جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

الكلية المتعددة التخصصات
FACULTÉ POLYDISCIPLINAIRE

Licence fondamentale en Sciences Économiques et Gestion
Module échantillonnage et estimation/Semestre 3

Chapitre II
Échantillonnage: de la population à l'échantillon

Professeur **Mohamed AIT HOU**
Département des Sciences économiques

Année universitaire 2023/2024

Plan

I • Échantillon et statistiques d'échantillon

II • Lois de probabilités des statistiques d'échantillon

III • Construction d'intervalle de Pari ou de fluctuation

Objectif

Connaître les lois de probabilités de variables aléatoires définies sur l'échantillon;

Être capable de construire des intervalles qui encadrent les statistiques d'échantillon;

Maîtriser toutes les étapes de la construction d'un intervalle de pari;

Structure du chapitre et problème concret d'échantillonnage

Problème concret

Kilométrage moyen de tous les véhicules vaut m = 18740 avec un écart type e = 2100. on tire un échantillon de taille 1000. Quel sera:

- Kilométrage moyen dans l'échantillon?
- Écart type du kilométrage dans l'échantillon?

Solution concrète

- Il y a 95% de chances pour que le kilométrage moyen dans l'échantillon soit compris entre 18610 et 18870 km.
- il y a 95% de chance pour que l'écart type du kilométrage dans l'échantillon soit compris entre 2021 et 2174 km.

Section 1: Outil pour la modélisation

Monde réel

A- Modélisation

C- Traduction

Section 4: Modélisation Simulation Traduction Récapitulatif

Monde artificiel

B- Simulation

Modèle – énoncé

Soit X_S la variable statistique parente...
Soit X la variable aléatoire parente qui …
Construire un intervalle de pari de niveau 95%:

- Sur la moyenne empirique $\overline{X}_{1000}$
- Sur l'écart type empirique S_{1000}

Section 2: Simulation

Section 3: Construction d'intervalle de pari

Modèle – solution

$$p[18610 \leq \bar{X}_{1000} \leq 18870] = 95\%$$

$$p[2021 \leq S_{1000} \leq 2174] = 95\%$$

1-1. Population mère et variable statistique parente

Définition:

On appelle « population mère » la population totale sur laquelle porte l'étude, on la note $\mathcal{P}$.

La taille de cette population mère est notée $\mathcal{N}$.

On appelle « variable statistique parente » la variable statistique qui associe à chaque individu de la population mère une modalité. On la note X_S.

La moyenne et l'écart type de X_S sont respectivement notées m et e.

1-1. Population mère et variable statistique parente

Exemple:

Dans l'exemple introductif, la population mère P est constituée de tous les véhicules répertoriés dans le pays A en 2020.

La taille N est égale à 2657000.

La *variable statistique parente* est l'application X_S, qui associe à chaque véhicule du parc automobile le nombre de kilomètres parcourus en 2020.

La moyenne de X_S sur *P* est m = 18740 km.

L'écart type de X_S sur P est e = 2100 km.

1-2. Variable aléatoire parente

Définition:

Soit X_S une variable statistique parente définie sur une population P .

Soit ε l'expérience aléatoire qui consiste à choisir au hasard un individu dans la population mère *P.*

Soit Ω l'univers des possibles associé à ε.

1. La variable aléatoire parente X est la variable aléatoire qui associe à chaque événement élémentaire de Ω une réalisation suivant le même processus d'association que X_S.
2. L'espérance et l'écart type de X sont respectivement μ et σ.

1-2. Variable aléatoire parente

Exemple : ***match de Basket-ball et exclusion de 2 min***

20 équipes participent au championnat de Basket-ball. Lors d'une saison, ces équipes ont joué 50 matchs (38 matchs de championnat et 12 matchs de la coupe). Les arbitres de ces 50 matchs ont noté sur chaque grille de match le nombre d'exclusions pour 2 min lors de la partie.

La variable statistique parente X_s associe à chacun des matchs le nombre d'exclusions pour 2 min (voir tableau).

Nombre d'exclusion pour 2 min (modalités x_j)	0	1	2	3	4	5
Effectifs n_j	10	4	8	15	10	3
Fréquences observées f_j	0,20	0,08	0,16	0,30	0,20	0,06

1-2. Variable aléatoire parente

Le nombre moyen d'exclusion pour 2 min est:

$$m = \sum_{j=0}^{5} f_j x_j = (0,20 \times 0) + (0,08 \times 1) + (0,16 \times 2) + (0,30 \times 3) + (0,20 \times 4) + (0,06 \times 5) = 2,4$$

La variance du nombre d'exclusion est:

$$e^2 = \sum_{j=0}^{5} f_i x_j^2 - m^2 = (0,20 \times 0^2) + (0,08 \times 1^2) + (0,16 \times 2^2) + (0,30 \times 3^2) + (0,20 \times 4^2) + (0,06 \times 5^2) - 2,4^2 = 2,36$$

L'écart type du nombre d'exclusions est donc égal à $e = 1,54$

1-2. Variable aléatoire parente

Propositions

Soit :

1. X_S une variable statistique parente définie sur une population mère P,
2. ε l'expérience aléatoire qui consiste à choisir au hasard un individu dans la population mère P,
3. Ω l'univers des possibles associé à ε,
4. X la variable aléatoire parente qui associe à chaque événement élémentaire de Ω une réalisation suivant le même processus d'association que X_S.

