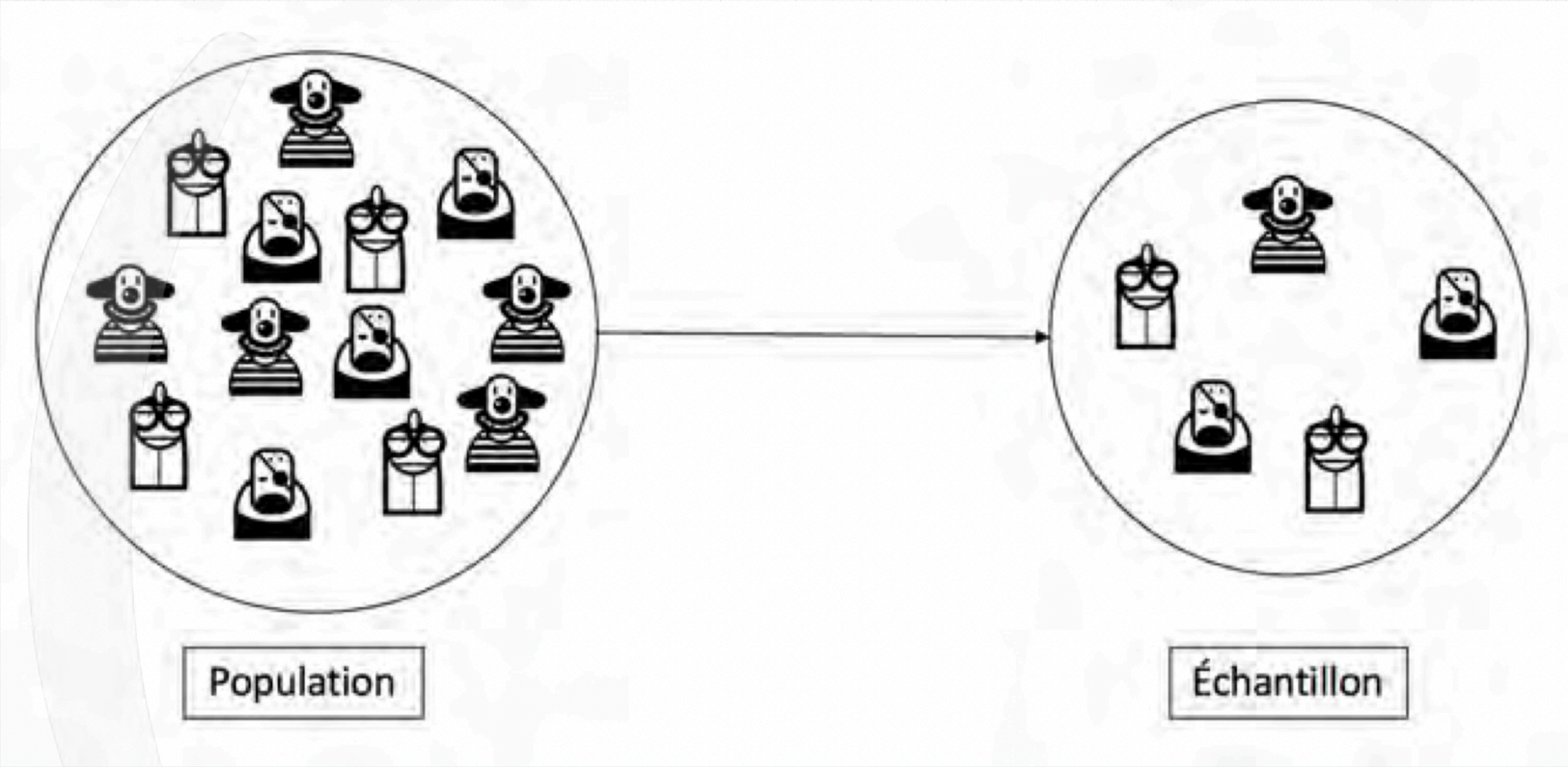
Une populations est définie comme l'ensemble de tous les éléments d'intérêt dans une étude particulière.

Un échantillon est défini comme un sous-ensemble de la population.

A population is defined as the set of all items of interest in a particular study.

A sample is defined as a subset of the population.

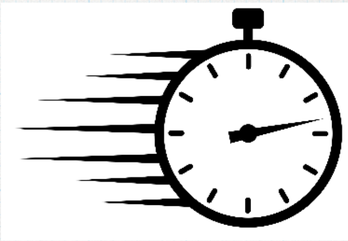
Illustration



Random Sample

Pourquoi recourir à un prélèvement d’échantillon dans une population ?

- Impossible de connaître toute la population.
- Collecter et traiter toutes les observations serait trop long.
- Le coût de collecte d’informations sur toute la population est prohibitif.



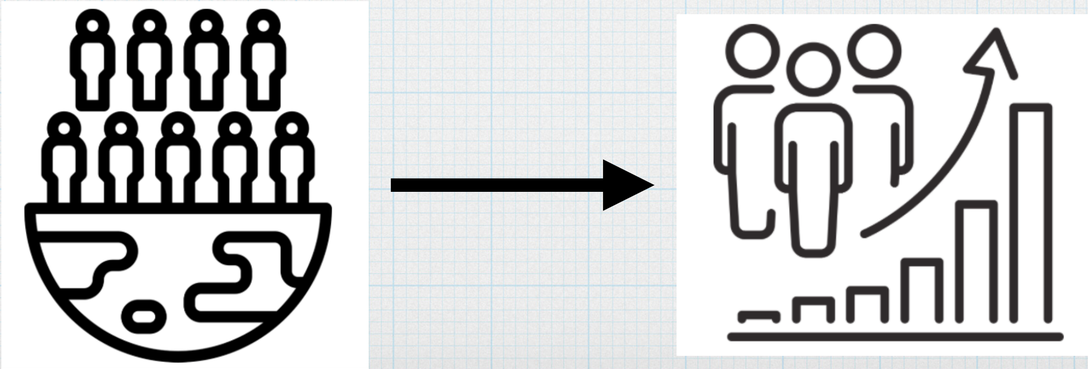
Why Use Random Sampling in a Population?

- It is impossible to know the entire population.
- Collecting and processing all observations takes too much time.
- The cost of collecting information on the entire population is prohibitive.

Random Sample

- Dans ce type de situation, l'échantillonnage permet de sélectionner une partie de la population qui reflète le plus fidèlement possible ses caractéristiques.
- Les résultats obtenus à partir de l'échantillon doivent être aussi proches que ceux que l'on aurait obtenus avec l'ensemble de la population.

- In such situations, sampling is used to select a part of the population that best represents its characteristics.
- The results from the sample should closely approximate what would have been obtained from the entire population.



Non-Probabilistic Sampling

On distingue deux grandes familles de méthodes de prélèvement d'un échantillon :

Les méthodes non-probabilistes :

- Dans les méthodes non probabilistes, le statisticien ou le responsable de l'échantillonnage sélectionne directement les individus.
- Cet échantillon est donc dit "de convenance", ou encore "subjectif".

Ex: Junia students sample

very useful in exploratory studies where speed and ease of data collection are crucial

There are two main families of methods for taking a sample:

Non-probabilistic methods:

- In non-probabilistic methods, the statistician or the sampling manager directly selects the individuals.
- This type of sample is called a "convenience sample," or "subjective" sampling.

Random Sample : probabilistic

On distingue deux grandes familles de méthodes de prélèvement d'un échantillon :

Les méthodes probabilistes :

- L'échantillon est sélectionné de manière aléatoire, ce qui signifie que chaque élément de la population a une chance égale d'être inclus.
- On parle d'échantillonnage aléatoire, parfois désigné par le terme "randomisé".

Ex: random selection

More reliable because they minimize bias and allow results to be generalized to the entire population.

There are two main families of methods for taking a sample:

Probabilistic methods:

- The sample is selected randomly, meaning every member of the population has an equal chance of being included.
- This is referred to as random sampling, or sometimes "randomized sampling"

	Non probabiliste	Probabiliste
Exemple	Demander seulement aux étudiants ayant une moyenne supérieure à 12 si le cours est intéressant.	Demander l'avis des étudiants ayant un numéro d'identification impair
Inconvénient majeur	Biais de sélection très important	Requiert des plans de sondage plus au moins sophistiqués

	Non Probabilistic	Probabilistic
Example	Ask only students with an average greater than 12 if the course is interesting.	Ask the opinion of students whose identification number is odd.
Major drawback	Very significant selection bias	Requires more or less sophisticated sampling plans.

Échantillonnage avec équiprobabilité de tirage :

- Chaque individu de la population a la même probabilité d'être inclus dans l'échantillon (équiprobabilité de tirage)
- Admettons que l'on veuille tirer un échantillon de taille n dans une population de taille N (avec $n < N$). Le Nombre d'échantillons différents que l'on peut former est :

$$C_N^n = \frac{N!}{n!(N-n)!}$$

Il s'agit d'une combinaison, l'ordre dans lequel les individus sont tirés n'importe pas.

Sampling with drawing equiprobability :

- Each individual in the population has the same probability of being included in the sample (draw equiprobability)
- Suppose that, we want to draw a sample of size n, from a population of size N (with $n < N$). The Number of different samples that can be formed is : $C_N^n = \frac{N!}{n!(N-n)!}$

It is a combination, it does not matter which order the individuals are drawn.

Dans un sondage aléatoire, le taux de sondage, c'est-à-dire la proportion de la population incluse dans l'échantillon, est donné par :

$$\frac{n}{N} = \tau$$

La probabilité pour qu'un individu i soit inclus dans l'échantillon est :

$$\pi_i = \frac{n}{N} = \tau$$

La probabilité qu'un échantillon spécifique soit sélectionné est : $\frac{1}{C_N^n}$

In a random survey, the sampling rate, i.e., the proportion of the population included in the sample, is given by:

$$\frac{n}{N} = \tau$$

The probability for an individual i to be included in the sample is:

$$\pi_i = \frac{n}{N} = \tau$$

The probability that a particular sample will be drawn is equal to: $\frac{1}{C_N^n}$

Random Sample

Il existe plusieurs types d’échantillonnage aléatoire, dont : :

- échantillonnage aléatoire simple (EAS)
- échantillonnage aléatoire stratifié
- échantillonnage aléatoire par grappe

There are different types of random sampling, such as:

- simple random sampling (EAS)
- stratified random sampling
- random sampling by cluster

Simple random sample (SRS)

Simple random sample (SRS)

L'échantillonnage aléatoire simple est facile à mettre en œuvre : il n'est pas nécessaire de connaître *au préalable* certaines caractéristiques de la population (comme les zones géographiques ou les catégories socio-professionnelles)

Il est toutefois nécessaire de connaître l'existence des individus de la population afin de pouvoir les inclure dans l'échantillon.

Simple random sampling is easy to implement: there is no need to know specific characteristics of the population *beforehand* (such as geographic areas or socio-professional categories).

However, prior knowledge of the existence of individuals in the population is necessary to include them in the sample.

