



Université  
de Limoges

IREM Institut de Recherche  
sur l'Enseignement des Mathématiques

IREM Limoges  
Année 2012

## Statistiques Inférentielles

---

Pierre DUSART



$$\begin{aligned}\sigma^2 &= \sum_{k=0}^n k^2 \binom{n}{k} p^k q^{n-k} - (np)^2 \\ &= \sum_{k=1}^n k \binom{n-1}{k-1} p^{k-1} q^{n-k} np - (np)^2 \\ &= \left( \sum_{k=1}^n k \binom{n-1}{k-1} p^{k-1} q^{n-k} - np \right) \cdot np \\ &= \left( \sum_{k=1}^n (k-1) \binom{n-1}{k-1} p^{k-1} q^{n-k} + \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-1-k} - np \right) \cdot np \\ &= ((n-1)p + 1 - np) \cdot np \\ \sigma^2 &= npq\end{aligned}$$

### Prérequis :

- Fluctuations d'échantillonnage (Stage 1)
- Convergence de la loi binomiale vers loi normale (Stage 2)
- Propriétés de la loi normale (Stage 2)

## 1 Intervalle de Fluctuation

### 1.1 Définitions

On considère une variable aléatoire  $X$ . Même si  $X$  suit une loi connue, sa réalisation peut prendre toutes les valeurs possibles de cette loi. Si on s'intéresse à des lois autres que la loi uniforme (équiprobabilité) et que l'on doit miser sur une valeur, autant prendre la valeur la plus probable. Si l'on a le droit de miser sur d'autres valeurs, cela conduit à miser sur un intervalle de la forme  $[a, b]$ . On maîtrise alors le risque (noté  $\alpha$ ) de perdre, c'est-à-dire la probabilité que  $X$  n'appartienne pas à l'intervalle choisi :

$$\alpha = P(X \notin [a, b]).$$

Comme on n'aime pas perdre, on préfère le voir sous la forme de la probabilité de gagner (événement contraire) soit

$$P(X \in [a, b]) = 1 - \alpha.$$

Cet intervalle est appelé intervalle de fluctuation au seuil  $1 - \alpha$  (ou coefficient de sécurité).

Remarque 1 : comme indiqué dans le document ressource, un intervalle de fluctuation au seuil 1 n'a pas d'intérêt et n'a pas de réalité.

Remarque 2 : Cette notion d'intervalle de fluctuation est à ne pas confondre avec intervalle de confiance que l'on verra par la suite. En effet, dans le premier cas, la loi de  $X$  – et ses paramètres – sont connus.

On peut chercher :

- un intervalle qui a l'amplitude minimale ( $IF_1$ )
- le plus petit intervalle centré autour de l'espérance ( $IF_2$ )
- un intervalle qui symétrise les probabilités que  $X$  soit à l'extérieur ( $IF_3$ )
- un intervalle approché (calcul approché de seconde, ou remplacement par une loi continue) ( $IF_4$ )

### 1.2 Application au cas binomial

Soit  $X$  une variable suivant une loi  $\mathcal{B}(n, p)$  et  $\alpha$  un réel dans l'intervalle  $]0, 1[$ . On prend, à titre d'exemple,  $n = 100$ ,  $p = 0,3$  et  $\alpha = 0,05$ . On cherche un intervalle de fluctuation au seuil 0,95 selon les différents critères :