Alors :

1. La moyenne m de X_S est égale à l'espérance μ de X.
2. L'écart type e de X_S est égal à l'écart type σ de X.

1-2. Echantillon aléatoire simple

Définition

Un échantillon aléatoire simple de taille n issu de la variable aléatoire parente X est un *n-uplet* (X_1, X_2,…, X_n), appelé également *vecteur aléatoire*, où :

1. Les *n* variables aléatoires X_i suivent la même loi de probabilité que la variable aléatoire parente X.
2. Les *n* variables aléatoires X_i sont indépendantes.

Un tel échantillon est noté *n* – EAS.

1-2. Echantillon aléatoire simple

Remarques

Les n expériences aléatoires sont réalisés :

1. Soit avec remises et les n variables aléatoires X_i sont alors indépendantes.
2. Soit sans remises. Dans ce cas :

Si le rapport $\frac{n}{N}$, appelé « *taux de sondage* », est suffisamment faible, alors les n variables aléatoires X_i sont considérées indépendantes.

Si le rapport $\frac{n}{N}$ est trop élevé, alors les variables aléatoires qui constituent l'échantillon ne sont pas indépendantes

1-2. Echantillon aléatoire simple

Exemple: Tirage sans remise et lois de probabilités

On dispos d'une urne contenant 3 boules noires et 2 boules rouges.

Soit ε l'expérience aléatoire qui consiste à tirer au hasard une boule dans l'urne.

Soit X la variable aléatoire parente indicatrice de l'événement « la boule est noire ». Par définition, X associe :

- La réalisation 1 à l'événement « la boule est noire »,
- La réalisation 0 à l'événement « la boule est non noire ».

D'où : $X \rightarrow B(3/5)$

ε est répétée deux fois sans remise.

Soit X_1 la variable indicatrice de l'événement « la première boule est noire ».

Soit X_2 la variable indicatrice de l'événement « la deuxième boule est noire ».

On sait que $X_1 \rightarrow B(3/5)$, mais qu'en est-il de la loi de X_2 ?

1-2. Echantillon aléatoire simple

$$p\left[X_2=1\right]=p\left[(X_2=1)\cap((X_1=1)\cup(X_1=0))\right]$$

D'où, d'après la distributivité de l'intersection par rapport à la réunion:

$$p\left[X_2=1\right]=p\left[((X_2=1)\cap(X_1=1))\cup((X_2=1)\cap(X_1=0))\right]$$

Or, $\left\{(X_2=1)\cap(X_1=1)\right\}=\varnothing et\left\{(X_2=1)\cap(X_1=0)\right\}=\varnothing$

$$p\left[X_2=1\right]=p\left[((X_2=1)\cap(X_1=1))\right]+p\left[((X_2=1)\cap(X_1=0))\right]$$

$$p\left[X_2=1\right]=p\left[(X_2=1)/(X_1=1)\right]\times p\left[X_1=1\right]+p\left[(X_2=1)/(X_1=0)\right]\times p\left[X_1=0\right]$$

$$p\left[X_2=1\right]=\frac{2}{4}\times\frac{3}{5}+\frac{3}{4}\times\frac{2}{5}=\frac{1}{2}\times\frac{3}{5}+\frac{1}{2}\times\frac{3}{5}=\frac{3}{5}$$

Finalement, la loi de X_2 est: $L(X_2)=\begin{pmatrix}0 & 1\\ \frac{2}{5} & \frac{3}{5}\end{pmatrix}$, autrement dit, $X_2\rightarrow B\left(\frac{3}{5}\right)$

N.B: Bien que le tirage soit sans remise, X_2 suit la même loi que X_1 et que la variable aléatoire parente X.

1-3. Statistiques d'échantillon ou statistiques d'échantillonnage

Reprenons l'exemple des kilomètres parcourus par les véhicules. Pour faciliter les calcules, on va travailler sur un échantillon de taille 4 (un *4* – EAS).

Soit X la variable aléatoire parente qui associe au véhicule tiré le nombre de kilomètres parcourus en 2016.

