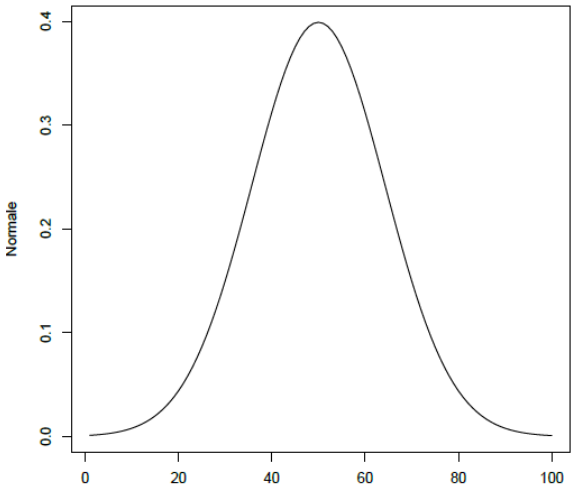
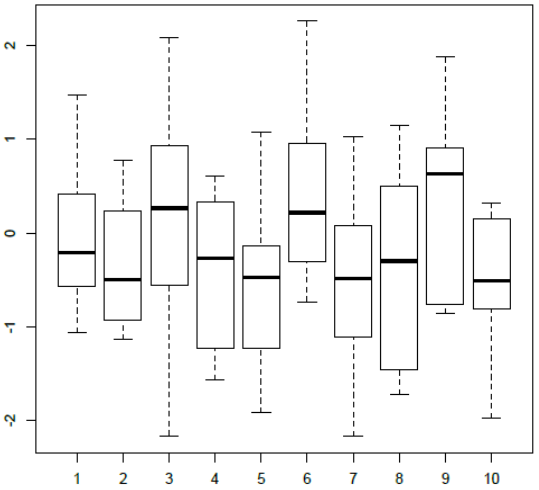
Institut de Recherche sur l'Enseignement des Mathématiques

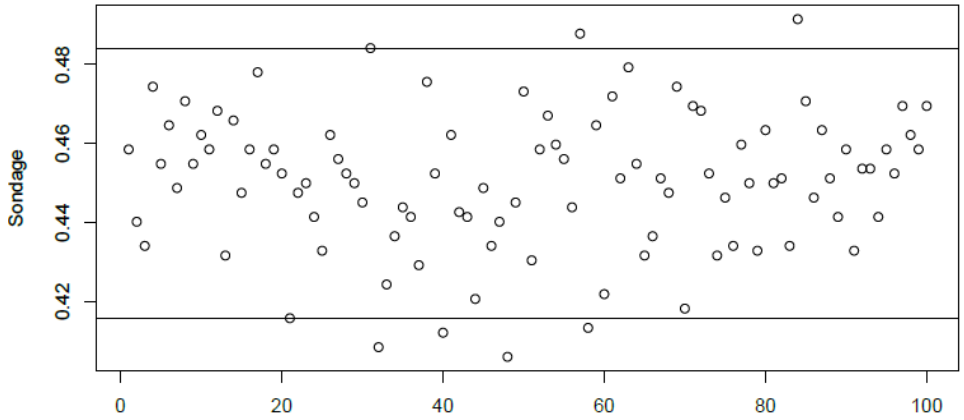
UNIVERSITÉ DE MONTPELLIER

I.R.E.M. de Montpellier

Probabilités et Statistique au Lycée







Auteurs : **BASCOU Noël, BRESSON Daniel, DELATOUR Françoise, LAVERGNE Christian, PLAZEN Marc, SCHADECK Jean-Marie.**

Mars 2016

Adresse : cc40, Place Eugène Bataillon
34095 MONTPELLIER CEDEX 05
Courriel : irem@math.univ-montp2.fr

☎ : (33) 04 67 14 33 83
Fax : (33) 04 67 14 39 09
Site : http://www.irem.univ-montp2.fr

Probabilités et Statistique au Lycée

Ce document a été réalisé par le groupe IREM Probabilités *et Statistique* de Montpellier.

Il est destiné aux enseignants de lycée général, technologique et professionnel. Pour sa conception, nous avons analysé les programmes et leurs différents documents d'accompagnement. Pour répondre aux questionnements de nos collègues, nous proposons ici quelques compléments.

À partir des notions abordées dans les programmes officiels, nous reprenons les fondamentaux de la statistique descriptive, des probabilités et de la statistique inférentielle.

Le document est composé de trois sections.

La première aborde les indices de la statistique univariée, à savoir les quantiles, et les diverses notions de dispersion et de résumés statistiques associés.

La deuxième section aborde la loi binomiale, la loi normale, le théorème de Moivre Laplace, le théorème central limite avec en annexe une preuve de la formule de Stirling et un aperçu sur l'intégrale de Lebesgue.