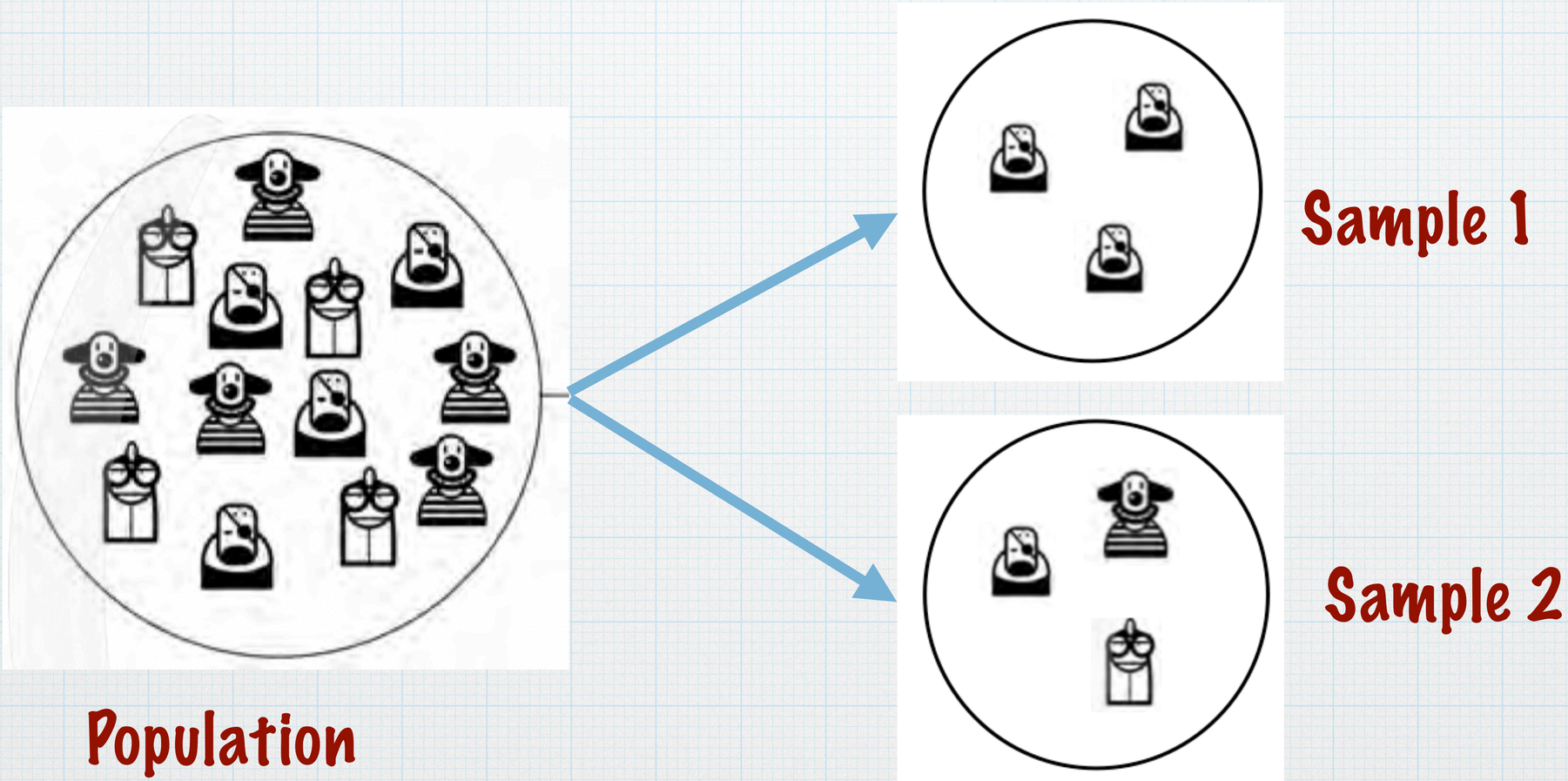
Simple random sample (SRS)

Si l'on dispose d'informations *préalables* sur la population, permettant de constituer un échantillon plus représentatif, l'échantillonnage aléatoire simple ne les prend pas en compte (chaque individu a la même probabilité d'être inclus dans l'échantillon).

If we have *prior* information about the population that would help create a more representative sample, simple random sampling does not take this into consideration (each individual has an equal probability of being included in the sample).

Engineering student (60%) Library Literature Student (40%)

Dans l'exemple ci-dessous, l'échantillon 2 est représentatif car il reflète la diversité de la population, tandis que l'échantillon 1 ne l'est pas



In the example below, Sample 2 is representative because it reflects the diversity of the population, while Sample 1 does not.

SRS and recovery

Si des informations auxiliaires n'ont pas été prises en compte lors de la sélection de l'échantillon (par oubli ou parce qu'elles n'étaient pas disponibles), et si elles permettent d'expliquer la variable d'intérêt, un redressement peut être effectué

Exemple d'information auxiliaire :

- genre, âge, catégories socio-professionnelles (pour des individus) ; taille en nombre d'employés (pour des entreprises)

Ce redressement permet d'aligner la structure de l'échantillon sur celle de la population.

If a priori information is available on the population that would allow for a more representative sample, simple random sampling does not consider this.

Example of auxiliary information:

- gender, age, socio-professional categories (for individuals); size in number of employees (for companies)

This adjustment aligns the sample structure to match that of the population.

SRS and recovery: rating

La proportion de ***la strate*** i dans la taille totale de l'échantillon

The proportion of ***stratum*** i in the total sample size

$$f_i = \frac{n_i}{\sum_i n_i} = \frac{n_i}{n}$$

La part de l'effectif de *la strate* i dans l'effectif total de la population

share of the number of stratum i in the total number of the population

$$F_i = \frac{N_i}{\sum_i N_i} = \frac{N_i}{N}$$

Le taux de sondage de la strate i

The stratum sampling rate i

$$s_i = \frac{n_i}{N_i}$$

Le taux de sondage global

The global sampling rate

$$s = \frac{n}{N}$$

SRS and recovery: rating

Un redressement possible à effectuer pour f_i est :

A possible adjustment to be made for f_i is :

$$f_i^* = \frac{f_i}{\frac{s_i}{s}} = \frac{f_i}{\frac{n_i/N}{n/N}} = f_i \times \frac{N_i/N}{n_i/n}$$

$$= f_i \times \frac{F_i}{f_i} = F_i$$

SRS and recovery: Example

Population: 60% women (180 women), 40% men (120 men).
Sample obtained: 170 women, 130 men.

Adjustment Formulas:

1. **Stratum share in the sample:**

$$f_{\textbf{women}} = \frac{n_{\textbf{women}}}{n} = \frac{170}{300} = 0{,}567$$

$$f_{\textbf{men}} = \frac{n_{\textbf{men}}}{n} = \frac{130}{300} = 0{,}433$$

2. **Stratum share in the population:**

$$F_{\textbf{women}} = \frac{N_{\textbf{women}}}{N} = \frac{180}{300} = 0{,}60$$

$$F_{\textbf{men}} = \frac{N_{\textbf{men}}}{N} = \frac{120}{300} = 0{,}40$$

SRS and recovery: Example

Population: 60% women (180 women), 40% men (120 men).
Sample obtained: 170 women, 130 men.

Adjustment Formulas:

3. **Sampling rate**:

$$s_{\text{women}} = \frac{n_{\text{women}}}{N_{\text{women}}} = \frac{170}{180} = 0{,}944$$

$$s_{\text{men}} = \frac{130}{120} = 1{,}083$$

4. Adjustment

$$f^*_{\text{women}} = f_{\text{women}} \times \frac{F_{\text{women}}}{f_{\text{women}}} = 0.567 \times \frac{0.60}{0.567} = 0.60$$

$$f^*_{\text{men}} = f_{\text{men}} \times \frac{F_{\text{men}}}{f_{\text{men}}} = 0.433 \times \frac{0.40}{0.433} = 0.40$$

Stratified random sample

Stratified random sample

- Il est possible de prendre en compte les informations auxilliaires utiles à l'explication de la variable d'intérêt lors de la constitution de l'échantillon, en réalisant un échantillon aléatoire stratifié.

- L'objectif est d'assurer la représentativité de l'échantillon. . .

- . . . en s'assurant que la structure de l'échantillon soit la même que celle de la population.

- It is possible to take into account the auxiliary information useful for explaining the variable of interest when constituting the sample, by carrying out a stratified random sample.

- The objective is to ensure the representativeness of the sample. . .

- . . . making sure that the sample structure is the same as that of the population.

Stratified random sample

- La population est divisée en groupes homogènes mutuellement exclusifs, que l'on appelle des strates.

- À l'intérieur de chaque strate, on procède à un tirage (par exemple un tirage aléatoire simple).

- Chaque individu de la population constituant cette strate a la même probabilité d'être inclus dans l'échantillon.

- The population is divided into mutually homogeneous and exclusive groups, which are called strata.

- Within each stratum, a selection is made (for example a simple random draw).

- Each individual in the population constituting this stratum has the same probability of being included in the sample.

Stratified random sample

Notre objectif est d’estimer une caractéristique.

Our goal is to estimate a characteristic.

Stratified random sample

- Admettons que l'on souhaite estimer le revenu annuel des actifs dans une population.
- Dans cette population, les revenus varient beaucoup.
- Admettons que l'on dispose de la répartition des individus en fonction de leur tranche d'âge et de leur niveau d'étude.
- L'échantillon peut alors **être stratifié** par tranche d'âge ***croisée*** par niveau d'étude : dans chaque strate, le revenu variera moins.

- Let us assume that we wish to estimate the annual income of active in a population.
- In this population, incomes vary a lot.
- Let us assume that we have the distribution of individuals according to their age group and their level of education.
- The sample can then **be stratified** by age group ***crossed*** by level of education: in each stratum, income will vary less.

Stratified random sample

Or, l'erreur type d'une estimation est étroitement liée à la variance de la caractéristique étudiée : si le caractère varie moins dans les strates, l'erreur type de l'estimation sera plus faible.

However, the standard error of an estimate is closely related to the variance of the characteristic studied: if the characteristic varies less in the strata, the standard error of the estimate will be lower.