Les trois réalisations suivantes du *4* – EAS ont été tirées au hasard:

$Réalisation\ 1:\ \ 4 - eas_1 = (18110; 16057; 19253; 21421)$

$Réalisation\ 2:\ \ 4 - eas_2 = (15721; 19399; 19171; 19781)$

$Réalisation\ 3:\ \ 4 - eas_3 = (17912; 19244; 17308; 16566)$

Ce qui permet de faire 6 calculs:

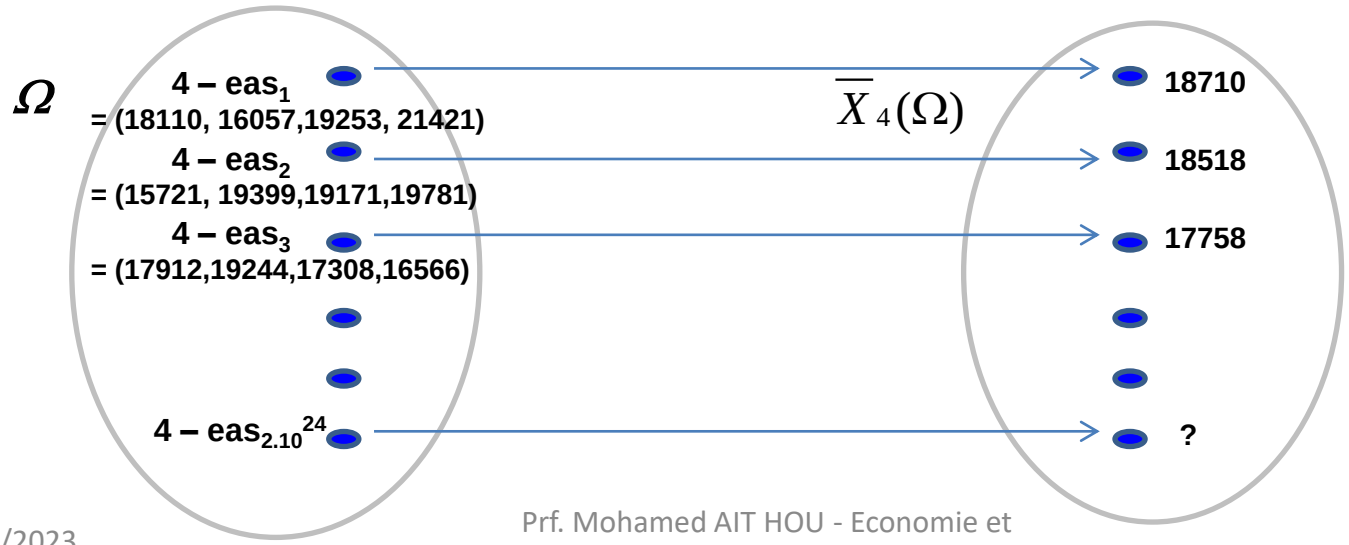
- 4 – eas_1 a pour moyenne 18710 et pour écart type 1939.
- 4 – eas_2 a pour moyenne 18518 et pour écart type 1629.
- 4 – eas_3 a pour moyenne 17758 et pour écart type 982.

1-3. Statistiques d'échantillon ou statistiques d'échantillonnage

18710, 18518, 17758 sont trois **réalisations** de la **variable aléatoire** $\overline{X}_4 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$ qui s'appelle la *moyenne empirique.* Il y a environ $\binom{2657000}{4} = C^4_{2657000} \cong 2.10^{24}$ réalisations différentes du 4 – EAS et autant réalisations différentes de la moyenne empirique.

De même, 1939, 1629, 982 sont trois **réalisations** de la **variable aléatoire** $S_4 = \sqrt{\frac{1}{4}\sum_{i=1}^{4}\left(X_i - \overline{X}_n\right)^2}$ appelée *écart type empirique*.

La variable aléatoire moyenne empirique



1-3. Statistiques d'échantillon ou statistiques d'échantillonnage

Définition

Soit (X_1, X_2,…, X_n) un échantillon aléatoire simple (EAS) de taille n issu de la variable aléatoire parente X.

1. La *moyenne empirique* est la variable aléatoire $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + ... + X_n)$
2. La *variance empirique* est la variable aléatoire $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$
3. L*'écart type empirique* est la variable aléatoire $S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$

En développant la formule de définition de la variance empirique, on obtient:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2\overline{X}_n}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X}_n^2$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2\overline{X}_n}{n}n.\overline{X}_n + \frac{1}{n}n\overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}_n^2 + \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^2$$

1-3. Statistiques d'échantillon ou statistiques d'échantillonnage

Proposition: Formule développée de la variance empirique

Soit $(X_1, X_2, \ldots, X_n)$ un *n* – EAS issu d'une variable aléatoire parente X.

$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^{\,2}$ = moyenne des carrés des X_i - carré de la moyenne empirique.

Définition

Soit $(X_1, X_2, \ldots, X_n)$ un *n* – EAS issu de la variable aléatoire parente X.