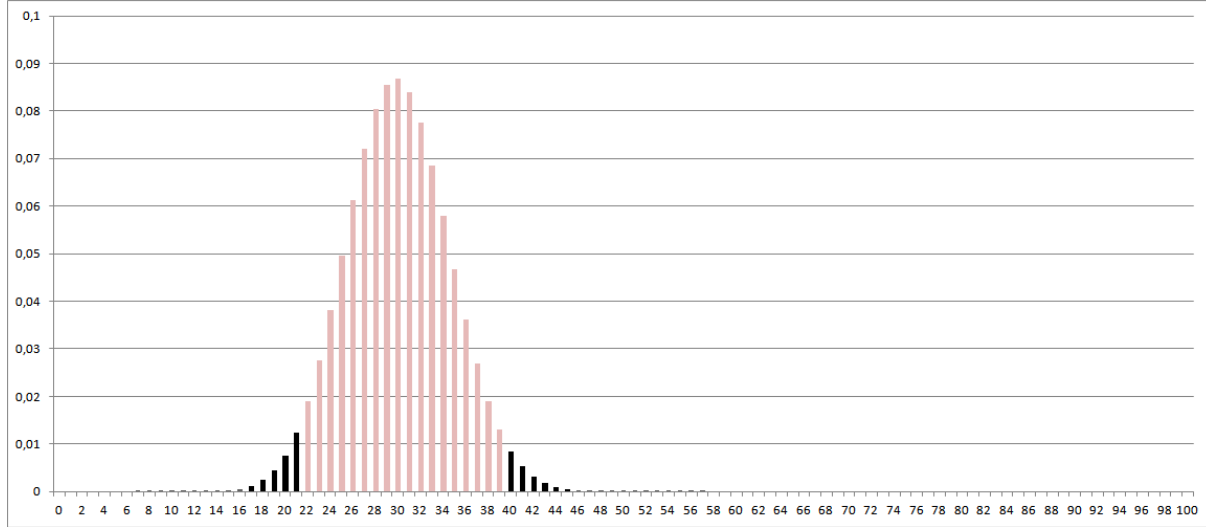
- ( $IF_1$ ), celui qui a l'amplitude minimale :  $[22, 39]$  de probabilité 0,9502
- ( $IF_2$ ), le plus petit intervalle centré autour de l'espérance (l'espérance vaut  $np$  pour une loi binomiale de paramètres  $n$  et  $p$ ) :  $[21, 39]$  de probabilité 0,9625
- ( $IF_3$ , niveau première), celui qui symétrise les probabilités que  $X$  soit à l'extérieur :  $[20, 39]$  avec une probabilité inférieure à 0,025 que  $X$  soit plus petit et inférieure à 0,025 que  $X$  soit plus grand que les valeurs de l'intervalle, de probabilité 0,9701.
- ( $IF_4$ , niveau seconde), intervalle de fluctuation approché au seuil 0,95, valable sous certaines conditions (approximation de  $\mathcal{B}$  par  $\mathcal{N}$  et  $p$  proche de 0.5) :  $[20, 40]$  de probabilité 0,9786.

#### Méthodes :

- Pour  $IF_1$  et  $IF_2$ , on prend un intervalle de départ restreint à une seule valeur (La valeur la plus probable pour  $IF_1$  et la valeur moyenne pour  $IF_2$ ). On étend ensuite cet intervalle de proche en proche (en incluant la probabilité d'une valeur possible contiguë à l'intervalle) jusqu'à obtenir la probabilité (seuil) choisie.
- Pour  $IF_3$ , on part de la plus petite valeur possible ( $X = 0$  pour le cas binomial), puis on incrémente cette valeur pour trouver l'entier  $a$  tel que  $P(X \leq a) = 0,025$  (ou le plus proche possible). On cherche ensuite la valeur  $b$ . On part de la plus grande valeur possible (ici  $X = n$ ), puis on décrémente cette valeur pour trouver l'entier  $b$  tel que  $P(X \geq b) = 0,025$ . On peut également partir de  $X = 0$  et incrémenter cette valeur jusqu'à trouver  $b$  tel que  $P(X \leq b) = 0,975$ .
- Pour le dernier intervalle, pas de recherche à faire puisque cet intervalle est fixé (lorsque  $n, p, \alpha$  sont donnés). Il peut donner une idée des valeurs de  $a$  et  $b$  pour  $IF_3$ .



