Cours de Statistiques

Focus sur le Chapitre 3: Echantillonnage

Licence 3 – Parcours Gestion et Finance

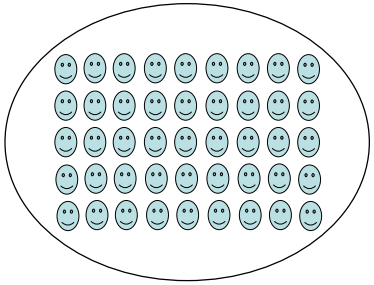
Stéphane ROBIN
Université Paris 1 Panthéon-Sorbonne
Département de Gestion – EM Sorbonne

Contenu du Chapitre 3

1. L'échantillonnage aléatoire
2. Distribution de la moyenne de l'échantillon
3. Forme de la distribution d'échantillonnage
4. Les proportions comme cas particulier de la moyenne
5. L'échantillonnage dans le cas d'une petite population
6. L'échantillonnage en pratique: exemple d'application.

1. De la population à l'échantillon

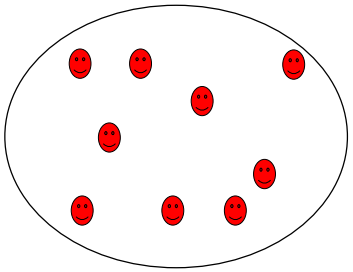
Population de référence



On s'intéresse au caractère X, considéré comme une VA.

X suit une loi $p(x)$ a priori inconnue, d'espérance m et de variance σ^2 inconnues.

Echantillon tiré au sort



Les observations $x_1, \ldots, x_n$ sont des réalisations de X. Chaque xi peut être considéré comme la réalisation d'une VA X_i de même loi $p(x)$ que X.

Echantillon aléatoire simple

On appelle **échantillon aléatoire simple (EAS)** un échantillon dont les n observations $x_1, \ldots, x_n$ sont des réalisations de X considérées comme les réalisations de n VA **indépendantes** $X_1, \ldots, X_n$ de même loi que X.

Ainsi, si la VA X suit la loi $p(x)$ dans la population, la loi $p_i(x)$ de chaque X_i sera **identiquement** égale à $p(x)$ pour tout x, ce qu'on note:

$$p_1(x) \equiv p_2(x) \equiv \ldots \equiv p_n(x) \equiv p(x)$$

Les X_i sont dites *indépendamment et identiquement distribuées* (i.i.d.).

Moyenne de la population et moyenne de l'échantillon

La moyenne m du caractère X dans la population est donnée par l'espérance de X, c'est-à-dire $E(X) = m$.

Dans l'échantillon aléatoire simple $\{X_1, \ldots, X_n\}$, la moyenne empirique "X barre" est plus proche de m que la plupart des valeurs des X_i.

On peut illustrer cette propriété à l'aide d'un exemple, où l'on considère une petite population dont tous les éléments sont connus (observés).

Illustration (1)

Soit une population de 100 étudiants (par exemple, les 100 étudiants inscrits en L3 Finance à l'EM Sorbonne une année donnée) dont on connaît l'âge. On calcule l'âge moyen m:

Distribution de la population			$xp(x)$ pour calcul de la moyenne
Age X (en années)	Effectifs	Fréquence ou probabilité $p(x)$	
19	1	0.01	0.19
20	6	0.06	1.20
21	24	0.24	5.04
22	38	0.38	8.36
23	24	0.24	5.52
24	6	0.06	1.44
25	1	0.01	0.25
	N = 100	**1.00**	**m = 22**

Illustration (2a)

Dans cette population de 100 étudiants, on tire au sort un échantillon aléatoire simple de $n = 5$ étudiants. On trouve les valeurs suivantes: x_1=20, x_2=21, x_3=22, x_4=22, x_5=24.

On calcule: $\overline{X} = 21.8$

La valeur 21.8 est plus proche de celle de $m = 22$ que la plupart des valeurs x_i, $i = 1, \ldots, 5$. En effet: $\left|m - \overline{X}\right| = 22 - 21.8 = 0.2$

Alors que: $|x_1 - m| = 2$
$|x_2 - m| = 1$
$|x_3 - m| = |x_4 - m| = 0$
$|x_5 - m| = 2$

Illustration (2b)

Le tirage au sort d'un second échantillon de 5 étudiants donne les valeurs : x_1=20, x_2=21, x_3=22, x_4=23, x_5=23.

