Chapitre 2. Distribution d’Échantillonnage

Khaled Jabeur

HDR (IHEC de Carthage, Tunisie), **Ph.D.** (Université Laval, Canada), **M.B.A** (Université Laval, Canada), **Ing.** (FST, Tunisie)

Professeur de l’enseignement supérieur à l’Institut Supérieur de Gestion (ISG) de Bizerte

Khaled.jabeur@uvt.tn

Estimer ne coûte presque rien,

Estimer incorrectement coûte cher.

Vieux proverbe chinois

Plan

Un peu d'histoire …

- Dans de nombreux domaines (scientifiques, économiques, épidémiologiques...), on a besoin de connaître certaines caractéristiques d'une population (ex. Salaire moyen des jeunes cadres dans les PME, Taux de chômage auprès des jeunes diplômés, …). Mais, en règle générale, ces caractéristiques sont difficiles à déterminer à cause du grand effectif des populations concernées. La solution consiste alors à estimer ces caractéristiques en se basant sur des observations d'un échantillon de taille plus réduit.

- L'idée de décrire une population à partir d'un échantillon réduit n'a été imaginée que dans la seconde moitié du XVIIIème siècle, notamment par l'école arithmétique politique anglaise. Ainsi, l'observation d'échantillons permettait d'éviter la complexité et le coût exorbitant de travailler sur des populations. Toutefois, ils se sont aperçu rapidement que les résultats manquaient d'exactitude : ils ne prenaient en considération ni la représentativité de l'échantillon, ni les fluctuations d'échantillonnage.

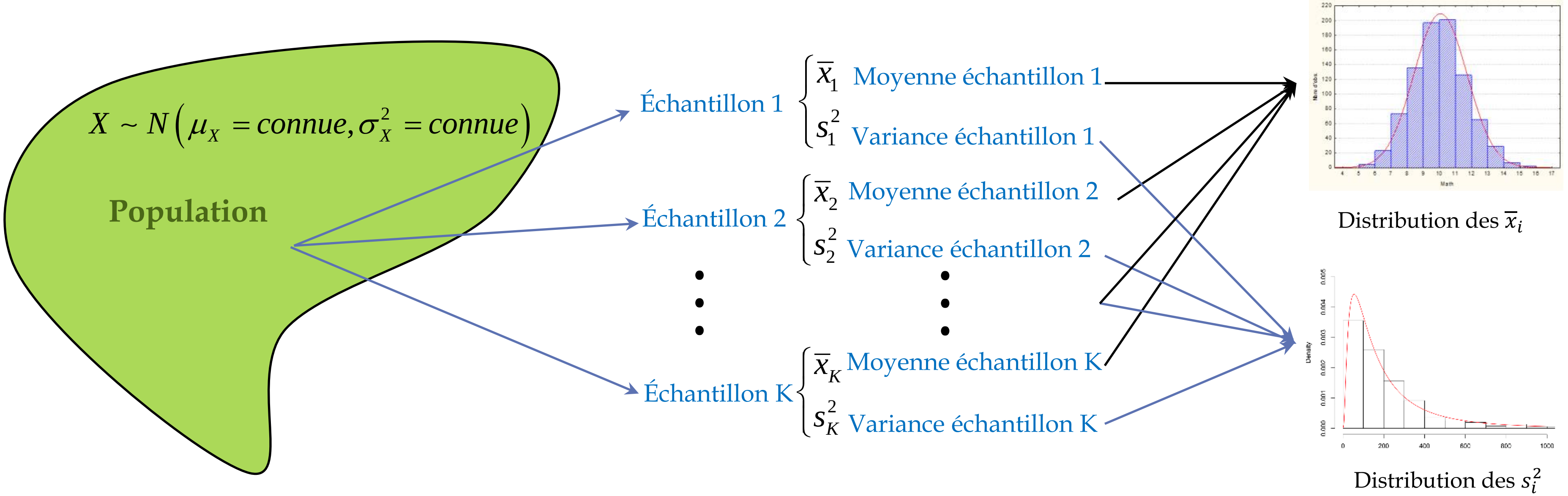
Cadre général

En effet, la question relative à la représentativité de l'échantillon fait partie de la théorie d'échantillonnage qui sera traitée dans le cadre du présent chapitre alors que la théorie de l'estimation (Chapitre 4) apportera des éléments de réponse au problème de fluctuations d'échantillonnage.

Théorie de l'échantillonnage : Définition

La théorie de l'échantillonnage étudie les liens entre les caractéristiques d'une population et celles des échantillons prélevés de cette population. En particulier, cette théorie suppose connaître la loi de probabilité (ainsi que ses paramètres) du caractère X étudié dans la population et se propose d'en déduire des renseignements sur des échantillons extraits de cette population (c'est purement théorique).