On retient donc comme intervalle de fluctuation au seuil de 95%, toutes les valeurs correspondantes aux probabilités suivantes :



On rappelle la forme de l'intervalle approché  $IF_4$ , (dont la justification se trouve au paragraphe 1.3) :

- en classe de **seconde**,

$$[np - \sqrt{n}; np + \sqrt{n}]$$

- et en classe de **terminale**,

$$[np - u_\alpha \sqrt{np(1-p)}; np + u_\alpha \sqrt{np(1-p)}]$$

où  $u_\alpha$  désigne l'unique réel tel que  $P(-u_\alpha < Z < u_\alpha) = 1 - \alpha$  avec  $Z$  suivant une loi normale  $\mathcal{N}(0, 1)$ . Soit ici,  $n = 100$ ,  $p = 0.3$  et  $\alpha = 0.05$  (soit  $u_\alpha = 1.96$ )

$$IF_4 = [30 - 10 ; 30 + 10] \text{ ou } [21, 018 ; 38, 982]$$

On peut vérifier que :

- pour des mêmes valeurs de  $p$  et de  $n$ , les intervalles augmentent si le seuil augmente.
- pour un seuil fixé et pour une même valeur de  $p$ , ces différents intervalles sont proportionnellement plus petits lorsque  $n$  augmente.

### 1.3 Intervalle de fluctuation asymptotique

On a vu (prérequis) par le théorème de Moivre-Laplace que, sous certaines conditions de convergence ( $n$  grand, ...), lorsque  $X_n$  suit la loi binomiale  $\mathcal{B}(n, p)$  alors

$$P(X_n \leq a) \approx P(Z_n \leq a)$$

où  $Z_n$  suit  $\mathcal{N}(np, \sqrt{np(1-p)})$ .

La recherche des bornes de l'intervalle est alors plus facile (on dispose d'une loi continue plutôt que d'une fonction en escalier). On cherche :

- $a$  tel que  $P(Z_n \leq a) = 0.025$  avec  $Z_n$  suivant  $\mathcal{N}(np, \sqrt{np(1-p)})$ .
- $b$  tel que  $P(Z_n \leq b) = 0.975$  avec  $Z_n$  suivant  $\mathcal{N}(np, \sqrt{np(1-p)})$ .

Cette recherche est un simple calcul de fonction réciproque de la fonction de répartition de la loi normale  $\mathcal{N}(np, \sqrt{np(1-p)})$ . Ce calcul peut se faire à l'aide d'un tableur :

$$= \text{LOI.NORMALE.INVERSE}(\text{PROBA}, \text{MOYENNE}, \text{ECART-TYPE})$$

où PROBA vaut respectivement 0,025 ou 0,975 pour trouver la valeur de  $a$  ou  $b$ , la MOYENNE vaut  $np$  et l'ÉCART-TYPE vaut  $\text{RACINE}(n * p * (1 - p))$ .

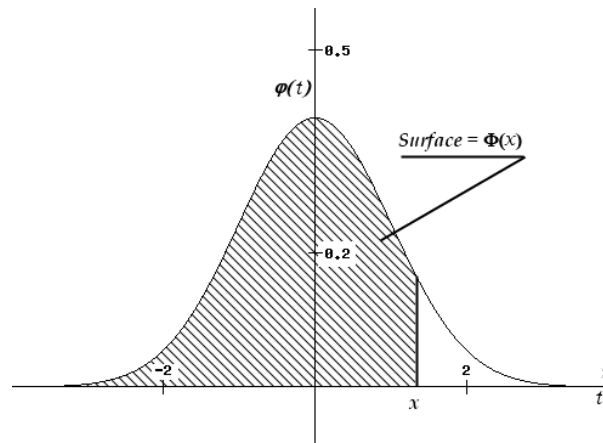
Grâce aux propriétés de la loi normale, on peut ramener ce calcul à une lecture de table de la loi normale centrée réduite  $\mathcal{N}(0, 1)$ . Si  $Z_n$  suit  $\mathcal{N}(np, \sqrt{np(1-p)})$  alors  $Z = \frac{Z_n - np}{\sqrt{np(1-p)}}$  suit  $\mathcal{N}(0, 1)$ . Ainsi