On calcule: $\overline{X} = 21.8$

La valeur 21.8 est plus proche de celle de $m = 22$ que la plupart des valeurs x_i, $i = 1, \ldots, 5$. En effet: $\left|m - \overline{X}\right| = 22 - 21.8 = 0.2$

Alors que: $|x_1 - m| = 2$
$|x_2 - m| = 1$
$|x_3 - m| = 0$
$|x_4 - m| = 1$
$|x_5 - m| = 1$

Illustration (2c)

Le tirage au sort d'un dernier échantillon de 5 étudiants donne les valeurs : x_1=20, x_2=21, x_3=22, x_4=23, x_5=24.

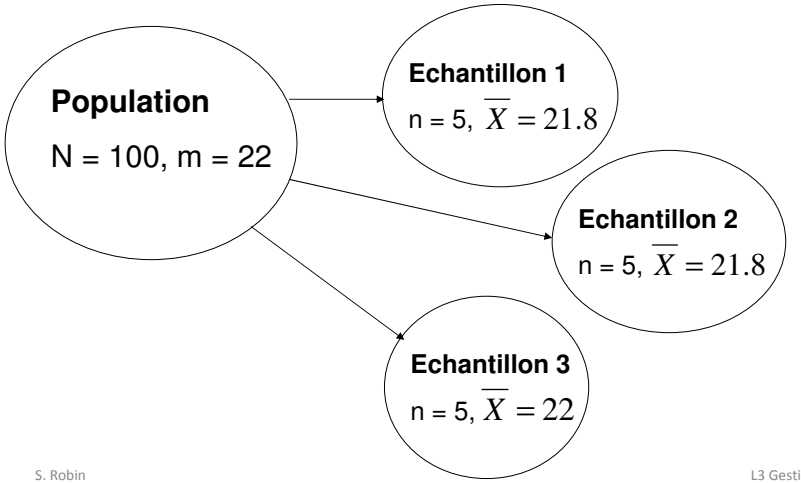
On calcule: $\overline{X} = 22$

La valeur 22 est plus proche de celle de m = 22 que la plupart des valeurs x_i, i = 1, …, 5. En effet: $|m - \overline{X}| = 22 - 22 = 0$

Alors que: $|x_1 - m| = 2$
$|x_2 - m| = 1$
$|x_3 - m| = 0$
$|x_4 - m| = 1$
$|x_5 - m| = 2$

Illustration (3)

On a tiré au sort trois échantillons (dont les individus ne sont pas forcément les mêmes) dans la même population.



La moyenne empirique comme VA

Comme on l'a vu en tirant au sort trois échantillons successifs, la valeur de "X barre" peut varier d'un échantillon à l'autre (tout en restant toujours "proche" de la valeur de m).

Comme la valeur de "X barre" résulte d'un tirage aléatoire, et varie d'un échantillon à l'autre, on peut considérer que la moyenne empirique "X barre" est elle-même une VA.

Il convient donc de s'intéresser à sa distribution, et à ses propriétés.

2. Distribution de la moyenne empirique

Considérons la moyenne empirique "X barre" comme une VA. Sa loi est a priori inconnue, mais on sait (par définition) que

$$\overline{X} = \frac{1}{n}(X_1 + ... + X_n)$$

où tous les X_i sont de même loi que la VA X. Cela implique que $\forall i$, X_i a la même espérance (et la même variance) que X :

$E(X_1) = \ldots = E(X_n) = E(X) = m$

$Var(X_1) = \ldots = Var(X_n) = Var(X) = \sigma^2$

On peut utiliser ces propriétés pour calculer l'espérance et la variance de la moyenne empirique "X barre".

Espérance de la moyenne empirique

On veut calculer l'espérance de la VA "X barre". En utilisant les propriétés de l'espérance, il vient:

$$E(\overline{X}) = E\left[\frac{1}{n}(X_1 + ... + X_n)\right] = \frac{1}{n}E(X_1 + ... + X_n)$$
$$= \frac{1}{n}\left[E(X_1) + ... + E(X_n)\right]$$
$$= \frac{1}{n}\left[m + ... + m\right] = \frac{1}{n}n.m$$
$$= m$$

L'espérance de la moyenne empirique est donc égale à la moyenne de la population.