La troisième section reprend les bases de la statistique inférentielle utiles pour l'enseignement de la statistique au lycée.

Les chapitres peuvent être lu indépendamment les uns des autres, ceci explique les redondances que l'on pourra constater.

Le groupe Probabilités Statistique de l'IREM de Montpellier :

Noël BASCOU

Daniel BRESSON

Françoise DELATOUR

Christian LAVERGNE

Marc PLAZEN

Jean-Marie SCHADECK

Nous remercions tous les collègues rencontrés dans nos établissements et lors de différents stages pour leurs remarques et questions qui ont permis de nourrir notre réflexion.

C'est particulièrement dans les jeux de hasard que paraît la faiblesse de l'esprit humain et la pente qu'il a à la superstition.

Pierre de Rémond de Montmort

mathématicien du XVII^{e} siècle.

Table des matières

PARTIE B : PROBABILITÉS

PARTIE C : STATISTIQUE INFÉRENTIELLE

Partie A

Statistique descriptive

LES QUANTILES

On étudie une variable quantitative X sur un échantillon de taille n . On considère la série statistique $(x_i)_{i=1,..,n}$ ordonnée des valeurs observées du caractère ; valeurs rangées par ordre croissant, la même valeur pouvant être répétée.

1. les quartiles

Intuitivement, l'idée est de partager la série en quatre parties de" même" effectif.

A. Définitions

Les définitions du programme sont les suivantes :

- le **premier quartile** notée $\boldsymbol{q_1}$, est la plus petite valeur d'une série statistique telle qu'au moins 25% des valeurs de la série sont inférieures ou égale à $\boldsymbol{q_1}$.

- le **troisième quartile** notée $\boldsymbol{q_3}$, est la plus petite valeur d'une série statistique telle qu'au moins 75% des valeurs de la série sont inférieures ou égale à $\boldsymbol{q_3}$.

Remarque 1
Le pourcentage des termes dont la valeur est strictement supérieur à q_3 est strictement inférieur à 25%.
En effet, si 25% des données étaient strictement supérieures à q_3, alors moins de 75% (strictement) seraient inférieures à q_3, ce qui n'est pas le cas.
De même le pourcentage des termes dont la valeur est strictement inférieur à q_1 est strictement inférieur à 25%.

Remarque 2
Il existe d'autres définitions qui ne donneront pas nécessairement le même résultat, et en particulier :
le **second quartile** noté $\boldsymbol{q_2}$ **qui** est alors la plus petite valeur de la série telle qu'au moins 50% des valeurs de la série sont inférieures ou égale à $\boldsymbol{q_2}$**,** ne correspond pas à la définition de la médiane du programme.

Cependant la définition du **deuxième quartile** $\boldsymbol{q_2}$ est une définition possible de la médiane m_e, on a alors :

$q_2 = m_e = x_k$ si n est pair donc $n = 2k$
$q_2 = m_e = x_{k+1}$ si n est impair donc $n = 2k+1$

Inversement, si on part de la définition de la médiane du programme, que l'on généralise pour définir les quartiles, on n'obtient pas nécessairement les mêmes résultats. De plus $\boldsymbol{q_1}$ n'est pas systématiquement la médiane de la première moitié de la série.

Revenons sur la médiane :

Définition du programme : la médiane m_e de la série ordonnée $(x_i)_{i=1,..,n}$ est

$m_e = \dfrac{x_k + x_{k+1}}{2}$ si n est pair donc $n = 2k$ (par convention)

$m_e = x_{k+1}$ si n est impair donc $n = 2k+1$

la médiane m_e vérifie alors

$\begin{cases} \textit{au moins 50\% des valeurs de la série sont inférieures ou égale à } m_e \\ \textit{au moins 50\% des valeurs de la série sont supérieures ou égale à } m_e \end{cases}$

Avec cette définition, la valeur de la médiane peut différer de celle du deuxième quartile $\boldsymbol{q_2}$.

La définition de la médiane et celle du deuxième quartile sont cohérentes avec celle qui définit la médiane **comme** la valeur (ou l'une des valeurs) qui minimise l'application définie sur $\mathbb{R}$ par

$$\varphi : x \longrightarrow \sum_{i=1}^{n} |x_i - x| \ .$$

La définition de la médiane et celle du deuxième quartile ne donnent un résultat différent que dans le cas où la taille de la série est paire et si $x_k \neq x_{k+1}$.

Chacune s'adapte à la définition des autres quartiles, et plus généralement des quantiles.