Exemples de questions posées en théorie de l'échantillonnage



Q1 :	Existe-t-il un lien entre μ_X (la moyenne du caractère X dans la population) et les $\bar{x}_i$ (les moyennes du caractère X dans les différents échantillons).
Q2 :	Existe-t-il un lien entre σ_X^2 (la variance du caractère X dans la population) et les s_i^2 (les variances du caractère X dans les différents échantillons).
Q3 :	Existe-t-il un lien entre la distribution de probabilité du caractère X dans la population (loi normale dans notre exemple) et la distribution de probabilité des $\bar{x}_i$ (les moyennes du caractère X dans les différents échantillons).
⋮	

Terminologie et définitions

- Une population se définit comme un ensemble d'éléments (individus, entreprises, dossiers, projets, ...) homogènes c'est-à-dire ayant des caractéristiques communes. On note par N la taille de la population.
- Un échantillon est tout sous-ensemble de la population. On note par n la taille de l'échantillon
- Un caractère ou une variable statistique c'est l'aspect que l'on désire étudier chez un élément (individus, entreprises, dossiers, projets, ...) de la population ou de l'échantillon.

Terminologie et définitions

Pour étudier les caractéristiques d'une population, on dispose de deux méthodes de collecte de données :

1. La **méthode exhaustive** ou recensement où chaque individu de la population est étudié selon le (ou les) caractère(s) d'intérêt.
2. La **méthode des sondages** ou échantillonnage qui conduit à n'examiner qu'une fraction (c'est-à-dire un échantillon) de la population.

Méthodes d'échantillonnage : Définition

Les méthodes d'échantillonnage : Ensemble des méthodes permettant de réaliser un sondage (de prélever un échantillon de données) au sein d'une population, de manière à reproduire un échantillon aussi représentatif que possible de cette population.

Échantillonnage : Raisons d'être

- On effectue l'échantillonnage essentiellement pour les raisons suivantes :

  1. Lorsque la population est infinie
  2. Par souci d'économie de coût
  3. Si le test est destructif
  4. Obtenir l'information le plus rapidement possible