Toute variable aléatoire T_n fonction de $(X_1, X_2, \ldots, X_n)$ est une ***statistique d'échantillon ou une statistique d'échantillonnage.***

1-4. Intervalles de pari ou de fluctuation sur des statistiques d'échantillon

Revenons sur le problème relatif au kilométrage des véhicules.

Un échantillon de 1000 véhicules est choisi au hasard.

- Quel sera le nombre moyen de kilomètres parcourus observé dans l'échantillon?

- Quel sera l'écart type du nombre de kilomètres parcourus observé dans l'échantillon?

La moyenne empirique $\overline{X}_{1000} = \frac{1}{1000}\left(X_1 + X_2 + \ldots + X_{1000}\right)$ = kilométrage moyen dans l'échantillon de taille 1000.

L'écart type empirique $S_{1000} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(X_i - \overline{X}_{1000}\right)^2}$ = écart type du kilométrage dans un échantillon de taille 1000.

NB: les statistiques d'échantillon, $\overline{X}_{1000}$ *et* S_{1000} , peuvent prendre un très grand nombre de réalisations, car l'échantillon (X_1, X_2,…, X_{1000}) qui issu d'une population de taille 2657000 peut prendre $C_{2657000}^{1000} = \frac{2657000!}{1000!(2657000-1000)!}$ réalisations différentes.

1-4. Intervalles de pari ou de fluctuation sur des statistiques d'échantillon

Les probabilités individuelles des réalisations de $\overline{X}_{1000}$ *et* S_{1000} sont donc très faibles. Ces probabilités sont même nulles, si l'on considère que la variable aléatoire parente X et donc les X_i sont des variables aléatoires continues.

Plutôt que de chercher la probabilité que les variables aléatoires $\overline{X}_{1000}$ *et* S_{1000} prennent chacune une valeur particulière, il convient de déterminer des intervalles du type:

$$p\left[a \leq \overline{X}_{1000} \leq b\right] = 95\%$$

$$p\left[c \leq S_{1000} \leq d\right] = 95\%$$

Ces expressions s'appellent des *intervalles de pari de niveau 95%*, respectivement sur

$$\overline{X}_{1000} \quad et \quad S_{1000}$$

1-4. Intervalles de pari ou de fluctuation sur des statistiques d'échantillon

Définition

Soit $T_n = f(X_1, X_2, ..., X_n)$ une statistique d'échantillon.

1. Un ***intervalle de pari (ou de fluctuation) bilatéral*** sur T_n de niveau $1 - \alpha$ est un intervalle [a ; b] tel que: $p[a \leq T_n \leq b] = 1 - \alpha$
2. Un ***intervalle de pari (ou de fluctuation) unilatéral*** sur T_n de niveau $1 - \alpha$ est un intervalle $[a; +\infty[$ ou $]-\infty; a]$ tel que:

$p[a \leq T_n] = 1 - \alpha$, ou bien $p[T_n \leq a] = 1 - \alpha$

NB: les intervalles de pari sur T_n *du niveau* 1 - α sont notés $ip_{1-\alpha}(T_n)$

II. Lois de probabilités des statistiques d'échantillon

Parmi toutes les statistiques d'échantillonnage, deux sont particulièrement utiles: la moyenne empirique $\overline{X}_n$ et la variance empirique S_n^2 (ou l'écart type empirique). Ces deux statistiques d'échantillon sont des variables aléatoires. Elles ont donc une espérance et une variance ce qui amène à calculer:

L'espérance de la moyenne empirique $E(\overline{X}_n)$

La variance de la moyenne empirique $Var(\overline{X}_n)$

L'espérance de la variance empirique $E(S_n^2)$

La variance de la variance empirique $Var(S_n^2)$

Définition

Soit X une variable aléatoire et r un nombre entier.

1. Le moment d'ordre r est défini par $\mu_r = E(X^r)$
2. Le moment centré d'ordre r est défini par $\mu_{Cr} = E(X - E(X))^r$

NB: Le moment d'ordre 1 est l'espérance E(X) et le moment centré d'ordre 2 est la variance.

2-1. Moments de la moyenne empirique et de la variance

Moyenne empirique

L'espérance de la moyenne empirique est: $E\left(\bar{X}_n\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}X_i\right)$

par linéarité de l'espérance.

$E\left(\bar{X}_n\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)$, par additivité de l'espérance.

Or, pour tout *i*, $E(X_i)=E(X)=\mu$, car les variables X_i suivent toutes la même lois que X puisque (X_1, X_2,…, X_n) un *n* – EAS .

D'où: $$E\left(\bar{X}_n\right)=\frac{1}{n}\sum_{i=1}^{n}\mu=\frac{1}{n}.n\mu=\mu$$

2-1. Moments de la moyenne empirique et de la variance

Moyenne empirique

La variance de la moyenne empirique est: $Var\left(\bar{X}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$

par linéarité de l'espérance.