Variance de la moyenne empirique

La variance de la moyenne empirique indique comment celle-ci varie autour de la moyenne de la population. En utilisant les propriétés de la variance, il vient:

$$Var(\overline{X}) = Var\left[\frac{1}{n}(X_1 + ... + X_n)\right] = \frac{1}{n^2}Var(X_1 + ... + X_n)$$
$$= \frac{1}{n^2}\left[Var(X_1) + ... + Var(X_n)\right]$$
$$= \frac{1}{n^2}\left[\sigma^2 + ... + \sigma^2\right] = \frac{1}{n^2}n.\sigma^2 = \frac{\sigma^2}{n}$$

NB : les X_i sont indépendantes deux à deux, donc la formule ci-dessus ne fait pas intervenir la covariance!

Moments de la moyenne empirique

Les principaux moments de la moyenne empirique sont donc:

Espérance : $E(\overline{X}) = m$

Variance : $Var(\overline{X}) = \frac{\sigma^2}{n}$

Ecart-type : (appelé également "écart-type d'échantillon") $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

On remarque que la dispersion autour de *m* diminue à mesure que *n* augmente (l'écart-type d'échantillon diminue)

3. Forme de la distribution d'échantillonnage

On a déterminé au point 2 l'espérance et l'écart-type de la moyenne empirique "X barre".

Il reste à déterminer la forme de la distribution de "X barre", également appelée "distribution d'échantillonnage"

On peut montrer que, **si** la distribution de X dans la population est normale **OU si** la taille de l'échantillon est grande (en général, à partir de n>30), **alors** la distribution d'échantillonnage a une forme approximativement normale.

Illustration (1)

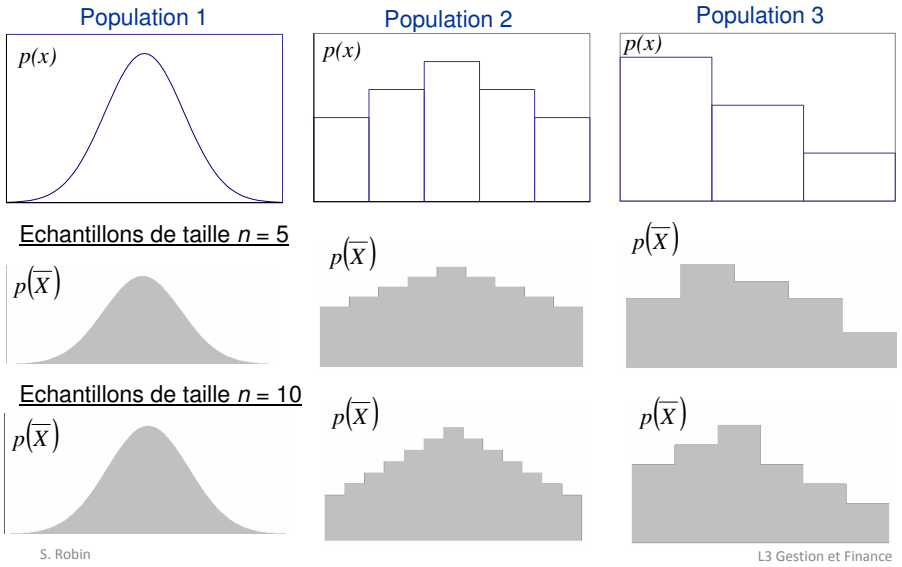
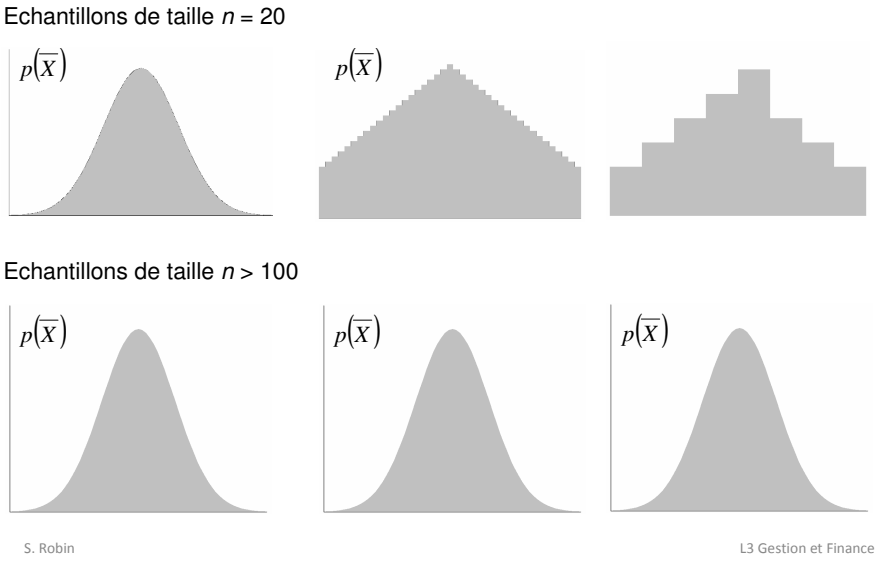


Illustration (2)



Règle de l'approximation normale

Cette règle, qui résume tout ce qui a été vu jusqu'à présent, peut s'énoncer ainsi :

Soit X une VA de loi inconnue, dont l'espérance et la variance dans une population de taille N sont respectivement m et σ.