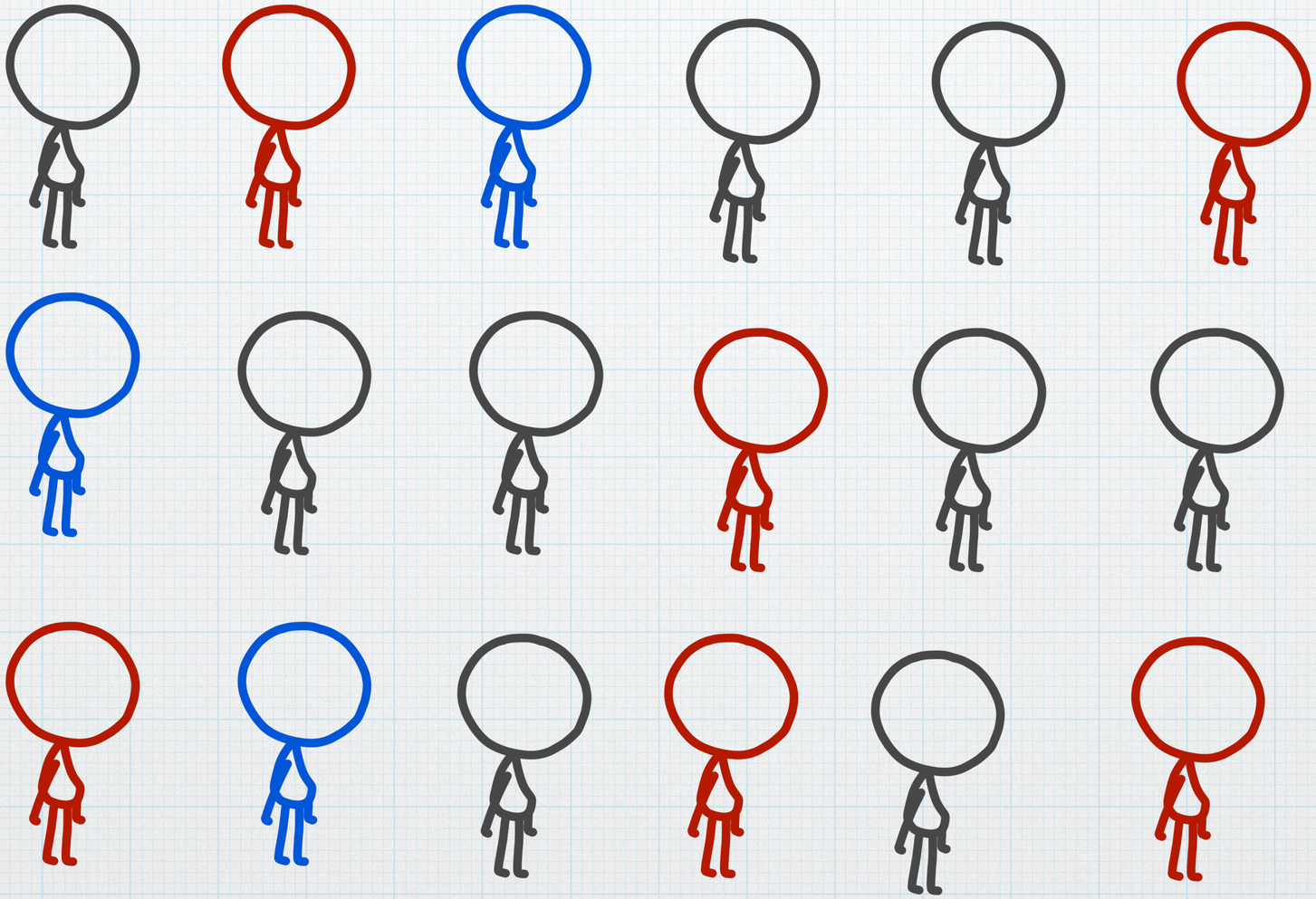
Stratified VS Simple

Dans notre exemple, si on effectuait un échantillonnage aléatoire simple, le hasard pourrait conduire à n'interroger que des personnes dans une strate donnée (e.g. les 18-25 ans sans diplôme) ;

In our example, if we carried out simple random sampling, chance could lead to interviewing only people in a given stratum (eg 18-25 year olds without a diploma);

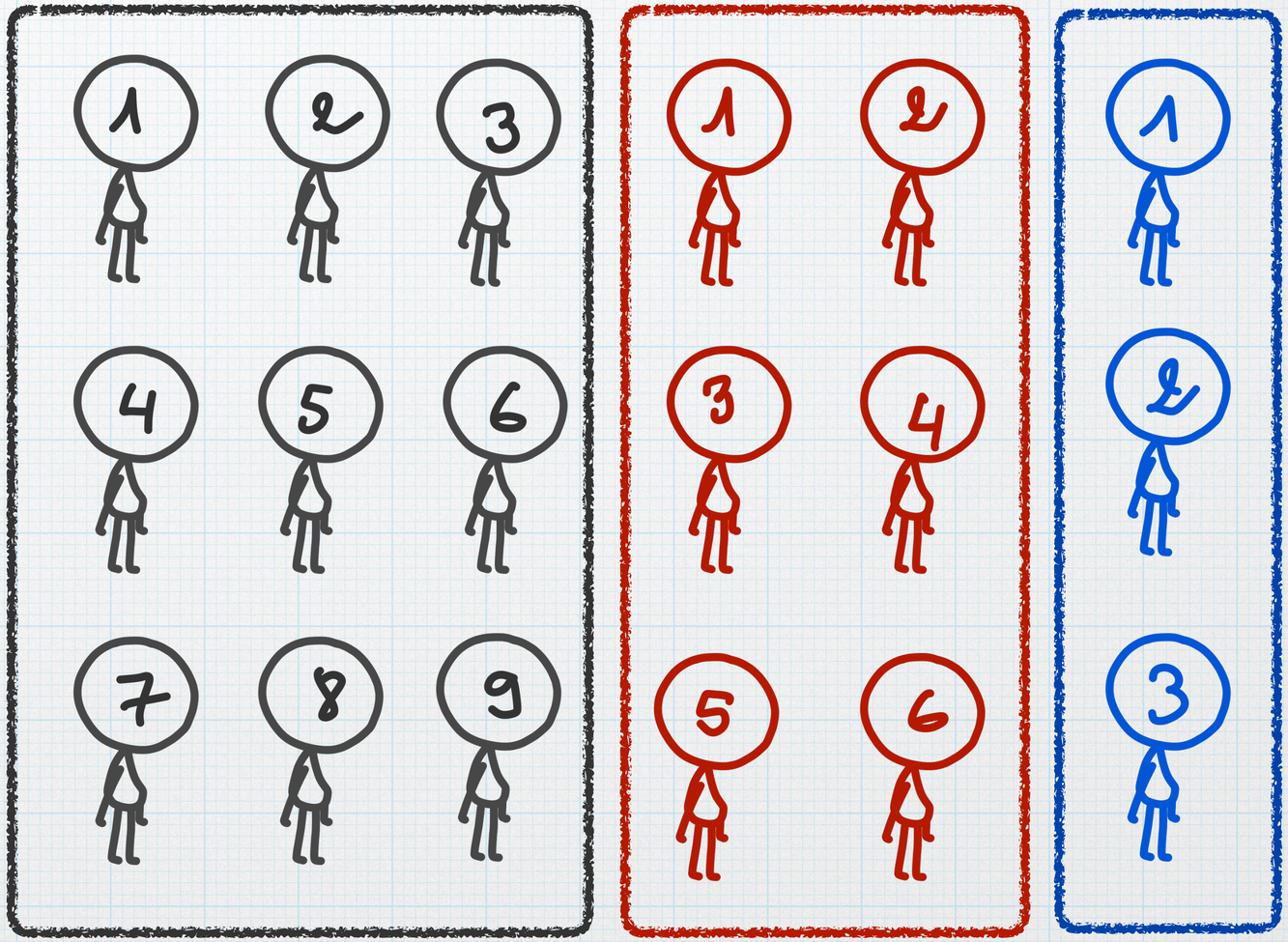
Example

Either the the following population



Example

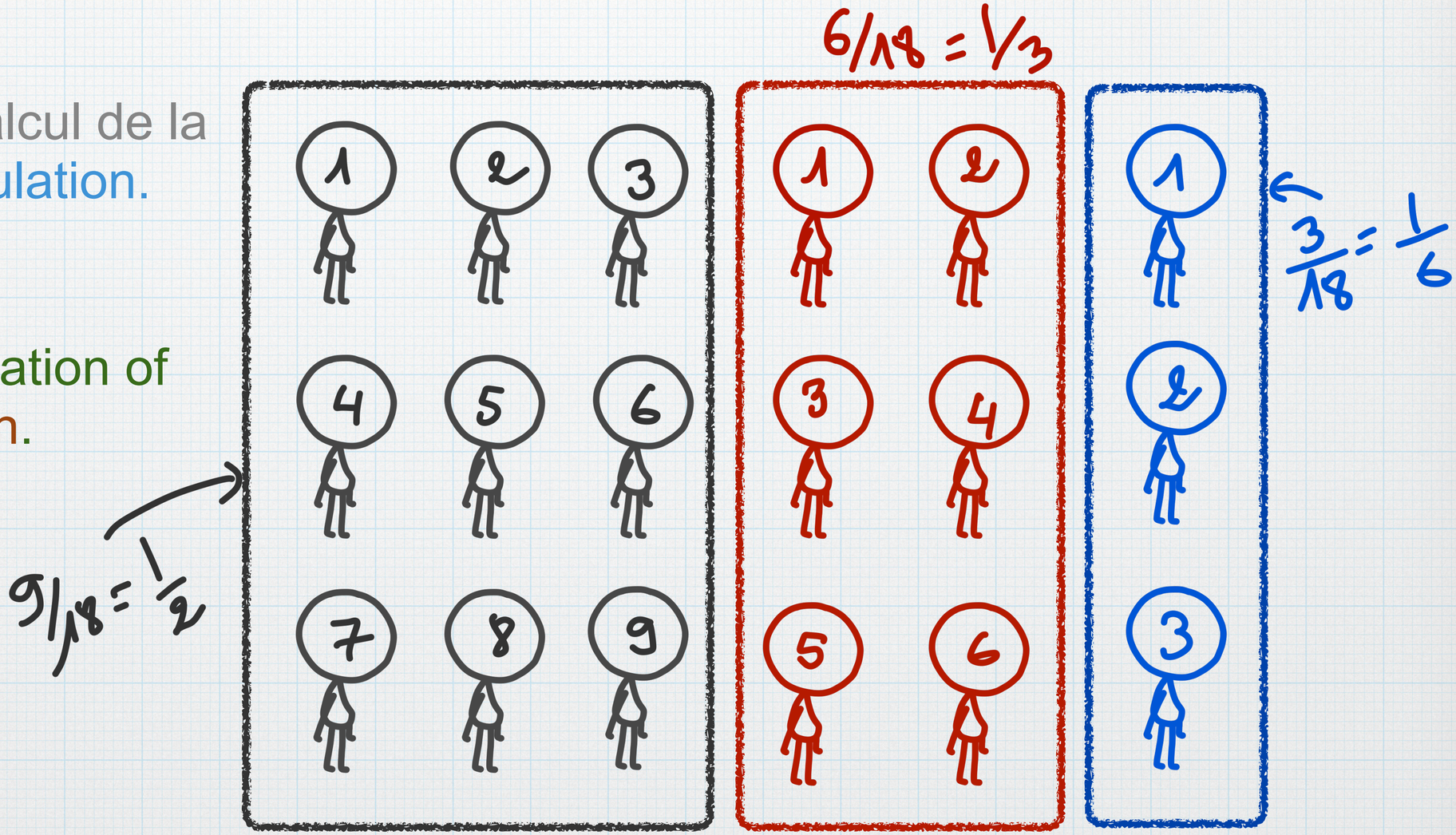
- Séparation de la population par strates selon une caractéristique commune qui influence potentiellement le résultat de l'estimation ;
- Numérotation des individus par strate.
- Separation of the population into strata according to a common characteristic that potentially influences the result of the estimate;
- Numbering of individuals by stratum.



Example

Pour chaque strate : calcul de la proportion dans la population.

For each stratum: calculation of the population proportion.



Example

Admettons que l'on désire un échantillon de taille $n = 5$.

On calcule le nombre d'individus à sélectionner dans chaque strate :

- dans la strate1(bleus): 0,5×5=2,5≈2;
- dans la strate2(noirs): 1/3×5≈2;
- dans la strate3(rouges): 1/6×5≈1

Suppose we want a sample size $n = 5$.