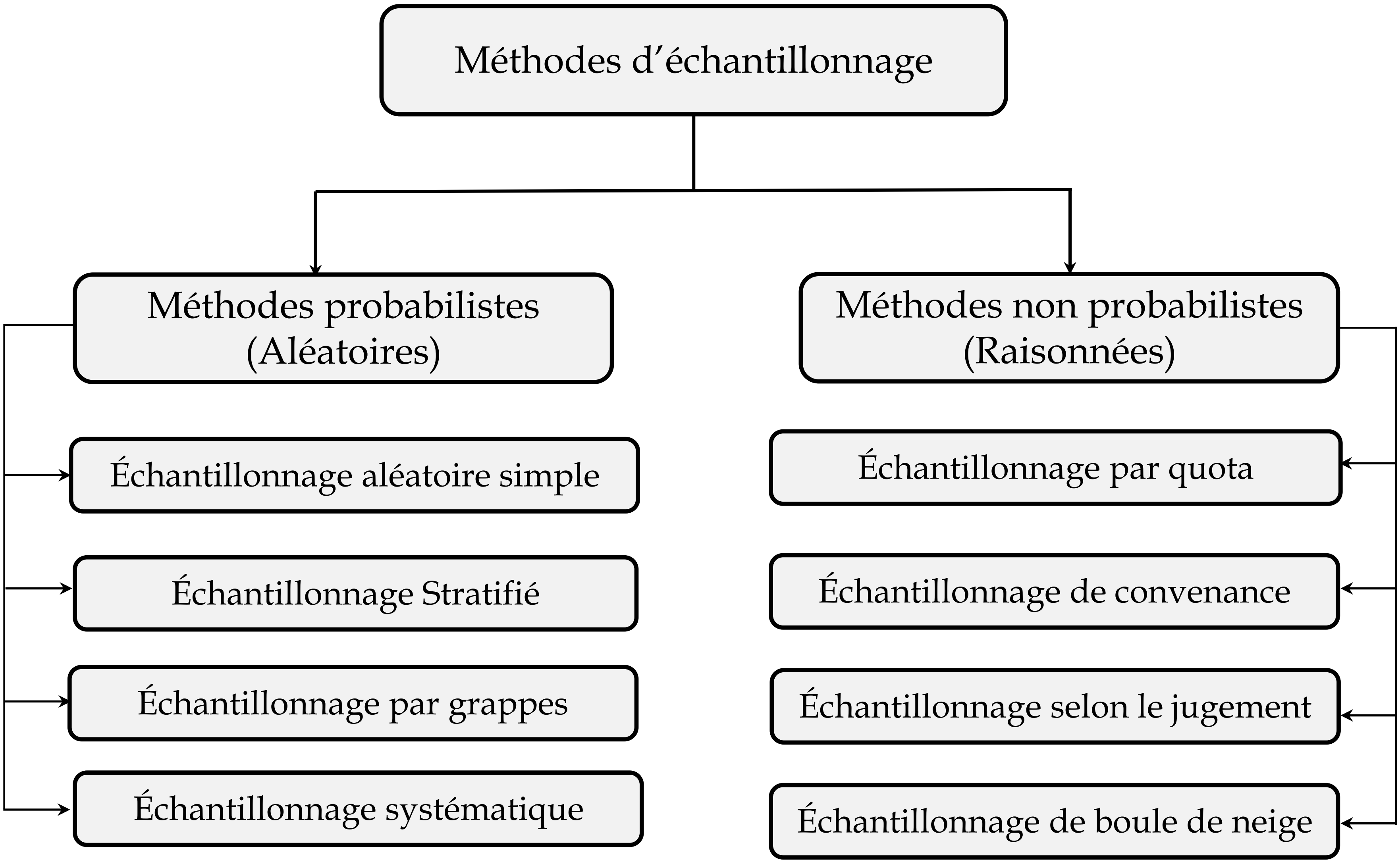
Échantillonnage : Familles de méthodes

- Les méthodes d'échantillonnage peuvent être regroupées en deux grandes familles :

    1. L'échantillonnage non aléatoire (ou non probabiliste) : L'analyste utilise son expérience et ses connaissances personnelles pour choisir parmi les unités de la population celles qui feront partie de l'échantillon et qui, à son avis, représentent adéquatement la population ;

    2. L'échantillonnage aléatoire (ou probabiliste) : Obtenu par l'intermédiaire d'un mécanisme probabiliste, de sorte que l'on sache à l'avance la probabilité (non nulle) qu'un individu quelconque de la population soit inclus dans l'échantillon.

Échantillonnage



Méthodes d'échantillonnage aléatoire simple

- Dans le cadre de ce cours on s'intéresse uniquement aux méthodes d'échantillonnage aléatoire simple.

- Définition : Dans une méthode d'échantillonnage aléatoire simple un échantillon est construit de telle sorte que chaque individu de la population ait la même probabilité d'être sélectionnée dans l'échantillon. On utilise souvent une table de nombres aléatoires pour s'assurer que le choix des individus s'effectue vraiment au hasard.

Tirage avec et sans remise

- La sélection des individus à inclure dans l'échantillon peut se faire selon deux procédures :
    - Procédure non exhaustif : On procède à n tirages successifs au hasard avec remise, c'est-à-dire que l'individu est remis dans la population après chaque tirage. Ainsi, avec cette procédure, **un individu peut être sélectionné plus qu'une fois dans l'échantillon**;
    - Procédure exhaustif : On procède à n tirages successifs au hasard sans remise ou à un tirage simultané de n individus. Ainsi, avec cette procédure, **un individu ne peut être sélectionné qu'au plus une seule fois dans l'échantillon.**

Échantillon aléatoire simple pour l'estimation des paramètres

- Soit X une variable aléatoire représentant un caractère que l'on désire étudier auprès des individus d'une population. Cette v.a. X est considérée comme parfaitement connue si on connaît la loi qu'elle suit ainsi que les différents paramètres (s'ils existent) de cette loi.

- Par exemple, si $X \sim N(-2,\ 4)$ on peut calculer n'importe quelle probabilité impliquant la v.a. X. De la même façon, si $Y \sim \mathcal{B}(5,\ 0.2)$ on peut calculer n'importe quelle probabilité impliquant la v.a. Y.

- En pratique, il arrive souvent qu'on connaît partiellement la loi d'une v.a. X, c'est-à-dire on connaît la nature de la loi (ex. Normale, Binomiale, Poisson, etc.) mais on ignore les paramètres de cette loi (ex. μ_X et σ_X^2 pour la loi Normale, λ pour la loi de Poisson, etc.).

Ce sont donc ces paramètres que l'on cherche à estimer.

Échantillonnage : Définition d'un paramètre

- Définition : Un paramètre est toute mesure caractéristique calculée sur la base des données de la population.
- Dans le cadre de ce chapitre, nous allons estimer les trois paramètres suivants :

μ_X	La moyenne du caractère X dans la population
σ_X^2	La variance du caractère X dans la population
p	La proportion d'individus ayant une certaine caractéristique dans la population

Exemples de paramètres, d'estimateurs et d'estimations

Paramètre	Estimateur possible (variable aléatoire)	Estimation (ou réalisation)
$E(X) = \mu_X$	$\bar{X}$ Moyenne échantillonnale (ou empirique)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$Var(X) = \sigma_X^2$	S^2 Variance échantillonnale (ou empirique)	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
p	$\hat{P}$ Proportion échantillonnale (ou empirique)	$\hat{p} = \frac{k}{n}$ où k est le nombre d'individus possédant la caractéristique étudiée dans l'échantillon