Or, $(X_1, X_2, \ldots, X_n)$ un n – EAS, donc les variables aléatoires X_i sont indépendantes, d'où, par additivité de la variance: $Var\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$

Or, pour tout i, $Var(X_i) = Var(X) = \sigma^2$, donc:

$$Var\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

II. Lois de probabilités des statistiques d'échantillon

2-1. Moments de la moyenne empirique et de la variance

Proposition

Si $(X_1, X_2, \ldots, X_n)$ est n – EAS issu d'une variable aléatoire parente X d'espérance μ et d'écart type σ, alors $E\left(\bar{X}_n\right) = \mu$ et $Var\left(\bar{X}_n\right) = \frac{\sigma^2}{n}$

Interprétation des moments de la moyenne empirique

La variable statistique parente est l'application X_s qui associe à chaque véhicule le nombre de kilomètres parcourus en 2016. la moyenne de X_s sur P est = 18740 km, l'écart type de X_s sur P et e = 2100 km.

En plus, ***m = μ = 18740 km*** et ***e = σ = 2100 km***

D'après l'IBT appliquée à la moyenne empirique, en choisissant k = 3, on a:

2-1. Moments de la moyenne empirique et de la variance

$$p\left[E\left(\bar{X}_n\right)-3\sigma_{\bar{X}_n} \le \bar{X}_n \le E\left(\bar{X}_n\right)+3\sigma_{\bar{X}_n}\right] \ge 1-\frac{1}{9}$$

Or, $E\left(\bar{X}_n\right)=\mu \quad et \quad Var\left(\bar{X}_n\right)=\frac{\sigma^2}{n}$

- $E\left(\bar{X}_{1000}\right)=18740$
- $\sigma_{\bar{X}_n}=\frac{\sigma}{\sqrt{n}}=\frac{2100}{\sqrt{1000}}=66$

D'où: $p\left[18740-3\times 66 \le \bar{X}_{1000} \le 18740+3\times 66\right] \ge 89\%$

Soit $p\left[18740-198 \le \bar{X}_{1000} \le 18740+198\right] \ge 89\%$

Autrement dit, on peut parier (avant de tirer l'échantillon), avec une probabilité d'au moins 89% de gagner, que le kilométrage moyen dans l'échantillon sera compris entre 18542 km et 18938 km.

2-1. Moments de la moyenne empirique et de la variance

Remarques

1. La moyenne empirique a une forte probabilité de prendre une valeur proche de 18740 km, c'est-à-dire proche du kilométrage moyen m dans la population mère.
2. Ce résultat est directement lié au fait que, d'une part, $E\left(\bar{X}_n\right) = \mu = m$ et que, d'autre part, l'écart type de la moyenne empirique est beaucoup plus petit (si n est grand) que l'écart type e de la variable statistique parente. Dans notre cas:

$$\sigma_{\bar{X}_n} = \frac{e}{\sqrt{n}} = \frac{2100}{\sqrt{1000}} = 66$$

2-1. Moments de la moyenne empirique et de la variance

Variance empirique

L'espérance de la variance empirique est:

$$E\left(S_n^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^{\,2}-\bar{X}_n^{\,2}\right)$$

$$E\left(S_n^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^{\,2}\right)-E\left(\bar{X}_n^{\,2}\right)$$ par application de l'additivité de l'espérance

$$E\left(S_n^2\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}X_i^{\,2}\right)-E\left(\bar{X}_n^{\,2}\right)$$ par application de la linéarité de l'espérance

$$E\left(S_n^2\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i^{\,2}\right)-E\left(\bar{X}_n^{\,2}\right)$$ par application de l'additivité de l'espérance

$$E\left(S_n^2\right)=\frac{1}{n}nE\left(X_i^{\,2}\right)-E\left(\bar{X}_n^{\,2}\right)=E\left(X_i^{\,2}\right)-E\left(\bar{X}_n^{\,2}\right)=E\left(X^2\right)-E\left(\bar{X}_n^{\,2}\right)$$

En effet, pour tout *i*, $L(X_i)=L(X)$ car (X_1, X_2,…, X_n) est *n* – EAS. Donc, $L(X_i^{\,2})=L(X^2)$ et $E(X_i^{\,2})=E(X^2)$

2-1. Moments de la moyenne empirique et de la variance

Variance empirique

Soit à calculer $E(X^2)$, $puis$ $E(\overline{X}_n^2)$

D'après la formule développée de la variance appliquée à *X*:

$$Var(X) = E(X^2) - [E(X)]^2 \Rightarrow E(X^2) = Var(X) + [E(X)]^2$$

$$Or, \quad Var(X) = \sigma^2 \quad et \quad [E(X)]^2 = \mu^2$$

$$Donc \quad E(X^2) = \sigma^2 + \mu^2$$

D'après la formule développée de la variance appliquée à $\bar{X}_n$:

$$Var(\bar{X}_n) = E\left(\bar{X}_n{}^2\right) - [E(\bar{X}_n)]^2 \Rightarrow E\left(\bar{X}_n{}^2\right) = Var(\bar{X}_n) + [E(\bar{X}_n)]^2$$