We calculate the number of individuals to be selected in each stratum:

- in stratum1 (blue): 0.5 × 5 = 2.5≈2;
- in stratum2 (blacks): 1/3 × 5≈2;
- in stratum3 (red): 1/6 × 5≈1

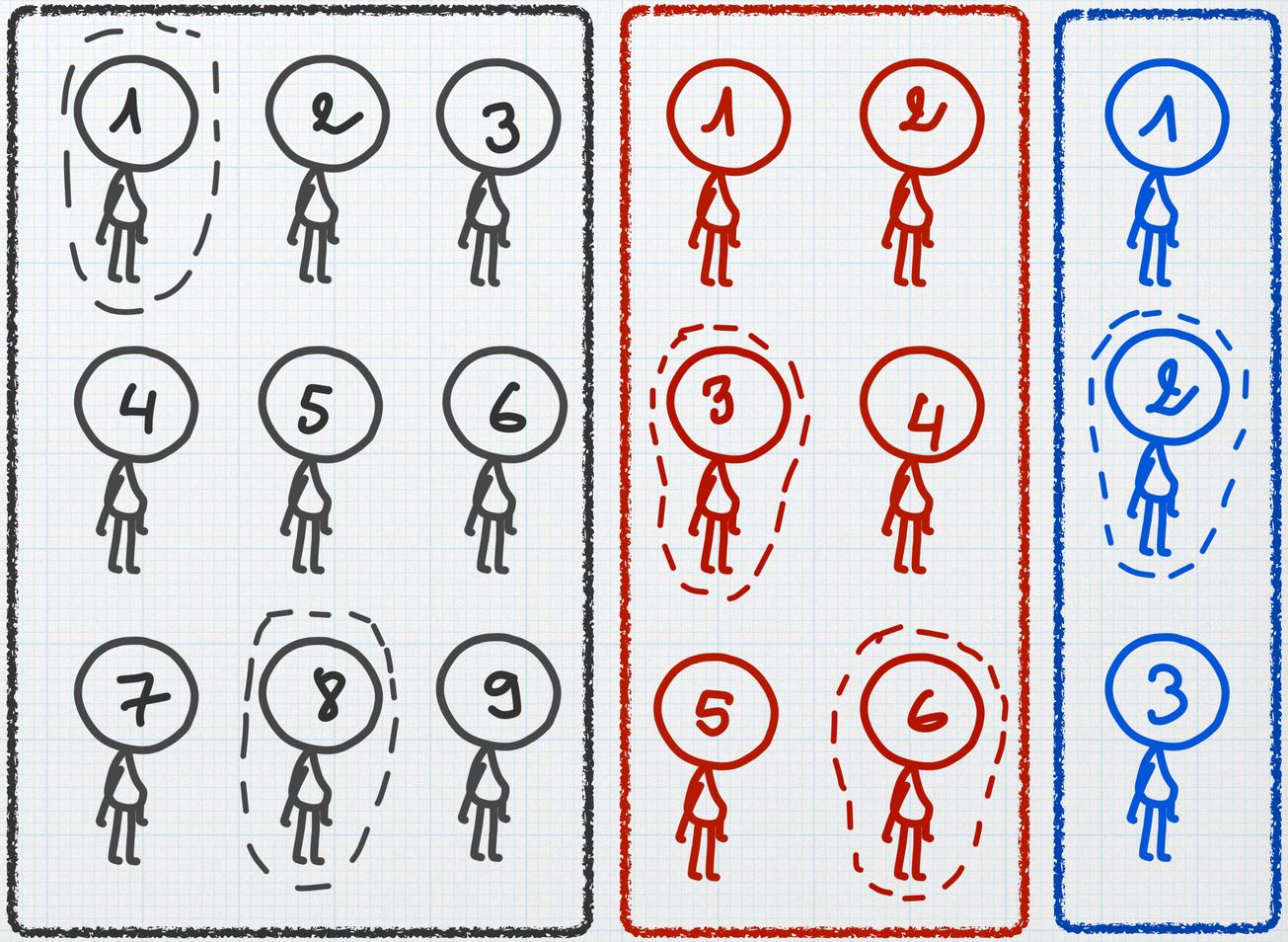
Example

Puis on sélectionne aléatoirement les individus pour chaque strate, par échantillon aléatoire simple.

Par exemple, on a obtenu 1 et 8 pour la strate des noirs, 3 et 6 pour celle des rouges et 2 pour celle des bleus.

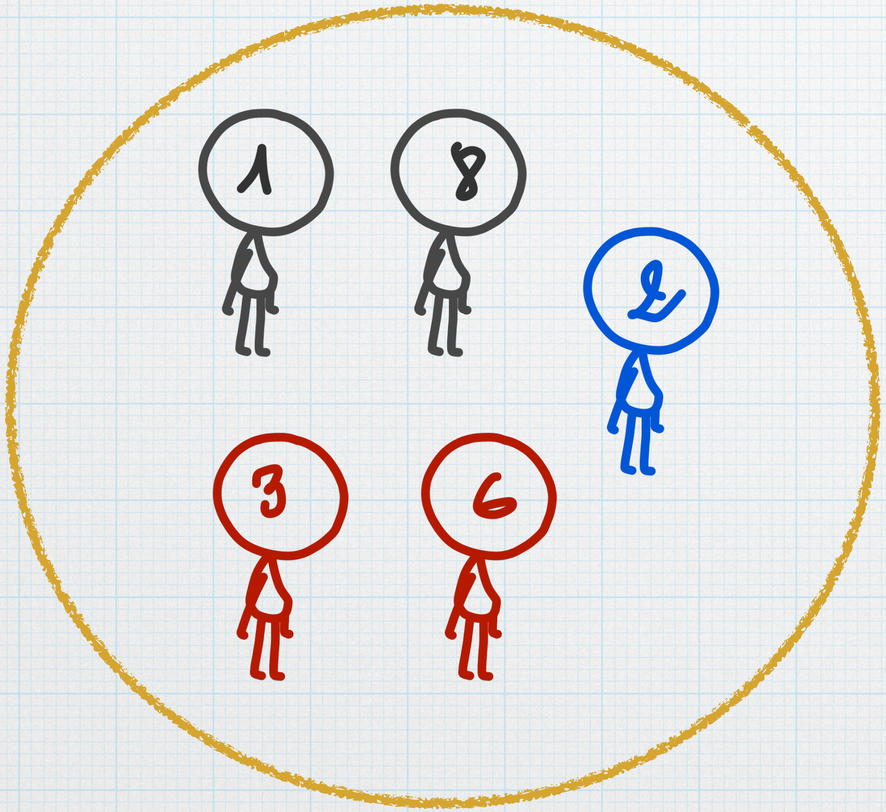
Then we randomly select the individuals for each stratum, by simple random sample.

For example, we obtained 1 and 8 for the black stratum, 3 and 6 for the red stratum and 2 for the blue stratum.



Example

L'échantillon final est bien représentatif de la population :
The final sample is well representative of the population:



Mais cette méthode de tirage aléatoire nécessite de connaître a priori les caractéristiques de la population. Par ailleurs, elle est souvent très coûteuse.

But this method of random selection requires to know a priori the characteristics of the population. Moreover, it is often very expensive.

Random cluster sampling

Il s'agit de constituer des groupes d'individus dans la population (que l'on va appeler des grappes), et d'en sélectionner seulement quelques-unes pour constituer l'échantillon.

Pour que l'échantillon ainsi constitué soit représentatif de la population, il est nécessaire :

1. que la taille des grappes soit à peu près homogène d'une grappe à l'autre ;
2. que les individus qui la composent possèdent des caractéristiques diversifées.

This involves forming groups of individuals in the population (which we will call clusters), and selecting only a few to form a sample.

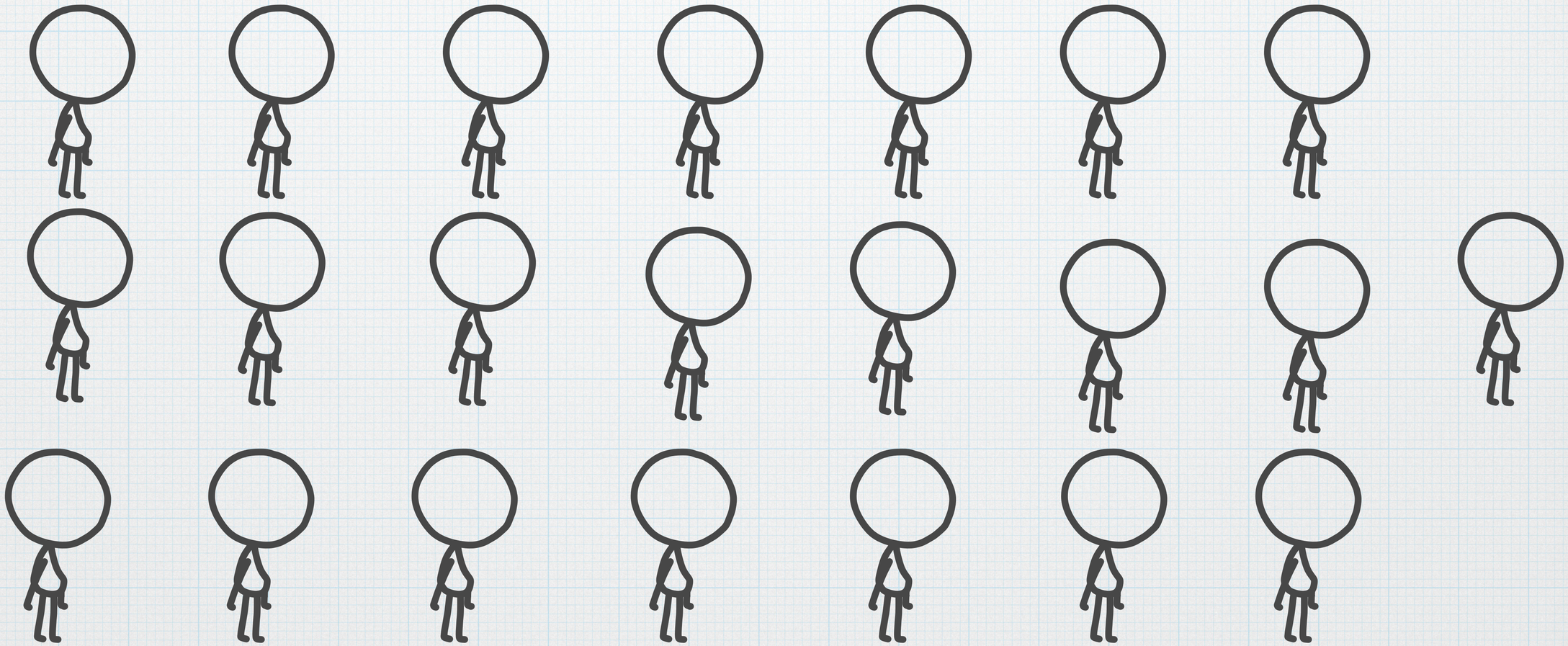
In order for the sample thus constituted to be representative of the population, it is necessary:

1. that the size of the clusters be roughly homogeneous from one cluster to another;
2. that the individuals that compose it have diverse characteristics.