La variable aléatoire $\overline{X}$

Population (N individus)

Tirages (avec ou sans remise) de K échantillons, chacun de taille n

$$\text{Échantillon } 1 : \underbrace{\left(x_1^1, x_2^1, x_3^1, \dots, x_i^1, \dots, x_n^1\right)}_{\text{Valeurs observées dans l'échantillon}} \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^1 = \overline{x}_1$$

Moyenne du caractère X observée dans l'échantillon 1

$$\text{Échantillon } 2 : \left(x_1^2, x_2^2, x_3^2, \dots, x_i^2, \dots, x_n^2\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x}_2$$

Moyenne du caractère X observée dans l'échantillon 2

$$\text{Échantillon } 3 : \left(x_1^3, x_2^3, x_3^3, \dots, x_i^3, \dots, x_n^3\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^3 = \overline{x}_3$$

Moyenne du caractère X observée dans l'échantillon 3

$\vdots$

$$\text{Échantillon } j : \left(x_1^j, x_2^j, x_3^j, \dots, x_i^j, \dots, x_n^j\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^j = \overline{x}_j$$

Moyenne du caractère X observée dans l'échantillon j

$\vdots$

$$\text{Échantillon } K : \left(x_1^K, x_2^K, x_3^K, \dots, x_i^K, \dots, x_n^K\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^K = \overline{x}_K$$

Moyenne du caractère X observée dans l'échantillon K

- Les moyennes $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_K$ constituent des valeurs possibles d'une nouvelle variable aléatoire, qu'on note $\bar{X}$, appelée moyenne échantillonnale (ou empirique). Ainsi, la variable aléatoire $\bar{X}$ décrit les moyennes du caractère X observées dans les différents échantillons de taille n qu'on peut construire (avec ou sans remise) à partir d'une population de taille N.

La variable aléatoire S^2

Tirages (avec ou sans remise) de K échantillons, chacun de taille n

Population (N individus)

Valeurs observées dans l'échantillon

$$\text{Échantillon } 1 : \underbrace{\left(x_1^1, x_2^1, x_3^1, \ldots, x_i^1, \ldots, x_n^1\right)} \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^1 - \bar{x}_1\right)^2 = s_1^2$$

Variance du caractère X observée dans l'échantillon 1

$$\text{Échantillon } 2 : \left(x_1^2, x_2^2, x_3^2, \ldots, x_i^2, \ldots, x_n^2\right) \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - \bar{x}_2\right)^2 = s_2^2$$

Variance du caractère X observée dans l'échantillon 2

$$\text{Échantillon } 3 : \left(x_1^3, x_2^3, x_3^3, \ldots, x_i^3, \ldots, x_n^3\right) \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^3 - \bar{x}_3\right)^2 = s_3^2$$

Variance du caractère X observée dans l'échantillon 3

$\vdots$

$$\text{Échantillon } j : \left(x_1^j, x_2^j, x_3^j, \ldots, x_i^j, \ldots, x_n^j\right) \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^j - \bar{x}_j\right)^2 = s_j^2$$

Variance du caractère X observée dans l'échantillon j

$\vdots$

$$\text{Échantillon } K : \left(x_1^K, x_2^K, x_3^K, \ldots, x_i^K, \ldots, x_n^K\right) \Rightarrow \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^K - \bar{x}_K\right)^2 = s_K^2$$

Variance du caractère X observée dans l'échantillon K

- Les variances s_1^2 , s_2^2, s_3^2, …, s_K^2 constituent des valeurs possibles d'une nouvelle variable aléatoire, qu'on note S^2, appelée variance échantillonnale (ou empirique). Ainsi, la variable aléatoire S^2 décrit les variances du caractère X observées dans les différents échantillons de taille n qu'on peut construire (avec ou sans remise) à partir d'une population de taille N.

La variable aléatoire $\hat{P}$

Tirages (avec ou sans remise) de K échantillons, chacun de taille n

Valeurs observées dans l'échantillon : $x_i^j = \begin{cases} 1 & si\ l'individu\ i\ de\ l'échantillon\ j\ possède\ la\ caractéristique\ étudiée \\ 0 & si\ non \end{cases}$

Population (N individus)

$$\text{Échantillon } 1 : \left(x_1^1, x_2^1, x_3^1, \ldots, x_i^1, \ldots, x_n^1\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^1 = \frac{k_1}{n} = \hat{p}_1$$

Proportion des individus ayant la caractéristique étudiée dans l'échantillon 1

$$\text{Échantillon } 2 : \left(x_1^2, x_2^2, x_3^2, \ldots, x_i^2, \ldots, x_n^2\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{k_2}{n} = \hat{p}_2$$