Attention : les tableurs et autres logiciels peuvent donner des résultats différents pour les quartiles car ils utilisent d'autres algorithmes.
Pour le premier quartile $\boldsymbol{q_1}$ et le troisième quartile $\boldsymbol{q_3}$, on obtient donc :

Taille de la série	définition de $\boldsymbol{q_1}$ du programme	définition de $\boldsymbol{q_3}$ du programme
$n = 4p$	x_p	x_{3p}
$n = 4p+1$	x_{p+1}	x_{3p+1}
$n = 4p+2$	x_{p+1}	x_{3p+2}
$n = 4p+3$	x_{p+1}	x_{3p+3}

q_1 est la borne inférieure de l'ensemble des solutions de l'inéquation $\frac{x}{4p+r} \geqslant \frac{1}{4}$, dans $\mathbb{N}^*$,

soit $x \geqslant p + \frac{r}{4}$.

q_3 est la borne inférieure de l'ensemble des solutions de l'inéquation $\frac{x}{4p+r} \geqslant \frac{3}{4}$ dans $\mathbb{N}^*$,

Soit $x \geqslant 3p + \frac{3r}{4}$.

B. On peut alors définir :

a. L'Intervalle interquartile : c'est l'intervalle dont les extrémités sont le premier et le troisième quartiles : $[q_1 ; q_3]$.

Au moins 50% des termes de la série ont une valeur qui appartiennent l'intervalle $[q_1 ; q_3]$

en effet :

Taille de la série	proportion minimale des termes dont la valeur appartient à $[q_1 ; q_3]$
$n = 4p$	$\frac{3p-p+1}{4p} = \frac{2p+1}{4p} = \frac{1}{2} + \frac{1}{4p}$
$n = 4p+1$	$\frac{(3p+1)-(p+1)+1}{4p+1} = \frac{2p+1}{4p+1} = \frac{1}{2} + \frac{1}{2}\frac{1}{4p+1}$
$n = 4p+2$	$\frac{(3p+2)-(p+1)+1}{4p+2} = \frac{2p+2}{4p+2} = \frac{1}{2} + \frac{1}{4p+2}$
$n = 4p+3$	$\frac{(3p+3)-(p+1)+1}{4p+3} = \frac{2p+3}{4p+3} = \frac{1}{2} + \frac{3}{2}\frac{1}{4p+3}$

Ainsi le pourcentage des termes dont la valeur appartiennent à $[q_1 ; q_3]$ est **strictement** supérieur à 50%.

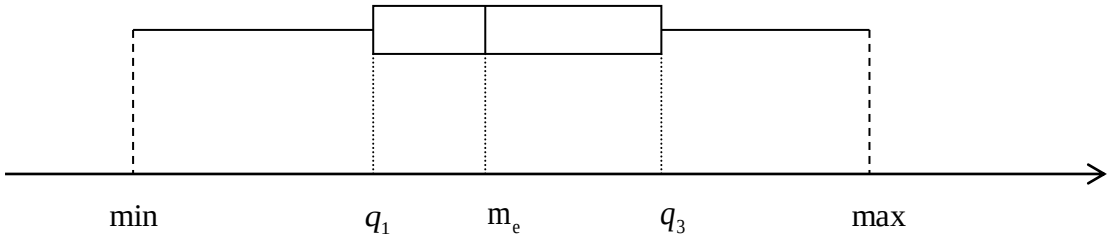
b. L'Écart interquartile, qui est la différence entre le troisième et le premier quartile $q_3 - q_1$, est une « mesure » de la dispersion de la série.

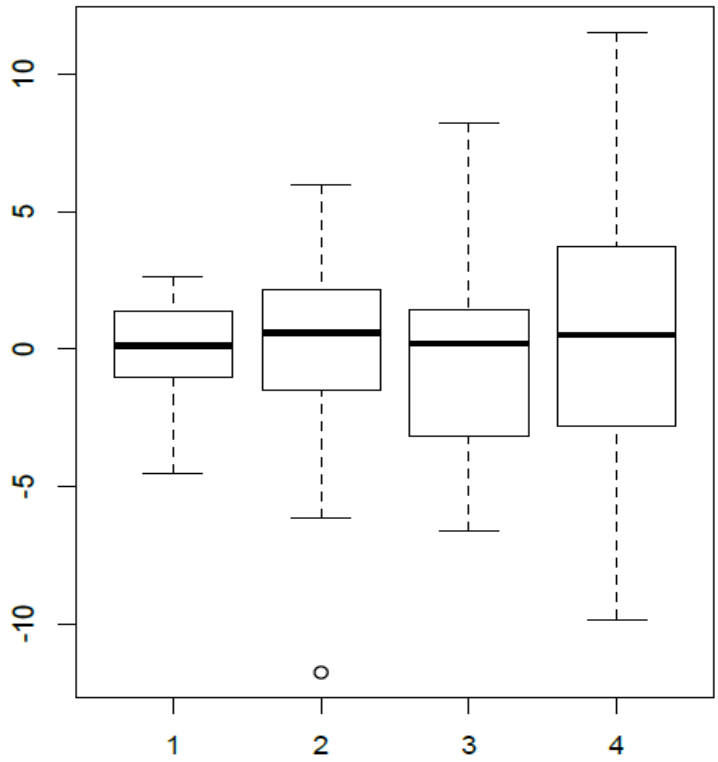
c. Le couple (médiane ; intervalle interquartile) est un résumé possible d'une série statistique