$$P(Z_n \leq a) = P(Z_n - np \leq a - np) = P\left(\frac{Z_n - np}{\sqrt{np(1-p)}} \leq \frac{a - np}{\sqrt{np(1-p)}}\right) = P(Z \leq a'), \text{ où } a' = \frac{a - np}{\sqrt{np(1-p)}}.$$

On cherche la valeur  $a'$  telle que  $P(Z < a') = 0.025$ .

Soit  $\Phi$  la fonction de répartition de  $\mathcal{N}(0, 1)$ . On cherche en fait  $\Phi^{-1}(0.025)$ . On trouve  $a' = -1.96$  (cette valeur est notée généralement  $-u_\alpha$  ou  $-z_{\alpha/2}$  avec  $\alpha = 0.05$ ). Cette valeur peut être lue dans les tables de  $\mathcal{N}(0, 1)$  ou calculée avec un tableur

$$= \text{LOI.NORMALE.STANDARD.INVERSE}(0.025) .$$



Ainsi  $\frac{a - np}{\sqrt{np(1-p)}} = -1.96$ , ce qui conduit à

$$a = np - 1.96\sqrt{np(1-p)}.$$

En suivant la même démarche pour  $b$ ,  $P(Z_n \leq b) = P\left(\frac{Z_n - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}\right) = P(Z \leq b')$  que l'on souhaite égal à 0,975. Comme  $\Phi^{-1}(0.975) = 1.96$ ,  $P(Z < 1.96) = 0.975$  et

$$b = np + 1.96\sqrt{np(1-p)}.$$

L'intervalle de fluctuation asymptotique de la variable  $X_n$  au seuil 0.95 vaut

$$I = [np - 1.96\sqrt{np(1-p)}; np + 1.96\sqrt{np(1-p)}].$$

Si  $p$  est proche de  $1/2$ , alors  $\sqrt{p(1-p)} \approx 1/2$ , son maximum. Ainsi, comme  $1.96\sqrt{p(1-p)} \leq 1$ , cet intervalle peut être "simplifié". On obtient l'intervalle de fluctuation asymptotique

$$I = [np - \sqrt{n}; np + \sqrt{n}].$$

En divisant par  $n$ , on retrouve l'intervalle au seuil (approximatif) de 95%, pour la variable fréquence, proposé dans le programme de seconde

$$I = \left[ p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right].$$

Remarque : l'intervalle de fluctuation n'est pas "unique" : par exemple, si  $X$  suit  $\mathcal{N}(0, 1)$  alors  $P(-2.576 < X < 1.696) \approx 0.95$  (Ex 2 p27 Doc Ressource).