Proportion des individus ayant la caractéristique étudiée dans l'échantillon 2

$$\text{Échantillon } 3 : \left(x_1^3, x_2^3, x_3^3, \ldots, x_i^3, \ldots, x_n^3\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^3 = \frac{k_3}{n} = \hat{p}_3$$

Proportion des individus ayant la caractéristique étudiée dans l'échantillon 3

$$\vdots$$

$$\text{Échantillon } j : \left(x_1^j, x_2^j, x_3^j, \ldots, x_i^j, \ldots, x_n^j\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^j = \frac{k_j}{n} = \hat{p}_j$$

Proportion des individus ayant la caractéristique étudiée dans l'échantillon j

$$\vdots$$

$$\text{Échantillon } K : \left(x_1^K, x_2^K, x_3^K, \ldots, x_i^K, \ldots, x_n^K\right) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^K = \frac{k_K}{n} = \hat{p}_K$$

Proportion des individus ayant la caractéristique étudiée dans l'échantillon K

- Notons que k_j désigne le nombre d'individus ayant la caractéristique étudiée dans l'échantillon j (j=1..K).
- Les proportions $\hat{p}_1, \hat{p}_2, \hat{p}_3, \ldots, \hat{p}_K$ constituent des valeurs possibles d'une nouvelle variable aléatoire, qu'on note $\hat{P}$, appelée proportion échantillonnale (ou empirique). Ainsi, la variable aléatoire $\hat{P}$ décrit les proportions des individus ayant la caractéristique étudiée dans les différents échantillons de taille n qu'on peut construire (avec ou sans remise) à partir d'une population de taille N.

Distribution d'échantillonnage

- Puisque $\bar{X}$, S^2 et $\hat{P}$ sont des variables aléatoires, alors on pourra s'intéresser à leur distribution de probabilité, appelée distribution d'échantillonnage.
- Dans le cadre de ce chapitre, les distributions d'échantillonnage (ou distribution de probabilité) des trois variables aléatoires suivantes seront étudiées :
  - Distribution d'échantillonnage de la moyenne $\bar{X}$
  - Distribution d'échantillonnage de la variance S^2
  - Distribution d'échantillonnage de la proportion $\hat{P}$
- Ainsi, pour chacune de ces trois statistiques, on calculera tout d'abord l'espérance et la variance : $\mathrm{E}(\bar{X})$, $\mathrm{Var}(\bar{X})$, $\mathrm{E}(S^2)$, $\mathrm{Var}(S^2)$, $\mathrm{E}(\hat{P})$ et $\mathrm{Var}(\hat{P})$. Ensuite, on discutera de la nature de la (ou des) loi(s) suivi par chacune de ces trois variables aléatoires.

Distribution d'échantillonnage de la moyenne $\bar{X}$

- Si $\bar{X}$ est la moyenne échantillonnale, alors on a :

1. $E\left(\bar{X}\right)=E\left(X\right)=\mu_X$ (On dit que $\bar{X}$ est un estimateur sans biais de μ_X)
2. $Var\left(\bar{X}\right)=\dfrac{Var\left(X\right)}{n}=\dfrac{\sigma_X^2}{n}$

- Exercice d'application : Une population est constituée de 4 étudiants en statistique (le faible effectif n'est pas dû à un manque d'intérêt pour la matière de la part des étudiants mais au désir de ne pas multiplier inutilement les calculs qui vont suivre !). Leur professeur s'intéresse au temps hebdomadaire consacré par chaque étudiant à l'étude des statistiques. Il a obtenu les résultats suivants :

Étudiant	Temps d'étude (en heures)
A	8
B	3
C	6
D	11
Total	**28**

1. Calculer la moyenne μ_X et la variance σ_X^2 du temps hebdomadaire consacré par chaque étudiant à l'étude des statistiques (c'est le caractère X étudié) de cette population de 4 étudiants.

2. Calculer $\mathrm{E}(\bar{X})$ et $\mathrm{Var}(\bar{X})$ sachant que les échantillons sont de taille 2 et que le tirage a été effectué avec remise.