d. Le diagramme de Tuckey ,ou « **diagramme en boîte** » ou « **box plot** » permet de visualiser une série, de comparer facilement plusieurs séries d'une même variable entre elles.

Il permet aussi de mettre en évidence une dissymétrie éventuelle de la série.

Exemples :





Remarque

Il se peut que la médiane et l'un ou l'autre des quartiles, ou les deux, soient confondus.

C. Avantages et inconvénient

Par construction, le couple (médiane ; intervalle interquartile) donne une bonne idée de la répartition des données autour de la valeur centrale, de plus, les données extrêmes (ou aberrantes) n'influent pas sur la valeur des quantiles.

Un inconvénient de ces résumés est que l'on ne peut pas obtenir les résumés du regroupement de plusieurs séries à partir des résumés de chacune des séries initiales.

2. Les déciles

On partage cette fois la série des observations en 10 parties.

Définition du décile d_k :

- pour *k* de l à 9, le *kième* décile noté $\boldsymbol{d_k}$ est la plus petite valeur d'une série statistique telle qu'au moins (*k* ×10) % des valeurs de la série sont inférieures ou égales à $\boldsymbol{d_k}$.

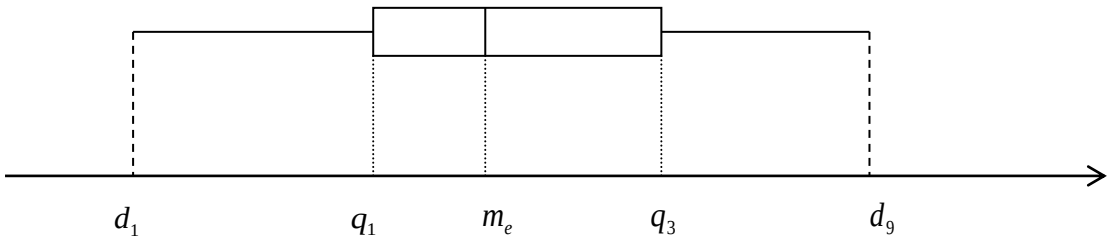
Et on définit de même :

a. L'intervalle de dispersion $[d_1 ; d_9]$ contient au moins 80% des termes de la série.

De la même façon l'intervalle de dispersion $[d_2 ; d_8]$ contient au moins 60% des termes de la série...

Ces intervalles sont centrés en pourcentage autour de la médiane mais pas en valeur.

b. On peut modifier le diagramme en boîte en remplaçant les valeurs extrêmes (min et max) par le premier et le dernier décile de manière à gommer sur le graphique les excès de la distribution



Enfin on peut définir d'autres quantiles de la même façon comme les centiles...

3. Transformation affine

Il est immédiat qu'une transformation affine conserve les quantiles. Par exemple :

- pour les quartiles :

  si $(x'_i ; n_i) = (ax_i + b ; n_i)$ $\quad \forall \mathrm{i} \in \{1;...;\mathrm{p}\}$ où $a \in \mathbb{R}_+$ et $b \in \mathbb{R}$,

  alors : $q'_1 = aq_1 + b$ *et* $q'_3 = aq_3 + b$

- pour l'écart interquartile :

  $q'_3 - q'_1 = (aq_3 + b) - (aq_1 + b) = a(q_3 - q_1)$

4. Exemple

On mesure la masse, en grammes, des lettres postales d'un échantillon de taille 500 :

x_i (en gr)	15	16	17	18	19	20	21	22	25	30
n_i	12	10	40	78	120	110	70	30	20	10
$\sum n_i$	12	22	62	140	260	370	440	470	490	500
Fréq cumul	0,024	0,044	0,124	0,28	0,52	0,74	0,88	0,94	0,98	1

- $q_1 = 18$, $q_3 = 21$ et $m_e = 19$ (faire la boîte)

$\frac{440-62}{500} = \frac{378}{500} = 0,756$ donc 75,6% des individus sont dans l'intervalle interquartile.

- $d_1 = 17$ et $d_9 = 22$

$\frac{470-22}{500} = \frac{448}{500} = 0,896$ donc 89,6% des individus sont dans l'intervalle de dispersion à 80%, $[d_1 ; d_9]$.