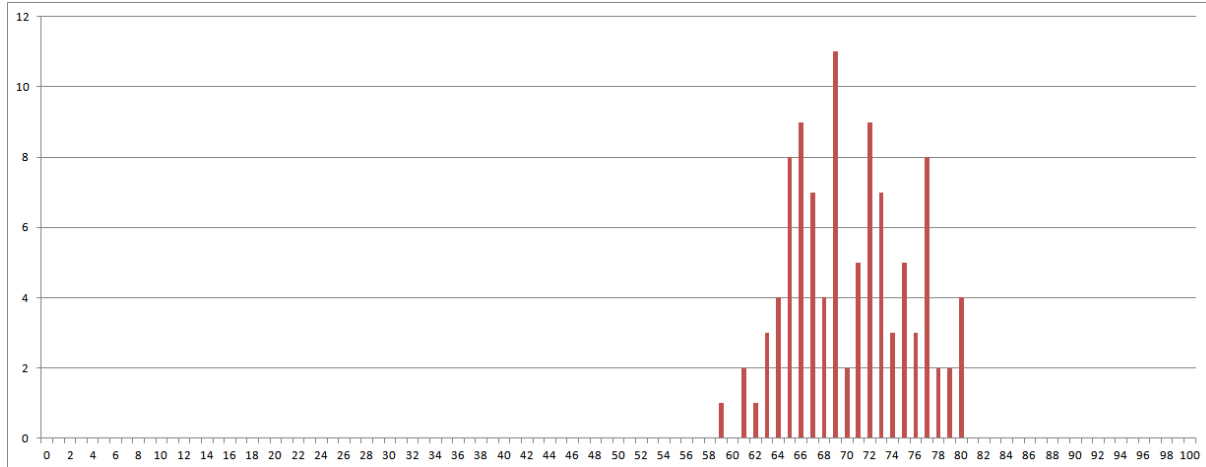
Calcul possible avec un tableur : =LOI.NORMALE.STANDARD(1,696)-LOI.NORMALE.STANDARD(-2,576)

## 2 Intervalle de confiance

### 2.1 Introduction

Cette fois, on sait que  $X$  suit une loi binomiale, on connaît  $n$  mais pas  $p$ , et on peut “observer”  $X$  autant de fois que l’on veut.

Observons !



On a répété l’observation de  $X$  100 fois (sous forme de lancers, tirages, étude épidémiologique, ...).

Questions :

1. Que vaut  $p$  ?
2. Que vaut  $p$  sans prendre de risque ?
3. Que vaut  $p$  en prenant un risque mesuré ?  
(C’est ma définition de l’intervalle de confiance)

Quelques pistes :

- On peut prendre la valeur qui apparaît le plus ( $x_0 = 69$  d’où  $f = 69/100$ ), ce qui est un peu risqué car assez instable.
- On peut prendre une valeur moyenne ( $\bar{x} = 70,31$  d’où  $f = 70,31/100$ ) qui est plus stable.
- On peut prendre un intervalle qui englobe les valeurs les plus probables.

Justifications théoriques :

- a) On montre qu’en moyenne la variable aléatoire  $F_n = \frac{X_n}{n}$  est une bonne approximation de  $p$  car

$$E(F_n) = E(X_n/n) = E(X_n)/n = np/n = p.$$

- b) On peut aussi se dire que si  $X_n$  suit  $\mathcal{B}(n, p)$  alors l’intervalle de fluctuation simplifié pour  $X_n$  est  $[np - \sqrt{n}; np + \sqrt{n}]$ , ou encore

$$P(np - \sqrt{n} < X_n < np + \sqrt{n}) = 0,95.$$

Cette fois, on “connaît”  $X_n$  (tout au moins par quelques réalisations) et on cherche  $p$  qui est inconnue. Alors

$$(np - \sqrt{n} < X_n < np + \sqrt{n}) \Leftrightarrow \left( \frac{X_n}{n} - \frac{1}{\sqrt{n}} < p < \frac{X_n}{n} + \frac{1}{\sqrt{n}} \right).$$

On remplace alors  $\frac{X_n}{n}$  par sa réalisation  $f$ , pour obtenir un intervalle de confiance simplifié pour  $p$  :

$$\left[ f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right].$$

C’est la forme d’intervalle de confiance pour une proportion qui sera présentée en Terminale.