- Réponse :

1. $$E(X)=\mu_X=\sum_{i=1}^{N}x_i\times P(X=x_i)=\frac{1}{N}\times\sum_{i=1}^{N}x_i=\frac{8+3+6+11}{4}=7$$

$$Var(X)=\sigma_X^2=E(X^2)-E(X)^2=\sum_{i=1}^{N}x_i^2\times P(X=x_i)-7^2=$$

$$=\left[\frac{8^2+3^2+6^2+11^2}{4}\right]-49=8.5$$

2. Nombre d'échantillons de taille 2 = 16 (Tirage avec remise) $N^n=4^2=16$

Numéro de l'échantillon	L'échantillon	Observations de l'échantillon	Moyenne de l'échantillon	Numéro de l'échantillon	L'échantillon	Observations de l'échantillon	Moyenne de l'échantillon
1	A, A	8, 8	8	**9**	C, A	6, 8	7
2	A, B	8, 3	5.5	**10**	C, B	6, 3	4.5
3	A, C	8, 6	7	**11**	C, C	6, 6	6
4	A, D	8, 11	9.5	**12**	C, D	6, 11	8.5
5	B, A	3, 8	5.5	**13**	D, A	11, 8	9.5
6	B, B	3, 3	3	**14**	D, B	11, 3	7
7	B, C	3, 6	4.5	**15**	D, C	11, 6	8.5
8	B, D	3, 11	7	**16**	D, D	11, 11	11

- Réponse (suite …) :

2.

Valeurs possibles de la v.a. $\bar{X}$	Fréquence de chaque valeur possible de $\bar{X}$	Probabilité à associer à chaque valeur possible de $\bar{X}$
3	1	0.0625
4.5	2	0.125
5.5	2	0.125
6	1	0.0625
7	4	0.25
8	1	0.0625
8.5	2	0.125
9.5	2	0.125
11	1	0.0625
Total	**16**	**1**

$$E(\bar{X}) = \sum_{i=1}^{9} x_i \times P(\bar{X} = x_i) = \left(3\times 0.0625 + 4.5\times 0.125 + \ldots + 11\times 0.0625\right) = 7 = E(X)$$

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = \sum_{i=1}^{9} x_i^2 \times P(\bar{X} = x_i) - 7^2 =$$

$$= \left(3^2\times 0.0625 + 4.5^2\times 0.125 + \ldots + 11^2\times 0.0625\right) - 49 = 53.25 - 49 = 4.25 = \frac{8.5}{2} = \frac{Var(X)}{n}$$

Distribution de probabilité de la variable aléatoire $\overline{X}$

(Tirage avec remise ou population infinie)

Taille de l'échantillon	Loi de X	Variance de X σ_X^2	Loi de $\bar{X}$	Loi réduite
$n \geq 30$	Quelconque	connue	$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \sim N(0,1)$
		inconnue	$\bar{X} \sim N\left(\mu_X, \frac{s^2}{n}\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{s}{\sqrt{n}}\right)} \sim N(0,1)$
$n < 30$	Quelconque		inconnue	
	Normale $X \sim N(\mu_X, \sigma_X^2)$	connue	$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \sim N(0,1)$
		inconnue		$\frac{\bar{X}-\mu_X}{\left(\frac{s}{\sqrt{n}}\right)} \sim t(n-1)dl$

Distribution d'échantillonnage (ou de probabilité) de la statistique $\bar{X}$

(Tirage sans remise ou population finie)

Taille de l'échantillon	Loi de X	Variance de X σ_X^2	Loi de $\bar{X}$	Loi réduite
$n \geq 30$	Quelconque	connue	$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n} \times \left[\frac{N-n}{N-1}\right]\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1)$
		inconnue	$\bar{X} \sim N\left(\mu_X, \frac{s^2}{n} \times \left[\frac{N-n}{N-1}\right]\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1)$
$n < 30$	Quelconque		inconnue	
	Normale $X \sim N(\mu_X, \sigma_X^2)$	connue	$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n} \times \left[\frac{N-n}{N-1}\right]\right)$	$\frac{\bar{X}-\mu_X}{\left(\frac{\sigma_X}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1)$
		inconnue		$\frac{\bar{X}-\mu_X}{\left(\frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right)} \sim t(n-1)dl$

- Remarque : Lorsqu'on dispose d'une population de taille N, le nombre d'échantillons de taille n qu'on peut construire à partir de cette population varie selon le mode de tirage, soit :

Le nombre d'échantillons lorsqu'on effectue des tirages avec remise $= N^n$

Le nombre d'échantillons lorsqu'on effectue des tirages sans remise $= C_N^n = \dfrac{N!}{n!(N-n)!}$

- Exemple 1: Supposons que les tailles des individus dans une population (caractère X) suivent une distribution normale de moyenne μ_X = 170 cm et de variance σ_X^2 = 25 cm^2. On tire avec remise un échantillon de taille 25 de cette population.

  Quelle est la probabilité pour que la taille moyenne dans l'échantillon soit supérieure à 172 cm ?