Example

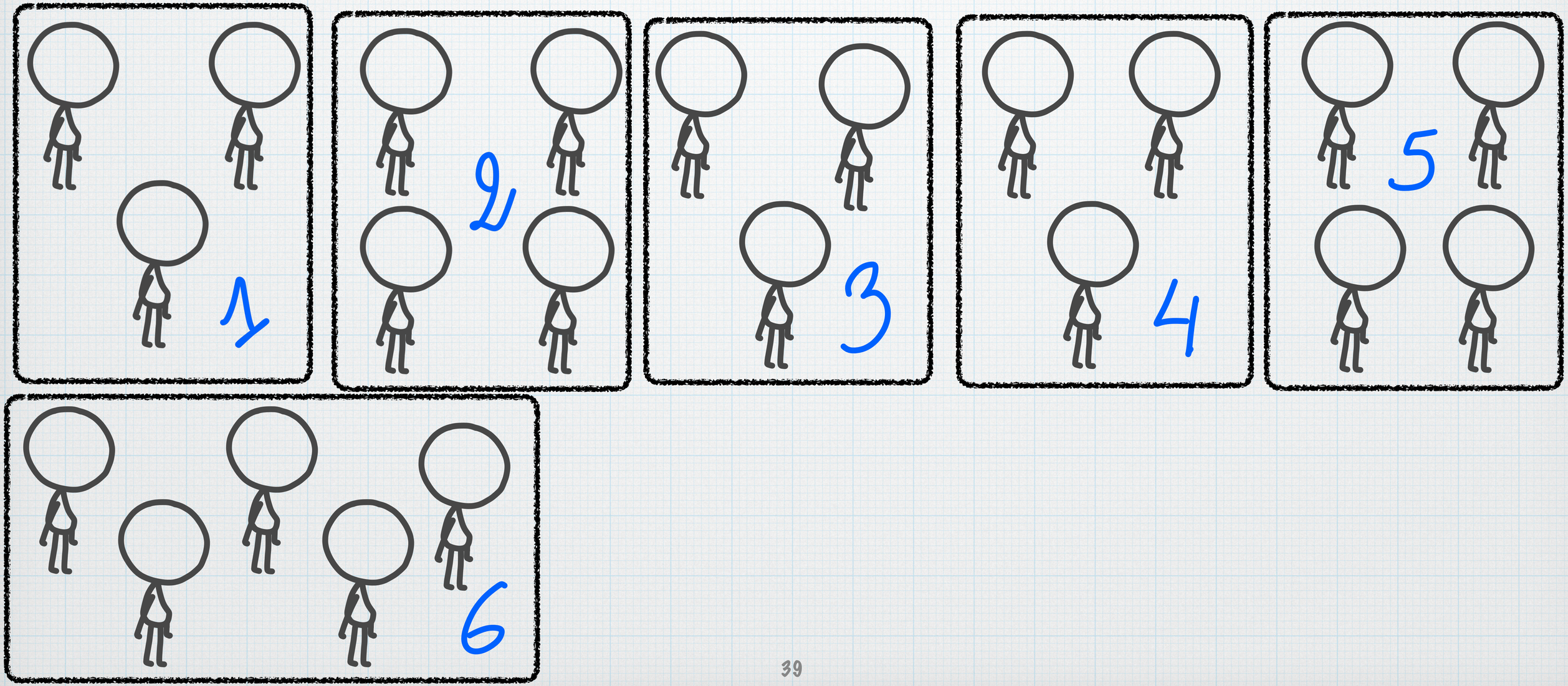
Considérons la population suivante :

Consider the following population:



Séparons la population en grappes numérotées, dont la taille (nb. individus) est semblable d'une grappe à l'autre.

Let us separate the population into numbered clusters, the size of which (no. Individuals) is similar from one cluster to another.



Example : Sample size

Admettons que l'on désire un échantillon de taille $n = 8$.

Suppose we want a sample size $n = 8$

$$\text{Cluster NB} = \frac{n}{\text{Average NB of individuals per cluster}}$$

$$= \frac{8}{\frac{3+4+3+3+4+5}{6}}$$

$$= \frac{8}{\frac{22}{6}} = \frac{24}{11} \approx 2{,}18$$

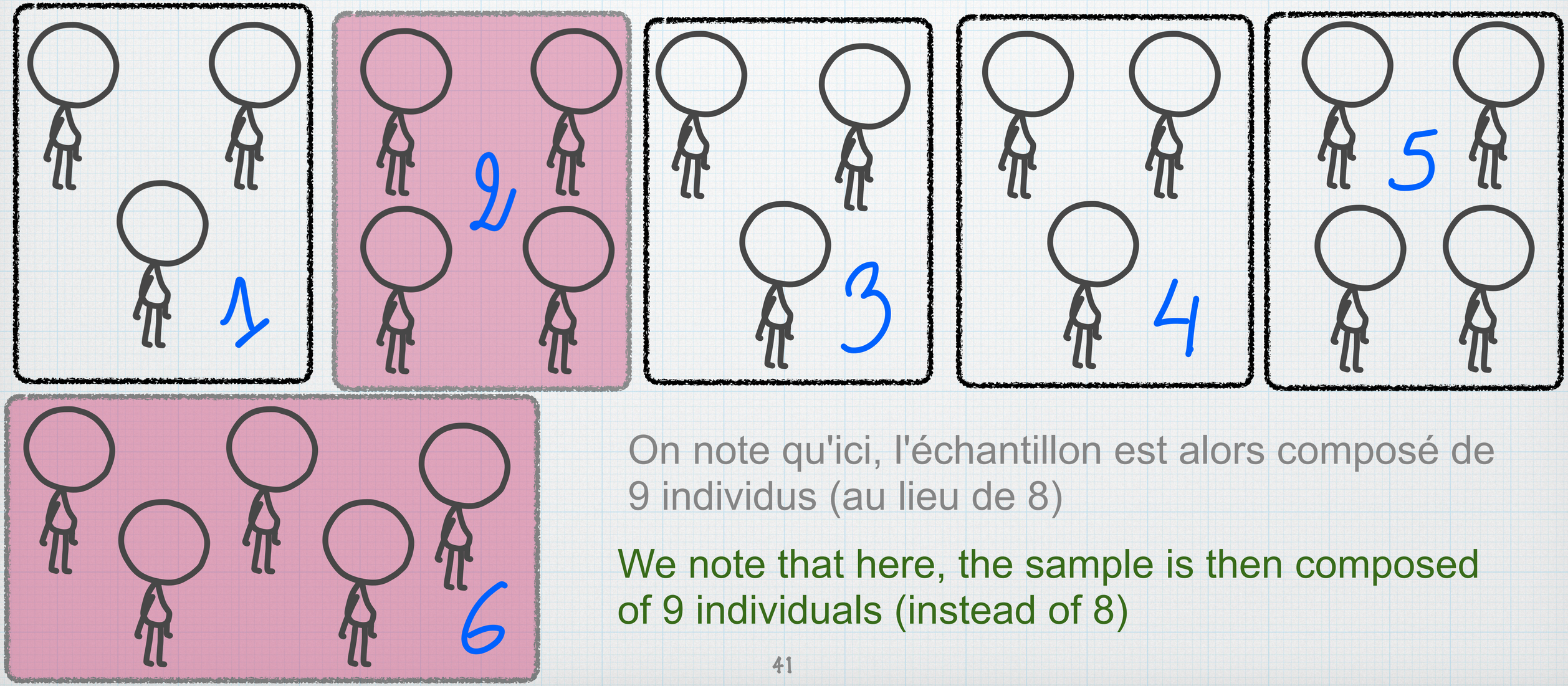
On choisit alors aléatoirement 2 grappes (méthode d'échantillonnage aléatoire simple).

We then randomly choose 2 clusters (simple random sampling method).

Example

Par exemple, nous avons sélectionné aléatoirement les strates 6 et 2 :

For example, we randomly selected strata 6 and 2:



On note qu'ici, l'échantillon est alors composé de 9 individus (au lieu de 8)

We note that here, the sample is then composed of 9 individuals (instead of 8)

Advantage

Le principal intérêt du recours à l'échantillonnage par grappes est qu'il permet de réduire les coûts liés au déplacement lors du sondage :

• par exemple, si on s'intéresse à la mesure d'une caractéristique dans les 101 quartiers de Lille, on peut regrouper les individus par quartiers (chaque quartier représente alors une grappe) ;

• lorsque l'on interroge uniquement quelques grappes, l'organisation de l'enquête est plus efficace et moins coûteuse.

The main advantage of using cluster sampling is that it reduces the costs associated with traveling during the survey:

• for example, if we are interested in the measurement of a characteristic in the 101 districts of Lille, we can group the individuals by district (each district then represents a cluster);

• When only a few clusters are interviewed, the organization of the survey is more efficient and less expensive.

Inconvenience

En revanche, si les grappes sont constituées d'individus statistiques homogènes, il y a risque de corrélation des observations au sein d'une même grappe :

- si on souhaite mesurer le revenu annuel des Lillois et que les grappes tirées sont situées uniquement dans les quartiers suds, on aura une mesure complètement différente que si les grappes tirées sont toutes situées dans le Vieux-Lille. Dans ce cas, un redressement de l'estimation est nécessaire.

On the other hand, if the clusters are made up of statistically homogeneous individuals, there is a risk of correlation of observations within the same cluster:

- if we want to measure the annual income of Lille and the drawn clusters are located only in the southern districts, we will have a completely different measure than if the drawn clusters are all located in Old Lille districts. In this case, a recovery of the estimation is necessary.

In this class

- Dans le cadre de ce cours, nous supposerons que les échantillons ont été constitués selon un tirage aléatoire simple (sauf indication contraire).

- L'indépendance des variables aléatoires est assurée par le fait que l'échantillonnage se réalise au moyen de tirages avec remise.

- For the purposes of this course, we will assume that the samples were drawn using a simple random sampling (unless otherwise specified).

- The independence of the random variables is ensured by the fact that the sampling is carried out by means of draws with replacement.

In this class

Ce que nous allons aborder dans la seconde partie de ce chapitre vise :

- à estimer des paramètres
- ? dont la valeur est établie pour la population entière
- Et à partir de statistiques
- ? dont la valeur est obtenue à partir d'un échantillon

What we are going to approach in the second part of this chapter aims to:

- estimate the parameters
- ? whose value is established for the entire population
- And from statistics
- ? whose value is obtained from a sample

In this class

La connaissance que nous aurons de la réalité sera systématiquement médiée par le calcul d'estimateurs échantillonnaux

The knowledge that we will have about reality, will be systematically mediated by the calculation of sample estimators

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	f

Distribution of means and proportions

Dans la mesure où la sélection d'échantillons suit un processus aléatoire, les statistiques de l'échantillon sont elles-aussi des variables aléatoires et suivent par conséquent une distribution de probabilité.

Nous allons être amenés à répondre aux interrogations suivantes :

➡ Comment se comportent les statistiques de l'échantillon par rapport aux paramètres correspondants de la population ?
➡ La statistique échantillonnale a-t-elle tendance, en moyenne, à être proche du paramètre vrai ou en diverge-t-elle systématiquement ?

To extent that, the selection of samples follows a random process, the sample statistics are also random variables and therefore follow a probability distribution.

We will have to answer the following questions:

➡ How do the statistics sample behave in relation to the corresponding parameters of the population?
➡ Does the sample statistic tend, on average, to be close to the true parameter or does it systematically diverge from it?

L’estimation d’une moyenne ou d’une proportion à partir d’un échantillon n’aura que très peu de chances (voire aucune) d’être égale à celle que nous aurions obtenue à partir de la population (sauf si la valeur présentée par tous les individus de la population est la même).

Il est même possible d’obtenir autant d’estimations échantillonnales différentes que d’échantillons à partir desquels ces estimations sont réalisées.

Aussi, nous devons être conscients qu’il existe une variation de la valeur des estimations échantillonnales, appelée variation d’échantillonnage.

Estimating a mean or proportion from a sample is unlikely (or none) to be equal to what we would have obtained from the population (unless, if the value presented by all individuals in the population is the same).

It is even possible to obtain as many different sample estimates as there are samples from which these estimates are made.

Also, we should be aware that there is variation in the value of the sample estimates, called the sampling variation.

