

# COURS N°5 : n-Échantillons.

Un **n-échantillon** ( $X_1, X_2, \dots, X_n$ ) est défini par un **ensemble de n variables aléatoires indépendantes et de même loi** (c'est des **répétitions indépendantes d'une même expérience**). On dit que les variables aléatoires sont **indépendantes et identiquement distribuées (iid)**.

Le n-échantillon va nous permettre d'effectuer des **inférences statistiques** [on va pouvoir tirer des conclusions statistiques à partir d'hypothèses statistiques] sur une **population inconnue**. Cependant, le plus souvent décrire la loi de probabilité du n-échantillon ne présente pas d'intérêt en soi. Ce qui nous intéresse, c'est la **loi de la somme des n variables aléatoires** ( $\sum X_i$ ) ou la **loi de la moyenne des n variables aléatoires** ( $\bar{X} = \frac{1}{n} \sum X_i$ ).

## I. Somme de n variables aléatoires indépendantes

Soit  $S_n$  l'addition de n variables aléatoires indépendantes  $X_i$ , de moyenne  $E(X)$  et de variance  $\text{Var}(X)$ . On peut utiliser les formules suivantes :

$$S_n = X_1 + X_2 + \dots + X_n$$

$$E(S_n) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = n \times E(X)$$

$$\text{Var}(S_n) = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) = n \times \text{Var}(X)$$

[on peut utiliser cette simplification pour la variance car les  $X_i$  sont indépendantes]

[Par exemple, la somme de n variables aléatoires indépendantes de Bernoulli donnent une loi Binomiale d'espérance  $np$  et de variance  $np(1-p)$ .

La somme de n variables aléatoires indépendantes de Poisson donne une loi de Poisson d'espérance  $n\lambda$  et de variance  $n\lambda$ .

La somme de n variables aléatoires indépendantes normales donne une loi normale d'espérance  $n\mu$  et de variance  $n\sigma^2$ .]

**Attention** : il ne faut pas confondre somme de n variables aléatoires indépendantes ( $S_n = X_1 + X_2 + \dots + X_n$ ) avec une affectation d'un coefficient multiplicateur pour une variable aléatoire ( $Y = n \times X_i$ ).

[Remarque : si Y est la somme de n variables aléatoires indépendantes de Gauss  $X_i$ , centrées et réduites telle que :

$$Y = \sum_{i=1}^n X_i^2$$

Alors Y suit la **loi du chi 2** à  $(n-1)$  degrés de liberté. Voir dernier cours de statistiques.]

## II. Théorème de la Limite Centrale (TLC) ou Théorème Central Limite

### A. Définition

Si on prend  $n$  variables aléatoires  $X_i$ , indépendantes et identiquement distribuées (iid), d'espérance  $E(X)$  et de variance  $\text{Var}(X)$ , alors « quelque soit la loi de  $X$ , la loi de  $S_n$  converge toujours vers une loi de Gauss quand  $n$  tend vers l'infini. »

$S_n$  suit alors une loi normale, d'espérance  $nE(X)$  et de variance  $n\text{Var}(X)$   $\{S_n \sim N[nE(X); n\text{Var}(X)]\}$   
C'est une loi approchée, qui est meilleure quand  $n$  est grand.

### B. Conditions d'application du TLC

#### 1. Si $X$ est continue

Si  $X$  est continue, l'approximation gaussienne est raisonnable si  $n \geq 30$

#### 2. Si $X$ est binaire

Si  $X$  est binaire [deux valeurs possibles pour l'expérience aléatoire], l'approximation gaussienne est raisonnable si  $n \times p$  ET  $np \times (1-p) \geq 5$

**Remarque** : il existe **deux approximations pour la loi Binomiale** :

- On peut approximer la loi Binomiale par une loi Normale, par l'intermédiaire du **TLC**, si  $n \times p$  ET  $np \times (1-p) \geq 5$  .
- On peut approximer la loi Binomiale par une loi de Poisson, si  **$p$  est petit et si  $n$  est grand** (on parle de **loi des événements rares**).

[On utilise ensuite la loi Normale comme on sait le faire pour résoudre les problèmes].

### III. Moyenne de n variables aléatoires

#### A. Définition

Soit  $n$  variables aléatoires  $X_i$ , indépendantes et identiquement distribuées, de moyenne  $\mu$  et de variance  $\sigma^2$ . ( $X_1 + X_2 + \dots + X_n$ )

$M_n$  sera la moyenne de ce n-échantillon si elle est caractérisée par la formule suivante :

$$M_n = \frac{S_n}{n} = \frac{(X_1 + X_2 + \dots + X_n)}{n}$$

$M_n$  est une VARIABLE ALEATOIRE.

$m_n$  est une VALEUR PARTICULIERE que prend  $M_n$  sur un  $n$ -échantillon particulier.

**Remarque** : une **proportion** est une **moyenne**. En effet, mesurer une proportion, c'est compter les événements sur  $n$  répétitions d'une expérience aléatoire, et les rapporter aux nombres de répétitions.