- Réponse : La taille moyenne dans l'échantillon est décrite par la variable aléatoire $\bar{X}$ et par conséquent, il nous demande de calculer la probabilité suivante :

  $$P\left(\bar{X} \geq 172\right) = ?$$

  Pour calculer cette probabilité, il faut tout d'abord déterminer la loi de probabilité de $\bar{X}$ ainsi que ses différents paramètres. En effet, selon le premier tableau (page 32), on a :

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \sim N\left(170, \frac{25}{25} = 1\right)$$

$$P\left(\bar{X} \geq 172\right) = P\left(Z \geq \frac{172 - 170}{\sqrt{1}}\right) = P\left(Z \geq 2\right) = 0.023$$

- **Exemple 2 :** Supposons que les tailles des individus dans une population de moyenne μ_X = 185 cm et de variance σ_X^2 inconnue. On tire avec remise un échantillon de taille 36 de cette population.

  Sachant que la variance de cet échantillon s^2 = 40 cm^2, quelle est la probabilité pour que la taille moyenne dans l'échantillon soit supérieure à 187 cm ?

- **Réponse :** La taille moyenne dans l'échantillon est décrite par la variable aléatoire $\bar{X}$ et par conséquent, il nous demande de calculer la probabilité suivante :

  $$P\left(\bar{X} \geq 187\right) = ?$$

  Pour calculer cette probabilité, il faut tout d'abord déterminer la loi de probabilité de $\bar{X}$ ainsi que ses différents paramètres. En effet, selon le premier tableau (page 32), on a :

  $$\bar{X} \sim N\left(\mu_X, \frac{s^2}{n}\right) \sim N\left(185, \frac{40}{36} = 1.11\right)$$

  $$P\left(\bar{X} \geq 187\right) = P\left(Z \geq \frac{187-185}{\sqrt{1.11}}\right) = P\left(Z \geq 1.90\right) = 0.029$$

- Exemple 3 : Supposons que les tailles des individus dans une population suivent approximativement une loi normale de moyenne $\mu_X = 185$ cm et de variance σ_X^2 inconnue. On tire avec remise un échantillon de taille 20 de cette population.

Sachant que la variance de cet échantillon $s^2 = 60$ cm^2, quelle est la probabilité pour que la taille moyenne dans l'échantillon soit supérieure à 188 cm ?

- Réponse : La taille moyenne dans l'échantillon est décrite par la variable aléatoire $\bar{X}$ et par conséquent, il nous demande de calculer la probabilité suivante :

$$P\left(\bar{X} \geq 188\right) = ?$$

Pour calculer cette probabilité, il faut tout d'abord déterminer la loi de probabilité de $\bar{X}$ ainsi que ses différents paramètres. En effet, selon le premier tableau (page 32), on a :

$$T = \frac{\bar{X} - \mu_X}{\left(\frac{s}{\sqrt{n}}\right)} \sim t\left(n-1\right) dl \sim t\left(19\ dl\right)$$

$$P\left(\bar{X} \geq 188\right) = P\left(T \geq \frac{188-185}{\left(\frac{\sqrt{60}}{\sqrt{20}}\right)}\right) = P\left(T \geq 1.73\right) = 0.05$$

Distribution d'échantillonnage de la variance S^2

- Soit X une variable aléatoire et S^2 la variance échantillonnale, alors on a :

1. $E\left(S^2\right)=Var\left(X\right)=\sigma_X^2$ (On dit que S^2 est un estimateur sans biais de σ_X^2)

2. $Var\left(S^2\right)=\sigma_X^4\left(\frac{2}{n-1}+\frac{\beta_2-3}{n}\right) avec\ \beta_2=\frac{\mu_4}{\sigma_X^4}\ et\ \mu_4=E\left(\left(X-E\left(X\right)\right)^4\right)$

La formule (3) est rarement utilisée en pratique. Toutefois, si la v.a. X suit une loi normale, alors la formule de Var(S^2) peut être estimée par une formule beaucoup plus simple (3^{bis}).

2^{bis}. $Var(S^2)\simeq\frac{2\sigma_X^4}{n-1}\quad si\ X\sim Normale$

- Exercice d'application : Reprenons l'exemple de la population de 4 étudiants et leur temps hebdomadaire consacré à l'étude des statistiques.

Étudiant	Temps d'étude (en heures)
A	8
B	3
C	6
D	11
Total	**28**

En utilisant des échantillons de taille 2 tirés avec remise, on vous demande de calculer $E(S^2)$ (σ_X^2 = 8.5 calculée précédemment).