Or, $Var(\bar{X}_n) = \frac{\sigma^2}{n}$ et $[E(\bar{X}_n)]^2 = \mu^2$

Donc: $\boldsymbol{E\left(\bar{X}_n{}^2\right) = \frac{\sigma^2}{n} + \mu^2}$

2-1. Moments de la moyenne empirique et de la variance

D'où :

$$E(S_n^2) = E(X^2) - E\left(\overline{X}_n^2\right) = (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \sigma^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n}$$

D'où: $E(S_n^2) = \frac{n-1}{n}\sigma^2$

Proposition

Si $(X_1, X_2, \ldots, X_n)$ est n – EAS issu d'une variable aléatoire parente X d'espérance μ et d'écart type σ, alors :

1. $E\left(S_n^2\right) = \frac{n-1}{n}\sigma^2$
2. $Var\left(S_n^2\right) \cong \frac{1}{n}\left(\mu_{C4} - \sigma^4\right)$, lorsque n est grand et où μ_{C4} est le moment centré d'ordre 4.

2-2. Loi de probabilité de la moyenne empirique et de la variance empirique

Définition

Un échantillon est gaussien s'il est issu d'une variable aléatoire parente qui suit une loi de Gauss.

Loi de probabilité (distribution d'échantillonnage) de la moyenne empirique $\bar{X}_n$

Par définition d'un échantillon aléatoire simple ($X_1, X_2, \ldots, X_n$), les X_i suivent la même loi que la variable aléatoire X et sont indépendantes.

•**Cas 1.** lorsque l'échantillon est gaussien, alors les X_i suivent une loi de gauss et sont indépendantes. D'après l'additivité de la loi de Gauss, la somme $Y = X_1 + \ldots + X_n$ suit une loi de Gauss. La linéarité de la loi de Gauss permet de conclure que $\bar{X}_n = \frac{1}{n} Y = \frac{1}{n} \sum_{i=1}^{n} X_i$ suit également une loi de Gauss.

2-2. Loi de probabilité de la moyenne empirique et de la variance empirique

•**Cas 2.** Lorsque l'échantillon n'est pas gaussien mais qu'il est de grande taille (n au moins égal à 30), alors les conditions du théorème central limite (TCL) sont satisfaites. Le TCL permet d'affirmer que $\bar{X}_n$ tend vers une loi de Gauss.

•**Cas 3.** Lorsque l'échantillon n'est ni Gaussien ni de grande taille, il convient de chercher au cas par cas la loi de $\bar{X}_n$. Par exemple, les X_i suivent une loi de Bernoulli, leur somme suit une loi binomiale et il est alors possible de déterminer la loi de $\bar{X}_n$.

Pour les cas 1 et 2, la moyenne empirique suit ou tend vers une loi normale, sachant que:

$$E\left(\bar{X}_n\right)=\mu \quad et \quad Var\left(\bar{X}_n\right)=\frac{\sigma^2}{n}$$

En centrant et en réduisant, on en déduit que $\dfrac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ suit ou tend vers la loi normal centrée réduite.

2-2. Loi de probabilité de la moyenne empirique et de la variance empirique

Loi de probabilité (distribution d'échantillonnage) de la variance empirique S_n^2

Les hypothèses de travail sont les mêmes que pour la moyenne empirique:

Les X_i suivent la même loi que la variable aléatoire parente X et sont indépendantes.

•**Cas1.** Lorsque l'échantillon est gaussien, alors les X_i suivent une loi de Gauss et sont indépendantes.

•**Cas 2.** Lorsque l'échantillon n'est pas gaussien mais qu'il est de grande taille (*n* au moins égal à 30), l'application du TCL permet de montrer que la variance empirique tend vers une loi normale.

•**Cas 3.** Lorsque l'échantillon n'est ni Gaussien ni de grande taille, il convient de chercher au cas par cas permet éventuellement de déterminer la loi de la variance empirique.

3-1. Intervalle de Pari sur la moyenne empirique

Intervalle de niveau 95% sur $\bar{X}_{1000}$

La question était de déterminer un intervalle de pari de niveau 95% sur $\bar{X}_{1000}$, c'est-à-dire déterminer les nombres ***a*** et ***b*** tel que: $p\left[a \le \bar{X}_{1000} \le b\right] = 95\%$

On sait que

- L'échantillon n'est pas gaussien, puisque rien ne permet d'affirmer que la variable parente suit une loi de Gauss.
- l'échantillon est de grande taille: n = 1000 >> 30.