#### B. Espérance de la moyenne d'un n-échantillon

$$M_n = \frac{S_n}{n}$$

$$E(M_n) = E\left(\frac{S_n}{n}\right) = \frac{E(S_n)}{n}$$

$$E(S_n) = E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) = n \times E(X)$$

**Car tous les  $X_i$  sont équidistribués [et donc ont la même espérance  $E(X)$ ]**

$$\frac{E(S_n)}{n} = \frac{n \times E(X)}{n} = E(X)$$

d'où :

$$E(M_n) = E(X)$$

## C. Variance de la moyenne d'un n-échantillon

$$M_n = \frac{S_n}{n}$$

$$\text{Var}(M_n) = \text{Var}\left(\frac{S_n}{n}\right) = \frac{\text{Var}(S_n)}{n^2}$$

$$\text{Var}(S_n) = \text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) = n \times \sigma^2$$

**Car tous les  $X_i$  sont équidistribués et INDEPENDANTS** [et donc ont la même variance  $\sigma^2$ ]

$$\frac{\text{Var}(S_n)}{n^2} = \frac{n \times \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

D'où :

$$\boxed{\text{Var}(M_n) = \frac{\text{Var}(X)}{n}}$$

## D. Loi de la moyenne des n variables aléatoires

### 1. Loi de la moyenne de n variables aléatoires gaussiennes indépendantes

Soit  $X_i$  des variables aléatoires gaussiennes indépendantes d'espérance  $\mu$  et de variance  $\sigma^2$ .  
« Toute combinaison linéaire de variables aléatoires gaussiennes est une variable aléatoire gaussienne » donc  **$M_n$  suit une loi normale.**

On a vu que  $E(M_n) = E(X)$

Dans ce cas, on a  $E(M_n) = E(X) = \mu$

On a vu que  $\text{Var}(M_n) = \frac{\text{Var}(X)}{n}$

Dans ce cas, on a  $\text{Var}(M_n) = \frac{\text{Var}(X)}{n} = \frac{\sigma^2}{n}$

Donc  **$M_n$  suit une loi normale d'espérance  $\mu$  et de variance  $\frac{\sigma^2}{n}$**  [ $M_n \sim N(\mu; \frac{\sigma^2}{n})$ ]

C'est une **loi exacte de  $M_n$**  (il n'y a pas d'approximation).

## 2. Loi de la moyenne de n variables aléatoires non gaussiennes (indépendantes)

Soit  $X_i$  des variables aléatoires indépendantes d'espérance  $E(X)$  et de variance  $\text{Var}(X)$ .

On utilise le **TLC** : « quelque soit la loi de  $X$ , la loi de  $M_n$  converge toujours vers une loi de Gauss quand  $n$  tend vers l'infini. »

[On rappelle que pour les variables aléatoires continues, il faut que  $n \geq 30$  et pour les variables aléatoires binaires, il faut que  $n \times p \geq 5$  et  $np \times (1-p) \geq 5$  ]

Donc  $M_n$  suit une **loi normale** d'espérance  $E(X)$  et de variance  $\frac{\text{Var}(X)}{n}$

$$[M_n \sim N(E(X); \frac{\text{Var}(X)}{n})]$$

C'est une **loi approchée de  $M_n$**  (et cette approximation est meilleure pour des valeurs de  $n$  élevées, car la variance sera d'autant plus petite que le  $n$  sera grand).

## IV. Intervalle de pari (ou de fluctuation)

### A. Définition

On appelle **intervalle de pari**, de **niveau  $1 - \alpha$**  (ou au **risque  $\alpha$** ) de  $X$  (variable aléatoire), l'intervalle **centré sur l'espérance de  $X$**  pour lequel  $P(a \leq X \leq b) = 1 - \alpha$   
Ainsi, l'intervalle de pari s'écrit :

$$IP_{1-\alpha}(X) = [\mu \pm \varepsilon_\alpha \sigma] = [\mu - \varepsilon_\alpha \sigma; \mu + \varepsilon_\alpha \sigma]$$

[On peut aussi remplacer  $\sigma$  par  $\sqrt{\text{Var}(X)}$ , c'est la même chose !]

**Remarque** : Généralement, on fait des intervalles de Pari de **niveau 95%**, ou de **risque 5%**, et l' $\varepsilon_\alpha$  correspondant est de 1,960 [voir formulaire]. On arrondit cette valeur à 2 et on a :

$$IP_{95}(X) = [\mu \pm 2\sigma] = [\mu - 2\sigma; \mu + 2\sigma]$$

## **B. Largeur de l'intervalle de pari de niveau 95%**

La **largeur** de l'intervalle de pari, notée généralement **l**, est définie par la **différence entre la borne supérieure de l'intervalle et la borne inférieure de l'intervalle** :

$$l = \text{borne supérieure} - \text{borne inférieure}$$

A partir de cette définition, on a :

$$l = 2 \varepsilon_{\alpha} \sigma$$

## **C. Précision de l'intervalle de pari de niveau 95%**

La **précision** d'un intervalle de pari correspond à **l'écart à la moyenne**. Elle est définie par la relation :

$$\text{précision} = \varepsilon_{\alpha} \sigma$$

**Remarque** : Pour les intervalles de pari de **moyenne de variables aléatoires**, qui a pour **écart-type**  $\sqrt{\frac{\text{Var}(X)}{n}}$  on a :

$$IP_{95}(M_n) = [\mu \pm 2 \sqrt{\frac{\text{Var}(X)}{n}}] = [\mu - 2 \sqrt{\frac{\text{Var}(X)}{n}}; \mu + 2 \sqrt{\frac{\text{Var}(X)}{n}}]$$

Ainsi, on va pouvoir **déterminer la taille du n-échantillon nécessaire pour constituer une certaine largeur**, ou une certaine **précision**, d'intervalle de pari **souhaitée**.

On partant de  $l = 2 \times 2 \sqrt{\frac{\text{Var}(X)}{n}}$  (pour un **intervalle de pari de niveau 95%**), on arrive mathématiquement à :

$$n = \left( \frac{2 \times 2}{l} \right)^2 \text{Var}(X)$$

Pour la **précision**, on a :

$$n = \left( \frac{2}{\text{précision}} \right)^2 \text{Var}(X)$$