Dans des EAS de taille n, la moyenne de l'échantillon "X barre" varie autour de la moyenne de la population m avec un écart-type égal à $\sigma/\sqrt{n}$.

Quand n augmente, la distribution d'échantillonnage (de "X barre") est de plus en plus concentrée autour de m, et devient de plus en plus proche de la distribution normale.

Cette règle n'est qu'une autre façon d'énoncer le Théorème Central Limite rencontré dans le Chapitre 2.

4. Le cas particulier des proportions

On conçoit généralement une proportion comme une fréquence. On oublie alors qu'une proportion est aussi une moyenne.

Il s'agit en effet de la moyenne des réalisations de n VA de Bernoulli $X_1, \ldots, X_n$ de même loi $\boldsymbol{B}(1, p)$. Par définition, $\forall i$, $X_i = 1$ (avec la probabilité p) ou 0 (avec la probabilité $1-p$). On a alors, pour k succès parmi les n épreuves :

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}(1+0+0+1+\ldots+1) = \frac{k}{n}$$

→ Proportion !

Ces n VA peuvent être tirées parmi une population de taille N. Si on tirait un autre échantillon, on pourrait obtenir une autre valeur de "X barre" (en fonction du nombre de succès k). La moyenne "X barre" est donc elle-même une VA.

Espérance et variance d'une proportion

La somme de n VA de Bernoulli (de même paramètre p) est une VA binomiale $Y \sim B(n, p)$. On sait alors que $E(Y) = np$.

Comme $Y = X_1 + ... + X_n$, on peut écrire $\overline{X} = \frac{1}{n} Y$

On a donc $$E(\overline{X}) = E\left(\frac{1}{n} Y\right) = \frac{1}{n} E(Y) = p$$

On sait de même que $Var(X) = np(1 - p)$, et on peut écrire:

$$Var(\overline{X}) = Var\left(\frac{1}{n} Y\right) = \frac{1}{n^2} Var(Y) = \frac{p(1-p)}{n}$$

Moments d'une proportion

En résumé, une proportion dans un échantillon de taille n est la moyenne des réalisations de n variables de Bernoulli $\boldsymbol{B}(1, p)$.

Cette moyenne (particulière car les variables ne prennent que les valeurs 0 ou 1) est elle-même une VA, notée ici $\overline{X}$, et plus souvent $f_n = Y/n = (X_1+...+X_n)/n = (\sum_i X_i)/n$

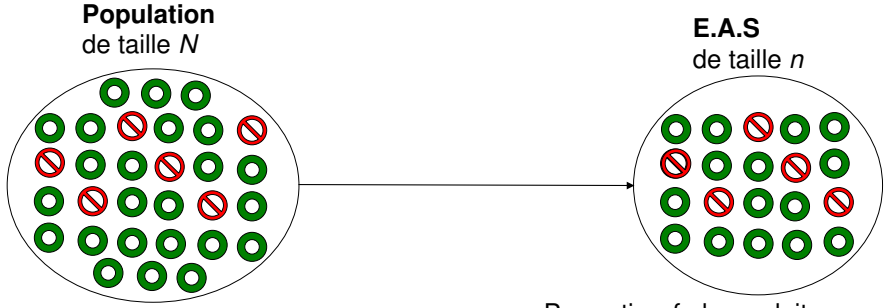
- Y : VA binomiale $\boldsymbol{B}(n,p)$
- $(X_1+...+X_n)$: Somme de n VA de Bernoulli $\boldsymbol{B}(1,p)$
- $(\sum_i X_i)/n$: Moyenne de n VA de Bernoulli $\boldsymbol{B}(1,p)$

On a $E(f_n) = p$, $Var(f_n) = p(1-p)/n$ et $\sigma(f_n) = \sqrt{\frac{p(1-p)}{n}}$

Le paramètre p représente la moyenne de la population (proportion dans la population), autour de laquelle varie f_n.