Distribution of means and proportions

Toutefois, nous sommes en mesure de calculer la distribution d'échantillonnage de ces estimations, qu'il s'agisse :

➡de moyennes; on parle alors de Distribution d'Echantillonnage des Moyennes (ou DEM) ;
➡de pourcentages; on parle alors de Distribution d'Echantillonnage des Pourcentages (ou DEP).

However, we are able to calculate the sampling distribution of these estimates, whether they are:

➡of averages; this is then referred to as the Sampling Distribution of Deans (or SDM);
➡Of percentages; this is referred to as the Percentage Sampling Distribution (or PSD).

Distribution of means and proportions

Une distribution d'échantillonnage de moyennes (*respectivement de pourcentages)* donne la distribution des moyennes échantillonnales (r*esp. des pourcentages échantillonnaux*) tirées de tous les échantillons différents de même taille n à partir d'une population de taille N.

La distribution d'échantillonnage (de moyennes ou de pourcentages) se définit ainsi comme une distribution de probabilités affectant à chaque réalisation échantillonnale de l'estimateur (de la moyenne ou du pourcent- age) une probabilité d'occurrence.

A sampling distribution of means (*respectively percentages*) gives the distribution of sample means (resp. *Sample percentages*) drawn from all different samples of the same size n from a population of size N.

The sampling distribution (of means or percentages) is thus defined as a probability distribution assigning to each sampling realization of the estimator (of the mean or of the percentage) an occurrence probability.

Sampling distribution of means

The SDM Average $\mathbb{E}(\bar{X})$

Population

Considérons une population de 5 étudiants auxquels on demande le temps qu'ils consacrent chaque semaine au travail personnel dans le cadre de leurs études. Les réponses sont présentées dans le tableau ci-dessous :

Consider a population of 5 students who are asked how much time they devote each week to personal work as part of their studies. The responses are presented in the table below:

	NB of hours
Student A	7
Student B	3
Student C	6
Student D	10
Student E	4

Le calcul du temps moyen consacré au travail personnel pour l'ensemble des 5 étudiants est immédiat :

The calculation of the average time spent on personal work for all 5 students is immediate:

$$\mu = (7 + 3 + 6 + 10 + 4)/5 = 30/5 = 6h$$

Question…

✓Si on ne dispose pas des moyens nécessaires à l'interrogation des 5 étudiants et si on doit se contenter d'un échantillon de 3 étudiants pris au hasard parmi les 5, quelles estimations échantillonnales de la moyenne peut-on calculer ?

✓If we do not have the means to interview the 5 students and if we have to draw a sample of 3 students among 5 (random), what the sample estimation of the mean can we calculate?

Notations

➡ μ : moyenne du caractère étudié dans la population

➡$\overline{X}$: variable aléatoire mesurant la moyenne du caractère étudié dans un échantillon de taille n (que l'on appelle moyenne d'échantillon)

➡ $\overline{x}$: moyenne obtenue (mesurée) dans un échantillon de taille n, réalisation de la v.a. $\overline{X}$.

➡μ mean of the trait studied in the population

➡$\overline{X}$ random variable measuring the mean of the characteristic studied in a sample of size n (called the sample mean)

➡$\overline{x}$: mean obtained (measured) in a sample of size n, realization of the v.a. $\overline{X}$.

All possible samples

On détermine tout d'abord le nombre d'échantillons différents de 3 personnes envisageables à partir de la population des 5 étudiants :

Sample	Composition	Individual Values	$\overline{X}$
1	ABC	7 3 6	16/3
2	ABD	7 3 10	20/3
3	ABE	7 3 4	14/3
4	ACD	7 6 10	23/3
5	ACE	7 6 4	17/3
6	ADE	7 10 4	21/3
7	BCD	3 6 10	19/3
8	BCE	3 6 4	13/3
9	BDE	3 10 4	17/3
10	CDE	6 10 4	20/3

We first determine the number of different samples of 3 possible people from the population of 5 students:

$$C_5^3 = \frac{5!}{3!2} = \frac{5 \times 4 \times 3!}{3! \times 2} = 10$$

On calcule ensuite la moyenne du temps hebdomadaire consacré aux études pour chaque échantillon (voir le tableau)

We then calculate the average weekly time spent studying for each sample (see the table)

Law of the sample mean

On peut construire à partir du tableau précédent le tableau de la distribution des moyennes échantillonnales :

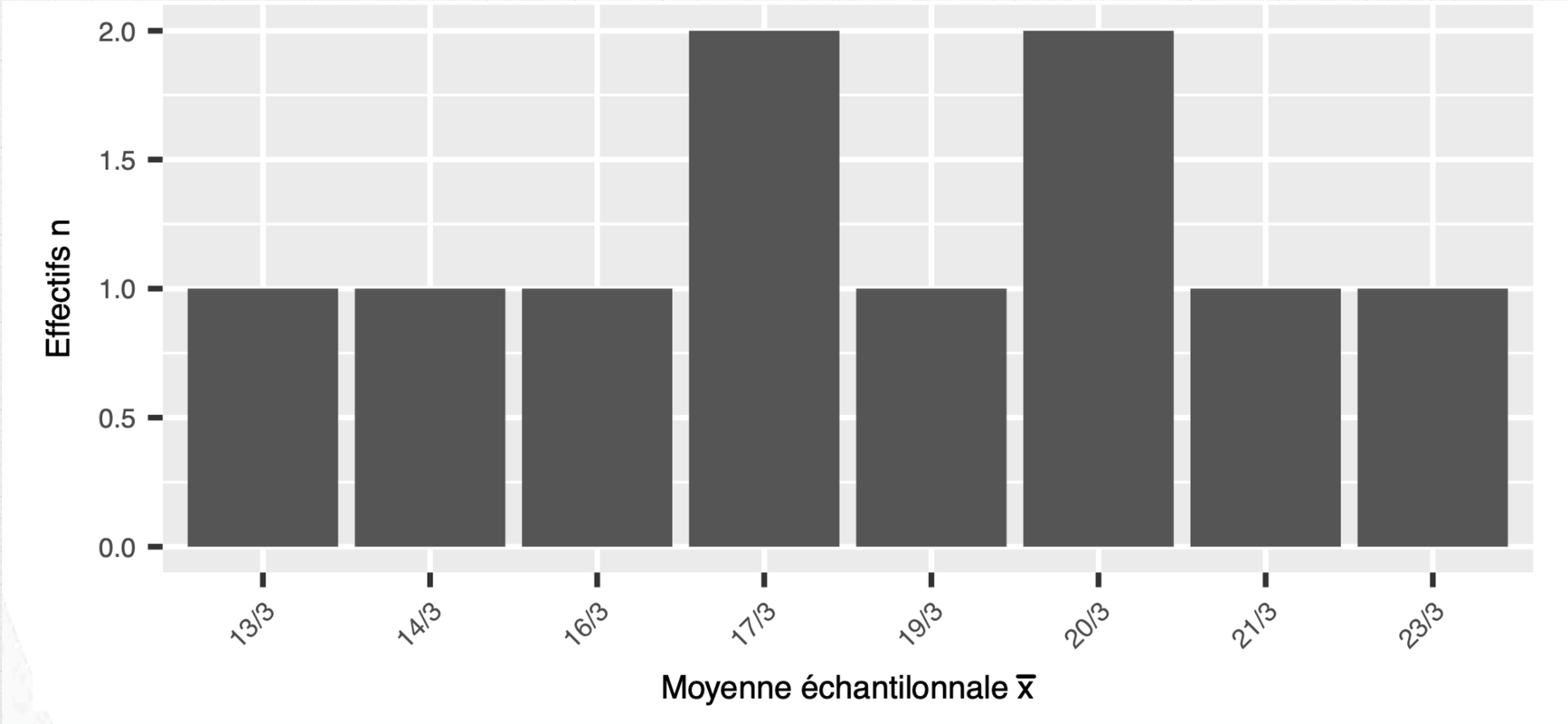
We can build from the previous table the table of the distribution of the sample means:

Sample mean $\overline{x}_i$	effective n_i	Occurrence Probability $\mathbb{P}(\overline{X} = \overline{x})$
13/3	1	1/10
14/3	1	1/10
16/3	1	1/10
17/3	2	2/10
19/3	1	1/10
20/3	2	2/10
21/3	1	1/10
23/3	1	1/10
Total	10	1

Graphical representation of the distribution of $\overline{X}$

La représentation graphique de la distribution des moyennes échantillonnales est donnée ci-dessous :

The graphical representation of the distribution of sample means is given below:



Expected Value of $\overline{X}$

Calculons la moyenne de la distribution d'échantillonnage des moyennes

Calculate the mean of the sampling distribution of means

$$\mu_{\overline{X}} = \mathbb{E}(\overline{X})$$

$$= \sum_{j=1}^{8} \overline{x} \times p_j$$

$$= \frac{13}{3} \times \frac{1}{10} + \frac{14}{3} \times \frac{1}{10} + \ldots + \frac{23}{3} \times \frac{1}{10}$$

$$= 6$$

Expected Value of $\overline{X}$

$$\mu_{\overline{X}} = \mathbb{E}(\overline{X}) = 6 = \mu$$

On note que la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne obtenue sur la population : $\mu_{\overline{X}} = \mu$.
Ainsi, *en moyenne*, la statistique échantillonnale de la moyenne de la variable est égale au paramètre de la population.

La moyenne d'un échantillon aléatoire est un estimateur non-biaisé de la moyenne de la population.

Note that, the mean of the means sampling distribution is equal to the mean obtained on the population: $\mu_{\overline{X}} = \mu$.
Thus, on average, the sample statistic of the mean of the variable is equal to the population parameter.

The mean of a random sample is an unbiased estimator of the population average.

Expected Value of $\overline{X}$

➡On s'intéresse aux valeurs prises par une v.a. X suivant une distribution stat. de moyenne μ et de variance σ^2, i.e., $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$

➡On prélève tous les échantillons de taille n dans une pop. supposée infinie :

- un premier échantillon de taille n contenant des valeurs de X : $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$
- un deuxième échantillon de taille n contenant des valeurs de X : $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$
- un troisième échantillon de taille n contenant des valeurs de X : $x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}$
- …

➡We are interested in the values taken by a X r.v. following a stat distribution with μmean and σ^2 variance , i.e. $\mathbb{E}(X) = \mu$ and $\mathbb{V}ar(X) = \sigma^2$

➡All samples of size n are taken in a pop. assumed to be infinite:

- a first sample of size n containing values of X : $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$
- a second sample of size n containing values of X : $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$
- a third sample of size n containing values of X : $x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}$
- …

Expected Value of $\overline{X}$

➡Dans les échantillons :

- Les valeurs $x_1^{(1)}, x_1^{(2)}, x_1^{(3)} \ldots$ peuvent être considérées comme les valeurs d'une v.a X_1 qui suit la loi de X.
- Les valeurs $x_2^{(1)}, x_2^{(2)}, x_2^{(3)} \ldots$ peuvent être considérées comme les valeurs d'une v.a X_2 qui suit la loi de X. ?
- Les valeurs $x_3^{(1)}, x_3^{(2)}, x_3^{(3)} \ldots$ peuvent être considérées comme les valeurs d'une v.a X_3 qui suit la loi de X.