- le 5^e centile est $c_5 = 17$, le 95^e centile est $c_{95} = 25$.

RÉSUMÉS NUMÉRIQUES D'UNE SÉRIE STATISTIQUE

I. Construction de quelques résumés numériques

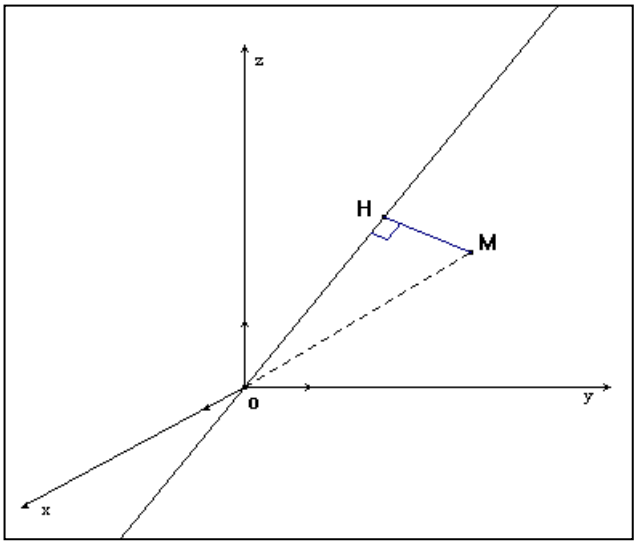
On cherche à « mesurer » la dispersion de la série $(x_1, x_2, ..., x_n)$.

L'idée est de la comparer à une série non dispersée, une série non dispersée est une série dont les *n* valeurs sont égales, c'est à dire du type : $(x, x, ..., x)$.

Et, de plus, de la comparer à la série non dispersée la plus proche d'elle, qu'il faut donc déterminer. Cette proximité se mesure à l'aide d'une distance *d*, qu'il faut choisir a priori, puis on **minimise** $d\left((x_1, x_2, ..., x_n);(x, x, ..., x)\right)$ qui est une fonction de la variable *x*.

1. Avec la distance d_2 : si $A(x_1;...;x_n)$ et $B(y_1;...;y_n)$, $d_2(A,B) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$

a. En dimension 3 (la taille de la série est 3)



On regarde la « série » $(x_1; x_2; x_3)$ et on cherche la « série » $(x ; x ; x)$ la plus « proche » , et on mesure cette proximité à l'aide de la distance entre les 2 points *M* et *H* dans l'espace « euclidien » en interprétant ces séries comme les coordonnées de ces points :

(*H* est le projeté orthogonal de *M* sur la trisectrice)

on cherche donc le *x* qui minimise

$d_2(H,M) = \left\|\overrightarrow{HM}\right\| = \sqrt{\sum_{i=1}^{3}(x_i - x)^2}$.

donc qui minimise $\left\|\overrightarrow{HM}\right\|^2$,

c'est à dire l'application définie sur $\mathbb{R}$ par $x \to \sum_{i=1}^{3}(x_i - x)^2$.

Pour caractériser le point *H* on peut utiliser le théorème de Pythagore ou le produit scalaire ou la dérivée ou la forme canonique. Avec le premier, par exemple, on obtient :

$$\left\|\overrightarrow{OM}\right\|^2 = \left\|\overrightarrow{OH}\right\|^2 + \left\|\overrightarrow{HM}\right\|^2 \qquad \text{soit}$$

$$x_1^2 + x_2^2 + x_3^2 = 3x^2 + (x_1 - x)^2 + (x_2 - x)^2 + (x_3 - x)^2$$

D'où, après calculs $x = \dfrac{x_1 + x_2 + x_3}{3}$ soit $x = \overline{x}$,

et donc pour le carré de la distance : $\left\| \overrightarrow{HM} \right\|^2 = \sum_{i=1}^{3} (x_i - \overline{x})^2 = \sum_{i=1}^{3} x_i^2 - 3\overline{x}^2$

b. Cas général : en dimension *n* (la taille de la série est *n*)

On considère alors l'application $\varphi : x \longrightarrow \sum_{i=1}^{n} (x_i - x)^2$ définie de $\mathbb{R}$ dans $\mathbb{R}$, qui mesure la dispersion de la série autour de la valeur *x*, et on cherche alors à minimiser cette application.