D'où on peut écrire $\bar{X}_{1000} \to N\left(\mu;\sigma/\sqrt{n}\right)$.

$E\left(\bar{X}_{1000}\right)=\mu=18740$ et $\sigma_{\bar{X}_{1000}}=2100/\sqrt{1000}=66$, donc $\bar{X}_{1000} \to N(18740;66)$

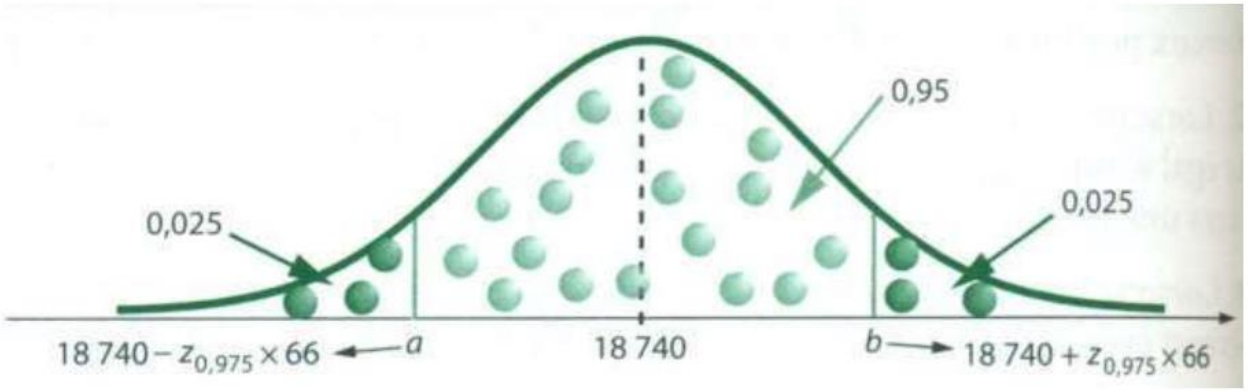
III. Construction d'intervalle de Pari ou fluctuation

3-1. Intervalle de Pari sur la moyenne empirique

L'intervalle en question n'est pas unique, en effet, pour avoir une aire de 95% sous la courbe de la fonction de densité de la loi normal, plusieurs valeurs de a et b sont possible.

Il est possible de chercher l'intervalle le plus petit. Un tel intervalle est forcément centré sur l'espérance de $\bar{X}_{1000}$, puisque c'est en cette valeur que la densité de probabilité est maximale.

Densité de probabilité de la moyenne empirique



3-1. Intervalle de Pari sur la moyenne empirique

Il s'agit donc de trouver l'intervalle de la forme (notation $z_{97,5\%}$ est justifiée ci-après):

$$p\left[18740 - z_{97,5\%} \times 66 \le \bar{X}_{1000} \le 18740 + z_{97,5\%} \times 66\right] = 95\%$$

Soit en centrant et en réduisant: $$p\left[-z_{97,5\%} \le \frac{\bar{X}_{1000} - 18740}{66} \le z_{97,5\%}\right] = 95\%$$

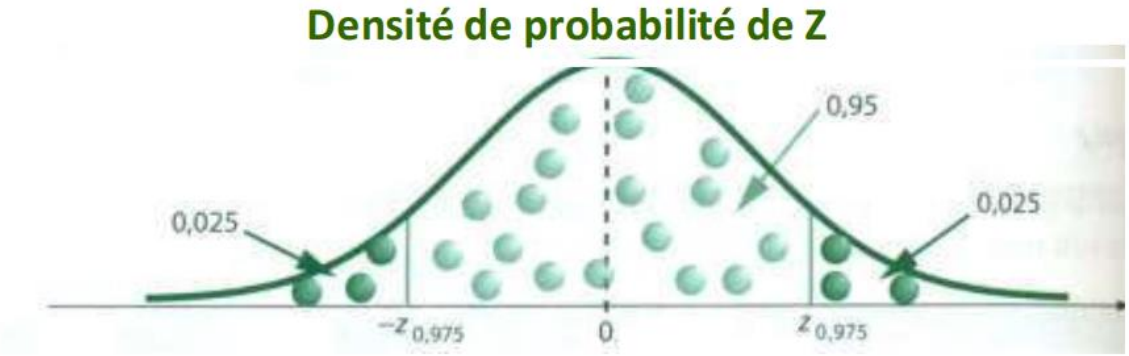
Avec $Z = \frac{\bar{X}_{1000} - 18740}{66}$ qui suit une loi normale centré réduite, on obtient:

$p\left[-z_{97,5\%} \le Z \le z_{97,5\%}\right] = 95\%$, d'où $2F_Z\left(z_{97,5\%}\right) - 1 = 95\% \Rightarrow F_Z\left(z_{97,5\%}\right) = 0,975$

Où F_Z est la fonction de répartition de la variable aléatoire Z. le choix de la notation $z_{97,5\%}$ est ainsi justifié: $z_{97,5\%}$ est le quantile d'ordre 97,5% de la loi de Gauss centrée réduite.