➡In the samples:

- The values $x_1^{(1)}, x_1^{(2)}, x_1^{(3)} \ldots$ can be considered as the values of a r.v. X_1 which follows the law of X
- The values $x_2^{(1)}, x_2^{(2)}, x_2^{(3)} \ldots$ can be considered as the values of a r.v. X_2 which follows the law of X
- The values $x_3^{(1)}, x_3^{(2)}, x_3^{(3)} \ldots$ can be considered as the values of a r.v. X_3 which follows the law of X

Expected Value of $\overline{X}$

➡ Pour chaque échantillon j, on calcule la moyenne échantillonnale :

➡For each sample j, we calculate the sample mean:

$$\overline{x}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(j)}$$

➡Chaque moyenne échantillonnale $\overline{x}^{(j)}$ (soit $\overline{x}^{(1)}, \overline{x}^{(2)}, \ldots$) est une combinaison linéaire de v.a. de moyenne μ et de variance σ^2

➡Each sample mean $\overline{x}^{(j)}$ (let $\overline{x}^{(1)}, \overline{x}^{(2)}, \ldots$) is a linear combination of r.v. with μ mean and σ^2 variance

➡Il s'agit de la valeur prise dans un échantillon par la variable aléatoire

➡This value is taken from a sample by the random variable :

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

➡Ainsi, la v.a. représentant la moyenne d'échantillon $\overline{X}$ est définie par :

➡Then, the r.v. representing the sample mean $\overline{X}$ is defined by :

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Expected Value of $\overline{X}$

➡L'espérance de cette variable aléatoire est alors :

➡The expectation of this random variable is then:

$$\mu_{\overline{X}} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

➡La moyenne de la distribution d'échantillonnage des moyennes de X est ainsi égale à la moyenne de X sur la population. $\overline{X}$ est un estimateur sans biais de μ

➡The mean of the sampling distribution of the means X is equal to the mean of X over the population. $\overline{X}$ is an unbiased estimator of μ

Reminder on bias

Soit $\hat{\theta}$ un estimateur (obtenu sur un échantillon) du paramètre θ (la vraie valeur dans la population).

La valeur de l'estimateur peut varier en fonction de l'échantillon que l'on a obtenu. Aussi, l'estimateur $\hat{\theta}$ est une variable aléatoire (X).

Le biais de l'estimateur est défini comme suit : $\textbf{Biais}(\hat{\theta}) \equiv \mathbb{E}(\hat{\theta}) - \theta$

Let $\hat{\theta}$ be an estimator (obtained on a sample) of the parameter θ (the true value in the population).

The value of the estimator may vary depending on the sample obtained. Also, the estimator $\hat{\theta}$ is a random variable (X).

The bias of the estimator is defined as follows: $\textbf{Biais}(\hat{\theta}) \equiv \mathbb{E}(\hat{\theta}) - \theta$

Reminder on bias

Si le biais vaut 0, l'estimateur est qualifié de non biaisé : dans ce cas, $\mathbb{E}(\hat{\theta}) = \theta$, i.e., en moyenne, la valeur de l'estimateur est égale à la vraie valeur.

Nous venons de voir que pour estimer la moyenne de la population μ, si on utilise comme estimateur la moyenne échantillonnale $\mu_{\overline{X}}$, alors cet estimateur est non biaisé.

If the bias is 0, the estimator is qualified as unbiased: in this case, $\mathbb{E}(\hat{\theta}) = \theta$, i.e., on average, the value of the estimator is equal to the true value.

We have just seen that to estimate the mean of the population μ, if we use as estimator the sample mean $\mu_{\overline{X}}$, then this estimator is unbiased.

The SDM Standard Deviation $\sqrt{\mathbb{V}ar(\bar{X})}$

Standard Deviation

À présent, intéressons-nous à l'écart-type de la distribution d'échantillonnage des moyennes :

Now, let's look at the standard deviation of the sampling distribution of the means:

➡quel lien existe-t-il entre l'écart-type de la distribution d'échantillonnage des moyennes et l'écart-type de la variable dans la population ?

➡What is the relationship between the standard deviation of the sampling distribution of the means and the standard deviation of the variable in the population?

Student	NB hours x_i	$x_i - \mu$	$(x_i - \mu)^2$
A	7	1	1
B	3	-3	9
C	6	0	0
D	10	4	16
E	4	-2	4
Total			30

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} = \sqrt{\frac{1}{5} \times 30} = \sqrt{6}$$

En moyenne, le temps passé par les étudiants à travailler chaque semaine s'écarte de 2,45h de la durée moyenne hebdomadaire égale à 6h.

On average, the time spent by students working each week deviates by 2.45 hours from the average weekly time of 6 hours.

Standard Deviation

Sample mean $\overline{x}_i$	effective n_i	$\overline{x}_i - \mu$	$(\overline{x}_i - \mu)^2$	$n_i(\overline{x}_i - \mu)^2$
13/3	1	-5/3	25/9	25/9
14/3	1	-4/3	16/9	16/9
16/3	1	-2/3	4/9	4/9
17/3	2	-1/3	1/9	2/9
19/3	1	1/3	1/9	1/9
20/3	2	2/3	4/9	8/9
21/3	1	3/3	9/9	9/9
23/3	1	5/3	25/9	15/9
Total	10			90/10=10

On a donc:

$$\sigma_{\overline{X}} = \sqrt{\frac{1}{C_N^n} \sum_{i=1}^{C_N^n} n_i(\overline{x}_i - \mu)^2}$$

So we have:

$$= \sqrt{\frac{1}{10} \times 10} = 1$$

Standard Deviation

$$\sigma^2_{\overline{X}} = \mathbb{V}ar(\overline{X}) = \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\mathbb{V}ar\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

La variance de la distribution d'échantillonnage des moyennes de X est égale à la variance de X sur la population à un coefficient $\frac{1}{n}$ près :

➡la variance échantillonnale diminue avec la taille de l'échantillon

The variance of the sampling distribution of the means of X is equal to the variance of X over the population up to a coefficient $\frac{1}{n}$:

➡ sample variance decreases with sample size

Standard Deviation

En population finie, on a

In finite population, we have

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

On vérifie dans l'exemple des étudiants :

Let we check in the students example:

$$\sigma_{\overline{X}} = \frac{\sqrt{6}}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}}$$

$$= \frac{\sqrt{6}}{\sqrt{3}}\sqrt{\frac{2}{4}}$$

$$= \sqrt{\frac{6}{3}}\sqrt{\frac{1}{2}} = \sqrt{\frac{6}{6}} = 1$$

Standard Deviation : remarques

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 1$ lorsque la population est infinie, le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ peut être négligé

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 0$ lorsque la taille de l'échantillon devient grand et se rapproche de la taille de la population, le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ tend vers 0.

$\lim_{N \to +\infty} \sigma_{\overline{X}} = 0$ L'écart-type de la distribution d'échantillonnage des moyennes tends vers 0 lorsque la taille de l'échantillon augmente. Plus l'information accumulée sur la population est importante, plus la variation d'échantillonnage est faible.

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 1$ when the population is infinite, the correction factor $\sqrt{\frac{N-n}{N-1}}$ can be overlooked

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 0$ as the sample size becomes large and approaches the population size, the correction factor $\sqrt{\frac{N-n}{N-1}}$ tends to 0.

$\lim_{N \to +\infty} \sigma_{\overline{X}} = 0$ The standard deviation of the sampling distribution of the means tends towards 0 as the sample size increases. The greater the information accumulated on the population, the smaller the sampling variation.

Percentage sampling distribution

Moyenne de la DEP $\mathbb{E}(F)$

Notations

➡ p: proportion dans la population d'un caractère que l'on étudie

➡F: variable aléatoire mesurant la fréquence d'apparition du caractère étudié dans un échantillon de taille n (que l'on appelle proportion d'échantillon)

➡ f : proportion obtenue dans un échantillon de taille n, réalisation de la v.a. F.

➡p : proportion in the population of a trait being studied

➡F random variable measuring the appearance frequency of the studied characteristic in a sample of size n (called the sample proportion)

➡f: proportion obtained in a sample of size n, realization of the v.a. F.

F, la fréquence d'apparition du caractère étudié dans un échantillon de taille n vaut $F = X/n$, avec X le nombre de fois où le caractère apparaît dans l'échantillon.

F, frequency of appearance of the characteristic studied in a sample of size n is $F = X/n$, with X lthe number of times the characteristic appears in the sample.

Example

Considérons la même population de 5 étudiants auxquels on demande s'ils ont lu au moins un livre au cours des douze derniers mois. Leur réponse est consignée dans le tableau suivant :

Consider the same population of 5 students asked if they have read at least one book in the past twelve months. Their response is recorded in the following table:

Student	Reading
A	Yes
B	No
C	Yes
D	No
E	No

Le calcul du pourcentage d'étudiants ayant lu au moins un livre est immédiat :

The calculation of the percentage of students who have read at least one book is immediate:

$$p = \frac{x}{N} = \frac{2}{5} = 40\,\%$$

Si on ne dispose pas des moyens nécessaires à l'interrogation des 5 étudiants et si on doit se contenter d'un échantillon de 3 étudiants pris au hasard parmi les 5, quelles estimations échantillonnales du pourcentage peut-on calculer ?

If we do not have the means to interview the 5 students and if we have to be satisfied with a sample of 3 students taken at random from the 5, what sample estimates of the percentage can we calculate?

All possible samples

On détermine tout d'abord le nombre d'échantillons différents de 3 personnes envisageables à partir de la population des 5 étudiants :

We first determine the number of different samples of 3 possible people from the population of 5 students:

Sample	Composition	Individual Values	Sample pourcentage f_j
1	ABC	Yes no yes	2/3
2	ABD	Yes no no	1/3
3	ABE	Yes no no	1/3
4	ACD	Yes yes no	2/3
5	ACE	Yes yes no	2/3
6	ADE	Yes no no	1/3
7	BCD	No yes no	1/3
8	BCE	No yes no	1/3
9	BDE	No non no	0/3
10	CDE	Yes no no	1/3

$$C_5^3 = \frac{5!}{3!2} = \frac{5 \times 4 \times 3!}{3! \times 2} = 10$$

On calcule ensuite la moyenne du temps hebdomadaire consacré aux études pour chaque échantillon (voir le tableau)

We then calculate the average weekly time spent studying for each sample (see the table)

F distribution

On peut construire à partir du tableau précédent le tableau de la distribution des pourcentages échantillonnage :