On trouve dans la littérature, un intervalle de confiance “non simplifié” pour la proportion  $p$  de la forme

$$\left[ f - u_\alpha \sqrt{\frac{f(1-f)}{n}}; f + u_\alpha \sqrt{\frac{f(1-f)}{n}} \right] \quad \text{ou} \quad \left[ f - u_\alpha \sqrt{\frac{f(1-f)}{n-1}}; f + u_\alpha \sqrt{\frac{f(1-f)}{n-1}} \right].$$

### Qu'est ce que l'on a finalement fait ?

En connaissant la loi de la variable aléatoire dans une population sans connaître ses caractéristiques (paramètre  $p$  pour une loi de Bernoulli ou binomiale s'il y a répétitions indépendantes), on cherche à estimer ce paramètre à l'aide des réalisations observées dans un échantillon. Pour que cette estimation soit valable, il faut :

1. Que l'échantillon soit représentatif de la population ...
2. Qu'il soit suffisamment grand pour que le remplacement de  $\mathcal{B}$  par  $\mathcal{N}$  n'influe pas trop.
3. Que les réponses ne sont pas biaisées ou perverses (pour les sondages d'opinion) ...

Pour essayer d'être le plus proche possible de la réalité, les instituts de sondage prennent des échantillons de 1000 personnes au minimum, ciblent les personnes (s'ils ont eu trop de réponses de femmes, ils cherchent à équilibrer avec des réponses de la part des hommes) et récupèrent l'information au travers de différents moyens de communication (téléphone, sondage de rue, ...). Ils modifient alors les résultats en fonction de résultats antérieurs, de confrontation à la réalité, etc.. C'est un métier.

**Application 1 :** Sondage d'opinion (Election au 1er tour ?) Un sondage est réalisé pour avoir une tendance du résultat d'une élection entre deux candidats A et B d'une région. Pour un total de 33 000 électeurs, le sondage portant sur 723 personnes interrogées donne 384 voix au candidat A. Dans quelle fourchette se trouve le pourcentage des intentions de vote des électeurs de la région pour le candidat A au seuil de 95% ?

Réponse : Taille de l'échantillon :  $n = 723$  (très supérieur à 25).  
Fréquence du caractère :  $\frac{384}{723} = 0,5311$  arrondie à 0,531 à  $10^{-3}$  près  
Intervalle de confiance au seuil de 95% :

$$\left[ f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right] = \left[ 0,531 - \frac{1}{\sqrt{723}}; 0,531 + \frac{1}{\sqrt{723}} \right] = [0,494; 0,568] \quad (\text{valeurs arrondies à } 10^{-3}).$$

On peut remarquer que les instituts de sondage donnent des pourcentages d'intention de vote, sans indiquer la « fourchette » dans laquelle se trouve cette valeur.

### Application 2 : Marge d'erreur

On peut retrouver le même type d'exercice sous la forme de l'étude de la longueur de l'intervalle de fluctuation, ou la détermination de l'effectif minimal pour une précision donnée.

Exemple : On dispose des observations suivantes  $p = 0.82$ . On cherche à obtenir un intervalle de fluctuation du type  $p \pm 5\%$  au seuil de 99%. Déterminer la taille minimale de l'échantillon pour avoir cette précision donnée.

Solution : on utilise l'intervalle de fluctuation asymptotique

$$IF_{1-\alpha}(p) = \left[ p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

Par identification,

$$u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq 5\%,$$

soit

$$n \geq u_\alpha^2 \frac{p(1-p)}{(5/100)^2}.$$

On applique à l'exercice :  $u_\alpha = 2,576$ ,  $p = 0,82$ , pour trouver  $n \geq 392$ . Le même style d'exercice peut être fait avec des intervalles de confiance.

**Application 3** : Comparaison d'intervalle de confiance (prélude aux tests d'hypothèse).

Exemple : Comparaison du taux de germination de semences de tomates de l'année avec celles de l'année précédente.

Un maraîcher achète un lot de semences de tomates pour produire ses plants de tomate. Il lui reste des semences de l'année passée, dont il doit contrôler le taux de germination pour pouvoir les utiliser avec les autres. En effet, des taux de germination trop différents provoquent des trous dans les plates-bandes de production, ce qui génère un coup de manutention plus élevé (il faut enlever les pots non germés avant de les conditionner). Il faut donc comparer les taux de germination des semences des deux années.