$\varphi(x) = nx^2 - 2\left(\sum_{i=1}^{n} x_i \right) x + \sum_{i=1}^{n} x_i^2$ est une fonction polynôme du second degré, sa représentation graphique est une parabole, et avec par exemple la forme canonique (chère aux élèves) on obtient facilement le minimum:

$$\sum_{i=1}^{n} (x_i - x)^2 = nx^2 - 2\left(\sum_{i=1}^{n} x_i \right) x + \sum_{i=1}^{n} x_i^2 = n\left(x - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 + \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2$$

le minimum est donc atteint pour : $x = \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i \right) = \overline{x}$, moyenne arithmétique des x_i.

et ce minimum est : $\sum_{i=1}^{n} (x_i - \overline{x})^2$. ou $\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$

Mais comme on comparera des séries de tailles différentes,

on multiplie la dispersion autour de la moyenne $\sum_{i=1}^{n} (x_i - \overline{x})^2$ par $\dfrac{1}{n}$ (pour annuler l'effet de taille), ce qui revient à faire la dispersion moyenne autour de la moyenne.

On utilisera donc l'expression : $\dfrac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$

c'est la **variance empirique** qui est la moyenne des carrés des écarts à la moyenne.

C'est aussi la différence entre « la moyenne des carrés » et « le carré de la moyenne » à l'aide de laformule de Huyghens-König,

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2$$

formule qui nous fournit une inégalité intéressante en elle même

de plus, comme la variance est positive, on obtient $\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \leqslant \frac{1}{n}\sum_{i=1}^{n} x_i^2$

soit $\left(\frac{x_1 + x_2 + ... + x_n}{n}\right)^2 \leqslant \frac{x_1^2 + x_2^2 + ... + x_n^2}{n}$

Enfin, pour revenir à l'unité de mesure des données : on considère l'expression :

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

c 'est l'**écart-type** qui est un indice de dispersion de la série.

Conclusion :

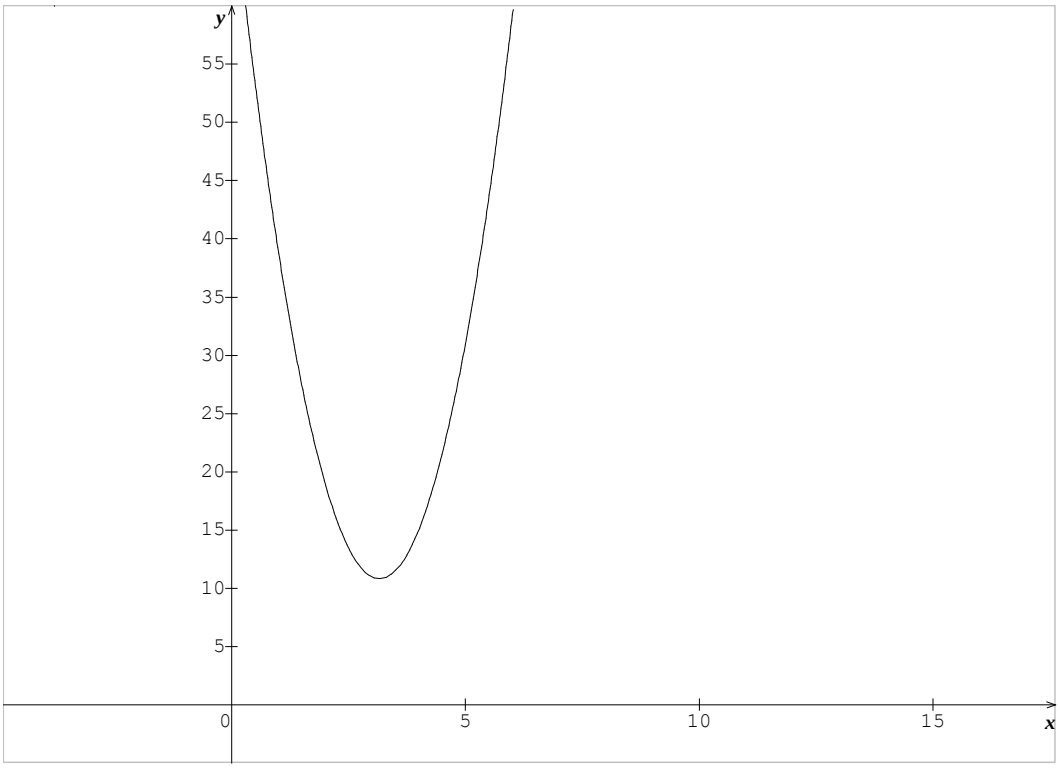
Avec la distance d_2, le couple de résumés numériques associé à la série est le couple

(moyenne , écart-type) : $\left(\ \overline{x}\ ;\ s(x)\ \right)$

Remarquons que cette distance est euclidienne (issue d'un produit scalaire), d'où son intérêt.

graphiquement

pour la série $\left(\ 1\ ;\ 2\ ;\ 3\ ;\ 4\ ;\ 4\ ;\ 5\ \right)$, la courbe de φ est



2. Avec la distance d_∞ : *si* $A(x_1;...;x_n)$ *et* $B(y_1;...;y_n)$, $d_\infty(A,B)=\sup\limits_{1\leqslant i\leqslant n}|y_i-x_i|$

et puique *n* est fini, $d_\infty(A,B)=\max\limits_{1\leqslant i\leqslant n}|y_i-x_i|$

On supposera de plus que la série $(x_1,x_2,...,x_n)$ est ordonnée: $x_1\leqslant x_2\leqslant ...\leqslant x_n$.