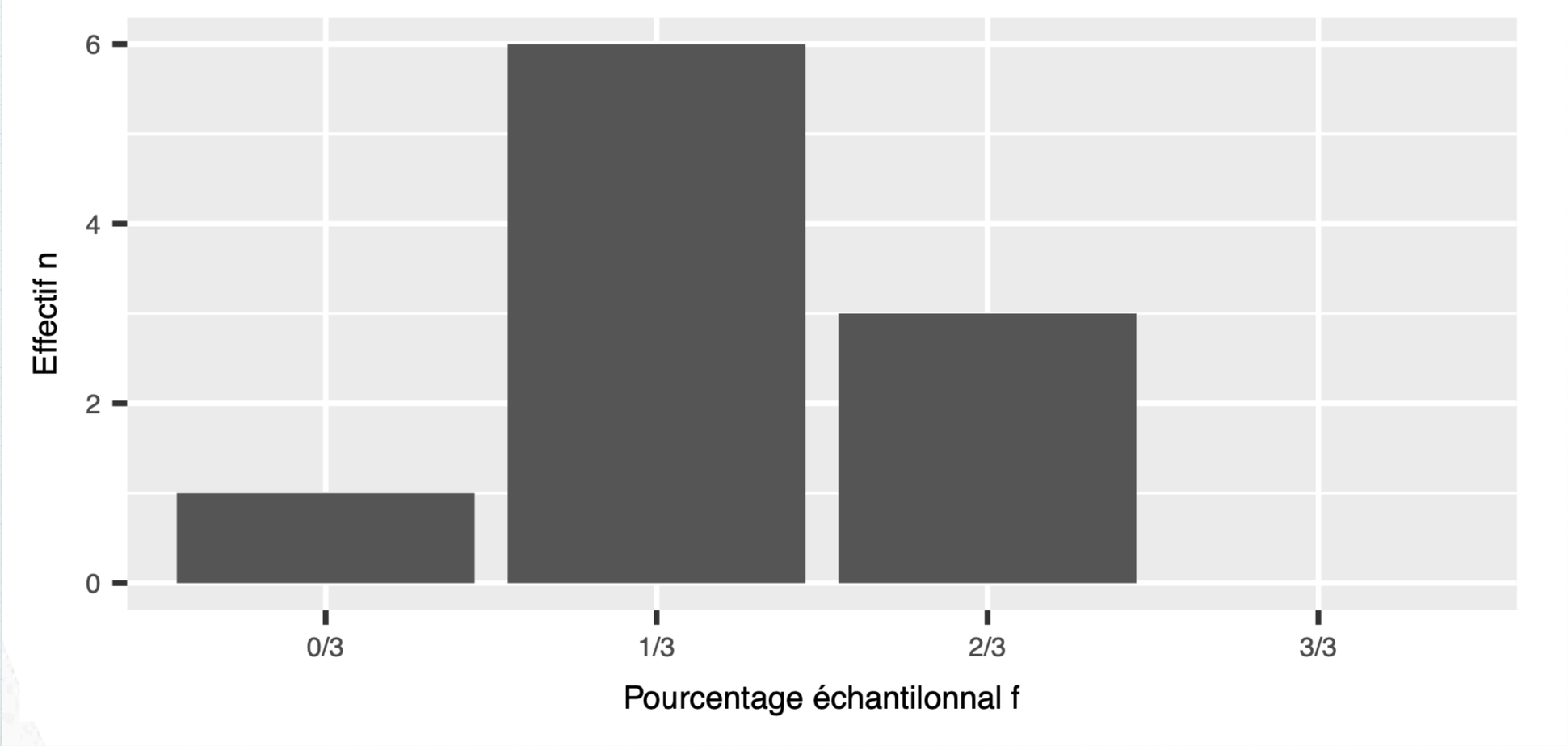
We can build from the previous table the table of the distribution of the sample pourcentage:

Sample pourcentage f_i	effective	Occurrence Probability $\mathbb{P}(F = f_i)$
0/3	1	1/10
1/3	6	6/10
2/3	3	3/10
3/3	0	0/10
Total	10	1

Graphical representation of the distribution of $\overline{X}$

La représentation graphique de la distribution des pourcentages échantillonnage est donnée ci-dessous :

The graphical representation of the distribution of sample pourcentage is given below:



Expected Value of F

Calculons la moyenne de la distribution d'échantillonnage des pourcentages

Calculate the mean of pourcentage sampling distribution

$$\mu_F = \mathbb{E}(F)$$

$$= \sum_{j=1}^{4} f \times \mathbb{P}(F = f_i)$$

$$= \frac{0}{3} \times \frac{1}{10} + \frac{1}{3} \times \frac{6}{10} + \frac{2}{3} \times \frac{1}{10} + \frac{3}{3} \times \frac{0}{10}$$

$$= 0.4 = 40\,\%$$

Expected Value of F

$$\mu_F = \mathbb{E}(F) = 0.4 = p$$

On note que la moyenne de la distribution d'échantillonnage des pourcentage est égale à la moyenne obtenue sur la population : $\mu_F = p$.
Ainsi, *en moyenne*, la statistique échantillonnale du pourcentage de la variable est égale au paramètre de la population.

Le pourcentage d'un échantillon aléatoire est un estimateur non-biaisé de la moyenne de la population.

Note that, the mean of the pourcentage sampling distribution is equal to the mean obtained on the population: $\mu_F = \mu$.
Thus, on average, the sample statistic of the pourcentage of the variable is equal to the population parameter.

The pourcentage of a random sample is an unbiased estimator of the population pourcentage.

Expected Value of F

➡On s'intéresse aux valeurs prises par une v.a. X suivant une distribution de Bernoulli de paramètre p

➡On prélève tous les échantillons de taille n dans une pop. supposée infinie :

- un premier échantillon de taille n contenant des valeurs de X : $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$
- un deuxième échantillon de taille n contenant des valeurs de X : $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$
- un troisième échantillon de taille n contenant des valeurs de X : $x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}$
- …

➡We are interested in the values taken by a X r.v. following a Bernoulli distribution with the p parameter.

➡All samples of size n are taken in a pop. assumed to be infinite:

- a first sample of size n containing values of X : $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$
- a second sample of size n containing values of X : $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$
- a third sample of size n containing values of X : $x_1^{(3)}, x_2^{(3)}, \ldots, x_n^{(3)}$
- …

Expected Value of F

➡Dans les échantillons :

- ? Les valeurs $x_1^{(1)}, x_1^{(2)}, x_1^{(3)} \dots$ peuvent être considérées comme les valeurs d'une v.a X_1 qui suit la loi de X, soit $X_1 \sim \mathfrak{B}(p)$
- ? Les valeurs $x_2^{(1)}, x_2^{(2)}, x_2^{(3)} \dots$ peuvent être considérées comme les valeurs d'une v.a X_2 qui suit la loi de X, ? soit $X_2 \sim \mathfrak{B}(p)$
- Les valeurs $x_3^{(1)}, x_3^{(2)}, x_3^{(3)} \dots$ peuvent être considérées comme les valeurs d'une v.a X_3 qui suit la loi de X, soit $X_n \sim \mathfrak{B}(p)$

➡In the samples:

- The values $x_1^{(1)}, x_1^{(2)}, x_1^{(3)} \dots$ can be considered as the values of a r.v. X_1 which follows the law of X, then $X_1 \sim \mathfrak{B}(p)$
- The values $x_2^{(1)}, x_2^{(2)}, x_2^{(3)} \dots$ can be considered as the values of a r.v. X_2 which follows the law of X, then $X_2 \sim \mathfrak{B}(p)$
- The values $x_3^{(1)}, x_3^{(2)}, x_3^{(3)} \dots$ can be considered as the values of a r.v. X_3 which follows the law of X, then $X_n \sim \mathfrak{B}(p)$

Expected Value of F

➡ Pour chaque échantillon j, on calcule la pourcentage échantillonnal :

➡For each sample j, we calculate the sample pourcentage:

$$f^{(j)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(j)}$$

➡Chaque pourcentage échantillonnal $f^{(j)}$ (soit $f^{(1)}, f^{(2)}, \ldots$) est une combinaison linéaire de v.a. de Bernoulli de paramètre p (donc une binomiale de paramètres (n, p).

➡Each sample mean $\overline{x}^{(j)}$ (let $\overline{x}^{(1)}, \overline{x}^{(2)}, \ldots$) is a linear combination of r.v. with μ mean and σ^2 variance

➡Il s'agit de la valeur prise dans un échantillon par la variable aléatoire

➡This value is taken from a sample by the random variable :

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

➡Ainsi, la v.a. représentant la moyenne d'échantillon F est définie par :

➡Then, the r.v. representing the sample mean F is defined by :

$$F = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

With $X_i \sim \mathfrak{B}(p)$, for all $i = 1,2,\ldots,n$

Expected Value of F

➡L'espérance de cette variable aléatoire est alors :

➡The expectation of this random variable is then:

$$\mu_F = \mathbb{E}(F) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}np = p$$

➡La moyenne de la distribution d'échantillonnage des pourcentage de X est ainsi égale à la pourcentage de X sur la population. F est un estimateur sans biais de p

➡The mean of the pourcentages sampling distribution of the X is equal to the pourcentage of X over the population. F is an unbiased estimator of p

The DEP Standard Deviation $\sqrt{\mathbb{V}ar(F)}$

Standard Deviation

À présent, intéressons-nous à l'écart-type de la distribution d'échantillonnage des pourcentages :

Now, let's look at the standard deviation of the pourcentages sampling distribution

Pourcentage sampling	Effectifs n_i	$f_i - \mu_F$	$(f_i - \mu_F)^2$	$n_i(f_i - \mu_F)^2$
0/3	1	-6/15	36/225	36/225
1/3	6	-1/15	1/225	6/225
2/3	3	4/15	16/225	48/225
3/3	0	9/15	81/225	0/225
Total	10			90/225

$$\sigma_F^2 = \frac{1}{C_5^3} \sum_{i=1}^{C_5^3} n_i(f_i - \mu_F)^2 = \frac{1}{10} \times \frac{90}{225} = \frac{9}{225} = \frac{1}{25} = \left(\frac{1}{5}\right)^2$$

Standard Deviation

$$\sigma^2_{\overline{X}} = \mathbb{V}ar(\overline{X}) = \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\mathbb{V}ar\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p)$$

$$= \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

La variance de la distribution d'échantillonnage des pourcentages de X est égale à la variance de $X \sim \mathfrak{B}(p)$ sur la population à un coefficient $\frac{1}{n}$ près :

➡la variance échantillonnale diminue avec la taille de l'échantillon

The variance of the pourcentage sampling distribution X is equal to the variance of $X \sim \mathfrak{B}(p)$ over the population up to a coefficient $\frac{1}{n}$:

➡ sample variance decreases with sample size

Standard Deviation

En petit échantillon, on a

In a small population, we have

$$\sigma_F = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

On vérifie dans l'exemple des étudiants :

Let we check in the students example:

$$\sigma_F = \frac{\sqrt{0{,}4 \times 0{,}6}}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}}$$

$$= \frac{\sqrt{4 \times 6}}{\sqrt{10 \times 10 \times 3}}\sqrt{\frac{2}{4}}$$

$$= \sqrt{\frac{4}{100}} = \frac{2}{10} = 20\,\%$$

Standard Deviation : remarques

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 1$ lorsque la population est infinie, le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ peut être négligé

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 0$ lorsque la taille de l'échantillon devient grand et se rapproche de la taille de la population, le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ tend vers 0.

$\lim_{N \to +\infty} \sigma_{\overline{X}} = 0$ L'écart-type de la distribution d'échantillonnage des moyennes tends vers 0 lorsque la taille de l'échantillon augmente. Plus l'information accumulée sur la population est importante, plus la variation d'échantillonnage est faible.

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 1$ when the population is infinite, the correction factor $\sqrt{\frac{N-n}{N-1}}$ can be overlooked

$\lim_{N \to +\infty} \sqrt{\frac{N-n}{N-1}} = 0$ as the sample size becomes large and approaches the population size, the correction factor $\sqrt{\frac{N-n}{N-1}}$ tends to 0.

$\lim_{N \to +\infty} \sigma_{\overline{X}} = 0$ The standard deviation of the sampling distribution of the means tends towards 0 as the sample size increases. The greater the information accumulated on the population, the smaller the sampling variation.

Central limit theorem

Pour des échantillons aléatoires simples, la distribution d'échantillonnage de F peut être approchée par une distribution de probabilité normale, lorsque la taille de l'échantillon devient importante (en pratique, dès lors que $n \geq 30$).

For simple random samples, the sampling distribution of F can be approximated by a normal probability distribution, when the sample size becomes large (in practice, since $n \geq 30$).

$$F \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$