- Réponse :

Numéro de l'échantillon	L'échantillon	Observations de l'échantillon	Variance s^2	Numéro de l'échantillon	L'échantillon	Observations de l'échantillon	Variance s^2
1	A, A	8, 8	0	**9**	C, A	6, 8	2
2	A, B	8, 3	12.5	**10**	C, B	6, 3	4.5
3	A, C	8, 6	2	**11**	C, C	6, 6	0
4	A, D	8, 11	4.5	**12**	C, D	6, 11	12.5
5	B, A	3, 8	12.5	**13**	D, A	11, 8	4.5
6	B, B	3, 3	0	**14**	D, B	11, 3	32
7	B, C	3, 6	4.5	**15**	D, C	11, 6	12.5
8	B, D	3, 11	32	**16**	D, D	11, 11	0

Valeurs possibles de la v.a. S^2	Fréquence des valeurs possibles de S^2	Probabilité à associer à chaque valeur possible de S^2
0	4	0.25
2	2	0.125
4.5	4	0.25
12.5	4	0.25
32	2	0.125
Total	**16**	**1**

$$E(S^2) = \sum_{i=1}^{5} x_i \times P(S^2 = x_i) =$$
$$= (0 \times 0.25 + 2 \times 0.125 + \ldots + 32 \times 0.125) = 8.5$$
$$= \sigma_X^2$$

Distribution d'échantillonnage (ou de probabilité) de la variable aléatoire S^2

Théorème

Soit X une variable aléatoire qui suit une loi normale $N(\mu_X,\ \sigma_X^2)$ et S^2 est la variance échantillonnale. On a alors la variable aléatoire :

$$Y = \frac{(n-1)S^2}{\sigma_X^2}$$

suit une loi du Khi-deux à $(n\text{-}1)$ *dl*.

- Exercice d'application : On fait l'hypothèse que la taille (en cm) des étudiants d'une école de génie est une variable aléatoire normale X de moyenne $\mu_X = 175$ et de variance $\sigma_X^2 = 100$. Un échantillon de taille 51 est sélectionné de cette population. Quelle est la probabilité que la variance des tailles dans l'échantillon soit d'au plus égale 112.66.

- Réponse : La variable aléatoire qui décrit la variance du caractère X (taille des étudiants d'une école de génie) dans l'échantillon est la variance échantillonnale S^2. Ainsi, la probabilité qu'on nous demande de calculer est :

$$P\left(S^2 \leq 112.66\right) = P\left(\frac{(n-1)S^2}{\sigma_X^2} \leq \left(\frac{(51-1)}{100}\right) 112.66\right) =$$

$$= P\left(Y \leq 56.33\right) \; avec \; Y \sim \chi^2\left(51-1=50\right) dl$$

$$= 1 - P\left(Y \geq 56.33\right) = 1 - 0.25 = 0.75$$

Distribution d'échantillonnage (ou de probabilité) de la variable aléatoire $\hat{P}$

Distribution d'échantillonnage de $\hat{P}$

Si $\hat{P}$ est la proportion échantillonnale, alors on a :

1. $E\left(\hat{P}\right) = p$
2. $Var\left(\hat{P}\right) = \dfrac{p\left(1-p\right)}{n}$ / $Var\left(\hat{P}\right) = \dfrac{p\left(1-p\right)}{n}\left(\dfrac{N-n}{N-1}\right)$

Tirage avec remise — Tirage sans remise

Distribution d'échantillonnage (ou de probabilité) de la statistique $\hat{P}$

Distribution d'échantillonnage de $\hat{P}$

Si $n \geq 30$, $n \times p \geq 5$ et $n \times (1 - p) \geq 5$ alors :

1. $\hat{P} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$ Tirage avec remise
2. $\hat{P} \sim N\left(p, \dfrac{p(1-p)}{n}\left(\dfrac{N-n}{N-1}\right)\right)$ Tirage sans remise

- Exercice d'application : Un fabricant de clous a déterminé que 3% des clous produits sont défectueux. On étudie un échantillon aléatoire de 300 clous (tirage avec remise). Quelle est la probabilité que la proportion de clous défectueux dans l'échantillon soit comprise entre 2% et 3,5% ?

- Réponse : la variable aléatoire qui décrit la proportion des clous défectueux dans l'échantillon est $\hat{P}$. Puisque n = 300 ≥ 30, $n \times p$ = 300×0.03 = 9 ≥ 5 et $n \times (1 - p)$ = 300 × 0.97 = 291 ≥ 5, alors on peut déduire la distribution de probabilité de $\hat{P}$, soit :

$$\hat{P} \sim N\left(p = 0.03, \frac{p(1-p)}{n} = 0.000097 \right)$$

- Ainsi, la probabilité qu'on nous demande de calculer est :

$$P\left(0.02 \leq \hat{P} \leq 0.035\right) = P\left(\frac{0.02 - 0.03}{\sqrt{0.000097}} \leq Z \leq \frac{0.035 - 0.03}{\sqrt{0.000097}} \right)$$

$$= P\left(-1.02 \leq Z \leq 0.51\right) = P\left(Z \geq -1.02\right) - P\left(Z \geq 0.51\right)$$

$$= 1 - P\left(Z \geq 1.02\right) - P\left(Z \geq 0.51\right) = 1 - 0.154 - 0.305 = 0.541$$