III. Construction d'intervalle de Pari ou fluctuation

3-1. Intervalle de Pari sur la moyenne empirique



D'après Excel, $z_{97,5\%}$ = ***LOI.NORMALE.INVERSE.N(0,975;0;1) = 1,96*** on obtient:

$$p\left[18740-1,96\times 66 \le \bar{X}_{1000} \le 18740+1,96\times 66\right]=95\%$$

$$p\left[18740-129,36 \le \bar{X}_{1000} \le 18740+129,36\right]=95\%$$

$$p\left[18610,64 \le \bar{X}_{1000} \le 18869,36\right]=95\%$$

Traduction: **avant** de tirer l'échantillon, le statisticien sait qu'il y a une probabilité de 95% que le kilométrage moyen dans l'échantillon soit compris entre 18610 et 18870 km. Il y a donc une probabilité de 5% que le km moyen dans l'échantillon ne soit pas compris dans [18610;18870]

III. Construction d'intervalle de Pari ou fluctuation

3-1. Intervalle de Pari sur la moyenne empirique

Proposition

Soit $(X_1, X_2, \ldots, X_n)$ est n – EAS issu d'une variable aléatoire parente X d'espérance μ et d'écart type σ. Si $n > 30$, alors les *intervalles des pari (ou de fluctuation) bilatéraux* sur $\bar{X}_n$ de niveau $1 - \alpha$ sont de la forme:

$$p\left[\mu - z_{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}} \leq \bar{X}_{1000} \leq \mu + z_{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$, où $z_{\left(1-\frac{\alpha}{2}\right)}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de la loi normale centrée réduite.

3-2. Intervalle de Pari sur la variance empirique

Déterminer un intervalle de pari du niveau 95% sur S_n consiste à déterminer les nombres ***c*** et ***d*** tel que: $p\left[c \le S_n \le d\right] = 95\%$

Comme la loi de l'écart type empirique n'est pas connue, l'idée est de construire – dans un premier temps – un intervalle sur la variance empirique. Les valeurs cherchées sont donc c^2 et d^2 tel que: $p\left[c^2 \le S_n^2 \le d^2\right] = 95\%$

Comme précédemment:

- L'échantillon n'est pas gaussien, puisque rien ne permet d'affirmer que la variable parente suit une loi de Gauss.
- l'échantillon est de grande taille: n = 1000 >> 30.

La statistique d'échantillon à utiliser est S_n^2 suivant une approximativement $N \to \left(E\left(S_n^2\right); \sigma_{S_n^2}\right)$

$$E\left(S_n^2\right) = \frac{n-1}{n}\sigma^2 = \frac{999}{1000} \times 2100^2 = 4405590 km$$

$$Var\left(S_n^2\right) \cong \frac{1}{n}\left(\mu_{C4} - \sigma^4\right)$$

3-2. Intervalle de Pari sur la variance empirique

Le calcule de la variance empirique demande de connaître la valeur du moment centré d'ordre 4. on admet que l'écart type de la variance empirique vaut 162500 km. Avec cette valeur: $S_n^2 \to N(4405590;162500)$.

Un raisonnement analogue à celui qui est développé pour la moyenne empirique conduit à:

$$p\left[4405590 - z_{\left(1-\frac{\alpha}{2}\right)} \times 162500 \le S_{1000}^2 \le 4405590 + z_{\left(1-\frac{\alpha}{2}\right)} \times 162500\right] = 1-\alpha$$

Où $z_{\left(1-\frac{\alpha}{2}\right)}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de la loi normale centrée réduite.

Pour un intervalle de pari de niveau 95%, α = 0,05 et, d'après Excel, $z_{97,5\%}$ = ***LOI.NORMALE.INVERSE.N(0,975;0;1) = 1,96*** on obtient:

$$p\left[4405590 - 1,96 \times 162500 \le S_{1000}^2 \le 4405590 + 1,96 \times 162500\right] = 95\%$$

3-2. Intervalle de Pari sur la variance empirique

D'où en on déduit: $p\left[\sqrt{4405590-1,96\times162500} \le S_{1000} \le \sqrt{4405590+1,96\times162500}\right]=95\%$

D'où: $p\left[2021,6 \le S_{1000} \le 2173,5\right]=95\%$

La précision des données initiales (moyenne et écart type dans la population mère) est le kilomètre. L'intervalle arrondi par valeur inférieure pour la borne de droite et supérieure pour la borne de gauche est: $ip_{95\%}\left(S_{1000}\right)=\left[2021;2174\right].$

Il y a ainsi une probabilité de 95% pour que l'écart type du kilométrage dans l'échantillon soit compris entre **2021 et 2174 km**.