On considère l'application $\varphi : x\longrightarrow \max\limits_{1\leqslant i\leqslant n}|x-x_i|$ définie de $\mathbb{R}$ dans $\mathbb{R}$,

distance de la série $(x_1,x_2,...,x_n)$ à la série $(x,x,...,x)$ d'étendue nulle (non dispersée).

et on cherche alors à minimiser cette application :

- par définition, $\forall x\in\mathbb{R}$, $|x-x_j|\leqslant \max\limits_{1\leqslant i\leqslant n}|x-x_i| \qquad \forall j\in 1;n$
- on remarque que pour le milieu $c=\dfrac{x_1+x_n}{2}$, $\max\limits_{1\leqslant i\leqslant n}|c-x_i|=|c-x_1|=|c-x_n|=\dfrac{x_n-x_1}{2}$

Supposons maintenant que x rende minimum $\max\limits_{1\leqslant i\leqslant n}|x-x_i|$

alors en particulier : $\max\limits_{1\leqslant i\leqslant n}|x-x_i|\leqslant \max\limits_{1\leqslant i\leqslant n}|c-x_i|$

donc $\forall j\in 1;n \qquad |x-x_j|\leqslant \max\limits_{1\leqslant i\leqslant n}|x-x_i|\leqslant \dfrac{x_n-x_1}{2}$

et en particulier encore :

$$|x-x_1|\leqslant \frac{x_n-x_1}{2} \qquad \text{et} \qquad |x-x_n|\leqslant \frac{x_n-x_1}{2}$$

$$-\left(\frac{x_n-x_1}{2}\right)\leqslant x-x_1\leqslant \frac{x_n-x_1}{2} \qquad \text{et} \qquad -\left(\frac{x_n-x_1}{2}\right)\leqslant x-x_n\leqslant \frac{x_n-x_1}{2}$$

$$x_1-\left(\frac{x_n-x_1}{2}\right)\leqslant x\leqslant x_1+\frac{x_n-x_1}{2} \quad \text{et} \quad x_n-\left(\frac{x_n-x_1}{2}\right)\leqslant x\leqslant x_n+\frac{x_n-x_1}{2}$$

$$\frac{-x_n+3x_1}{2}\leqslant x\leqslant \frac{x_n+x_1}{2} \qquad \text{et} \qquad \frac{x_n+x_1}{2}\leqslant x\leqslant \frac{3x_n-x_1}{2}$$

d'où $x = \frac{x_1 + x_n}{2}$, donc $x = c$ et le minimum est la demi-étendue $\frac{x_n - x_1}{2}$

Conclusion :

avec la distance d_∞ , le couple de résumés numériques associé à la série est le couple

(milieu des extrêmes, demi-étendue).

Remarque : On peut aussi expliciter φ :

- $x < c$

  $\forall\, i$ tel que $x_i \leqslant c$

  $$|x - x_i| = |x - c + c - x_i| \leqslant |x - c| + |c - x_i| = c - x + c - x_i = x_1 + x_n - x - x_i \leqslant x_n - x$$

  $\forall\, i$ tel que $x_i > c$

  $$|x - x_i| = |x - c + c - x_i| \leqslant |x - c| + |c - x_i| = c - x + x_i - c - = x_i - x \leqslant x_n - x$$

  donc $x < c \Rightarrow \varphi(x) = x_n - x$

- $c \leqslant x$

  $\forall\, i$ tel que $x_i \leqslant c$

  $$|x - x_i| = |x - c + c - x_i| \leqslant |x - c| + |c - x_i| = x - c + c - x_i = x - x_i \leqslant x - x_1$$