Synthèse graphique

Exemple: on s'intéresse à la proportion p (inconnue) de produits défectueux dans une population de produits de taille N.



Proportion p inconnue de produits défectueux dans la population

Proportion f_n de produits défectueux dans l'échantillon, $E(f_n) = p$, $Var(f_n) = p(1-p)/n$

Règle de l'approximation normale

La règle de l'approximation normale s'applique aux proportions comme à n'importe quelle moyenne:

Soit une proportion p inconnue dans une population de taille N.

Dans des EAS de taille n, la proportion de l'échantillon, f_n, varie autour de la proportion de la population p avec un écart-type égal à $\sqrt{p(1-p)/n}$.

Quand n augmente, la distribution d'échantillonnage (c-à-dire la distribution de f_n) est de plus en plus concentrée autour de p, et devient de plus en plus proche de la distribution normale.

Cela vient du fait qu'on peut approcher la loi $\boldsymbol{B}(n, p)$ de Y par la loi $\boldsymbol{N}(np, \sqrt{np(1-p)})$, et donc la loi de $f_n = Y/n$ par $\boldsymbol{N}(p, \sqrt{p(1-p)/n})$

5. Taille de la population à échantillonner

Dans une grande population, il importe peu, *a priori*, qu'un EAS soit tiré au sort avec ou sans remise car le tirage sans remise ne risque pas « d'épuiser » la population.

Un individu prélevé sur une population de grande taille N ne va pas affecter les fréquences de l'EAS au tirage suivant, réalisé sur $N-1$ individus. En effet, si N est de l'ordre plusieurs millions, $N-1$ le sera également.

On préfèrera dans ce cas, un tirage sans remise car il est plus « efficace »: on ne risque pas de tirer au sort deux fois le même individu et de réutiliser une information déjà connue.

Échantillonnage d'une petite population

Dans une « petite » population, en revanche, tirer au sort un EAS sans remise peut poser problème.

En effet, chaque individu tiré au sort réduit d'autant la taille de la population d'origine, ce qui affecte les probabilités relatives aux tirages suivants. Chaque tirage **dépend** du tirage précédent: les VA $X_1, \ldots, X_n$ ne sont plus i.i.d., car elles ne sont **plus indépendantes**. L'échantillon obtenu n'est plus un EAS.

Dans une « petite » population, il faut donc effectuer un tirage **avec remise** pour obtenir un EAS. Si chaque individu tiré au sort est remis dans la population, celle-ci revient à sa taille d'origine et les tirages successifs restent **indépendants**.

Petite population et facteur d'exhaustivité

Néanmoins, même avec une petite population, on préfèrerait tirer les échantillons sans remise, pour éviter les informations redondantes (Cf. « 5. Taille de la population à échantillonner »).

En effet, un échantillonnage **sans remise** donne une moyenne d'échantillon « X barre » **moins variable** (meilleure, plus proche de sa « cible » m) car, une fois prélevées, les valeurs extrêmes ne peuvent plus ressortir. Il n'y a plus à s'en préoccuper.

En général, si N et n sont les tailles respectives de la population et de l'échantillon, cette **réduction de la dispersion** dépend de l'importance relative de N et de n, mesurée par un **facteur d'exhaustivité**.

Facteur d'exhaustivité

Dans le cadre d'un tirage sans remise d'un échantillon de taille n dans une population de taille N, la dispersion de « X barre » (ou f_n) est réduite par le facteur d'exhaustivité:

$$0 \leq \sqrt{\frac{N-n}{N-1}} \leq 1$$

Autrement dit, l'écart–type d'échantillon devient

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \text{au lieu de} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

On appliquera ce facteur pour un tirage sans remise dans une population « petite » (mais pas trop !). Dans une « grande » population où $N >> n$, le facteur est ≈ 1: on peut le négliger.

6. L'échantillonnage en pratique

Ce point a été traité dans le cours à l'aide d'un exemple (fourni dans les annexes au cours en ligne sur l'EPI).

L'exemple montre comment on peut simuler le tirage d'un EAS dans une petite population à l'aide des tables de nombre au hasard (méthode de Monte Carlo).

La même méthode peut être appliquée un grand nombre de fois si l'on dispose d'un ordinateur équipé d'un logiciel adapté.

Dans ce cas, on peut faire apparaître que la moyenne (ou proportion) d'échantillon est une VA qui varie autour de la moyenne (ou proportion de la population).