Une stratégie (il en existe d'autres, hors programme, mais qui peuvent faire l'objet d'une recherche) consiste à calculer et à comparer les intervalles de confiance des taux de germination (qui sont des proportions) des plants de l'année et de l'année précédente. Si les deux intervalles ne se recoupent pas, on peut conclure à une différence de taux de germination entre les semences des deux origines. Il faudra alors les semer séparément. Pour faire cette comparaison, le maraîcher prélève, aléatoirement dans les semences de l'année, un échantillon de 200 graines qu'il met à germer. Il constate que 185 graines germent. Il prélève ensuite, aléatoirement dans les semences de l'année précédente, un échantillon de 200 graines qu'il met à germer. Il constate que 150 graines germent.

1. Déterminer un intervalle de confiance, au niveau de confiance de 95%, du taux de germination  $p_a$  du lot de semences de l'année.

*Solution*

$$IC_{95\%}(p_a) = [185/200 - 1/\sqrt{200}; 185/200 + 1/\sqrt{200}] = [0,925 - 0,071; 0,925 + 0,071] \approx [0,85; 0,99]$$

2. Déterminer (par la même méthode qu'à la question a)) un intervalle de confiance au niveau 95%, du taux de germination  $p_b$  du lot de semences de l'année précédente.

*Solution*

$$IC_{95\%}(p_b) = [150/200 - 1/\sqrt{200}; 150/200 + 1/\sqrt{200}] = [0,75 - 0,071; 0,75 + 0,071] \approx [0,68; 0,82]$$

3. Conclure.

*Solution* : les deux intervalles sont disjoints, on peut donc conclure à une différence entre les taux de germination  $p_a$  et  $p_b$  au niveau de confiance 0,95. (ou voir p.69 du document ressource)

**Application 4** : Intervalle de confiance pour une moyenne

La formule de l'intervalle de confiance change si la grandeur observée n'est plus un pourcentage mais une moyenne. Ces intervalles sont à relier avec le paragraphe "Les intervalles « Un, deux, trois sigmas »" (page 11 du document ressource)

Ref : [http://fr.wikipedia.org/wiki/Intervalle\\_de\\_confiance](http://fr.wikipedia.org/wiki/Intervalle_de_confiance)

1. l'intervalle  $\left[\bar{x} - \frac{\sigma(X)}{\sqrt{n}}; \bar{x} + \frac{\sigma(X)}{\sqrt{n}}\right]$  est un intervalle de confiance de la moyenne à un seuil d'environ 68%.
2. l'intervalle  $\left[\bar{x} - 2\frac{\sigma(X)}{\sqrt{n}}; \bar{x} + 2\frac{\sigma(X)}{\sqrt{n}}\right]$  est un intervalle de confiance de la moyenne à un seuil d'environ 95%.
3. l'intervalle  $\left[\bar{x} - 3\frac{\sigma(X)}{\sqrt{n}}; \bar{x} + 3\frac{\sigma(X)}{\sqrt{n}}\right]$  est un intervalle de confiance de la moyenne à un seuil d'environ 99,7%.

**Références en vrac :**

- Brigitte CHAPUT (APMEP), [http://www.apmep.tlse.free.fr/spip/IMG/pdf/Int\\_fluct.pdf](http://www.apmep.tlse.free.fr/spip/IMG/pdf/Int_fluct.pdf)
- Equipe Acad Bordeaux, [http://mathematiques.ac-bordeaux.fr/pedalyc/seqdocped/statproba/dia\\_flu\\_ech/fluct\\_echantillonnage.pdf](http://mathematiques.ac-bordeaux.fr/pedalyc/seqdocped/statproba/dia_flu_ech/fluct_echantillonnage.pdf)
- [http://smf.emath.fr/files/text\\_like\\_files/lyceegressourcesmathproba-stat207115.pdf](http://smf.emath.fr/files/text_like_files/lyceegressourcesmathproba-stat207115.pdf)