  $\forall\, i$ tel que $x_i > c$

  $$|x - x_i| = |x - c + c - x_i| \leqslant |x - c| + |c - x_i| = x - c + c - x_i = x - x_i \leqslant x - x_1$$

  donc $c \leqslant x \Rightarrow \varphi(x) = x - x_1$

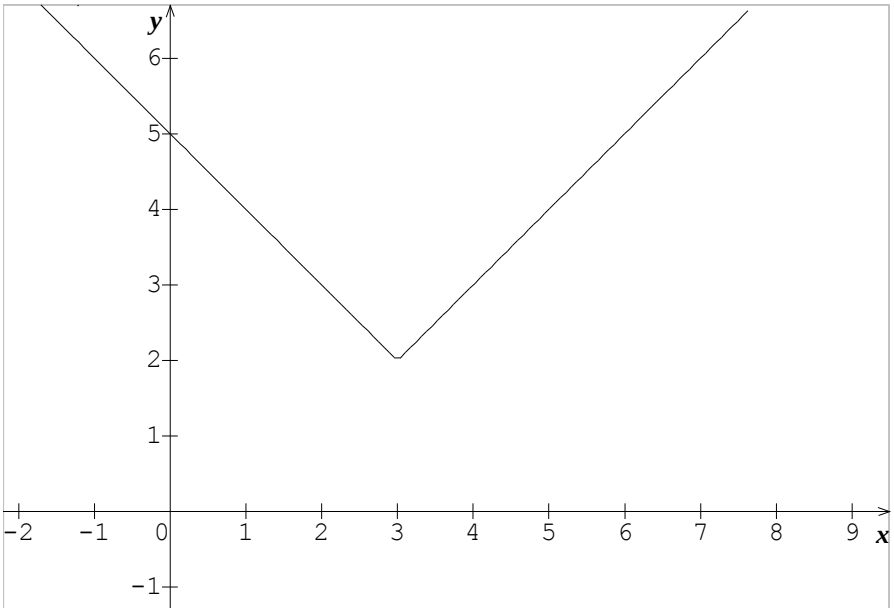
en résumé φ est définie sur $\mathbb{R}$ par

$\begin{cases} \varphi(x) = x_n - x & x < c \\ \varphi(x) = x - x_1 & x \geqslant c \end{cases}$ ou encore $\varphi(x) = |x - c| + \frac{x_n - x_1}{2}$ le minimum est alors en évidence.

la fonction ne dépend que des extrêmes , quelque soit la répartition des valeurs intercalées!

graphiquement

pour les séries $(1;2;3;4;4;5)$; $(1;...;...;...;...;5)$..., la courbe de φ est



3. Avec la distance d_1 : *si* $A(x_1;...;x_n)$ *et* $B(y_1;...;y_n)$, $d_1(A,B)=\sum_{i=1}^{n}|y_i-x_i|$

On supposera également que la série $(x_1,x_2,...,x_n)$ est ordonnée :

$$x_1 \leqslant x_2 \leqslant ... \leqslant x_n .$$

On considère l'application $\varphi : x \longrightarrow \sum_{i=1}^{n}|x_i - x|$ définie de $\mathbb{R}$ dans $\mathbb{R}$.

distance d_1 de la série $(x_1,x_2,...,x_n)$ à la série $(x,x,...,x)$ d'étendue nulle (non dispersée).

et on cherche à minimiser cette application.

- Recherche du minimum :

- tout d'abord, φ est continue sur $\mathbb{R}$, car somme de fonctions continues sur $\mathbb{R}$

(et minorée par 0) et on peut expliciter φ comme suit,

si $\ldots x_{j-1} \leqslant x_j \leqslant x_{j+1} \leqslant \ldots \leqslant x_{j+p} < x_{j+p+1}$, pour $j+p \leqslant n-1$,

$$x \in \left[x_j ; x_{j+p} \right[\Rightarrow \sum_{i=1}^{n} |x_i - x| = \sum_{i=1}^{j+p} (x - x_i) + \sum_{i=j+p+1}^{n} (x_i - x)$$

donc $\varphi(x) = \sum_{i=1}^{n} |x_i - x| = (2j + 2p - n)x + \sum_{i=j+p+1}^{n} x_i - \sum_{i=1}^{j+p} x_i$

en particulier si $x_j < x_{j+1}$, pour $1 \leqslant j \leqslant n-1$,

$$x \in \left[x_j ; x_{j+1} \right[\Rightarrow \sum_{i=1}^{n} |x_i - x| = \sum_{i=1}^{j} (x - x_i) + \sum_{i=j+1}^{n} (x_i - x)$$

$$\varphi(x) = \sum_{i=1}^{n} |x_i - x| = (j - n + j)x - \sum_{i=1}^{j} x_i + \sum_{i=j+1}^{n} x_i = (2j - n)x + \sum_{i=j+1}^{n} x_i - \sum_{i=1}^{j} x_i$$

et comme $\sum_{i=1}^{n} |x_i - x_{j+1}| = \sum_{i=1}^{j} (x_{j+1} - x_i) + \sum_{i=j+1}^{n} (x_i - x_{j+1})$,on a encore

$$x \in \left[x_j ; x_{j+1} \right] \Rightarrow \varphi(x) = \sum_{i=1}^{j} (x - x_i) + \sum_{i=j+1}^{n} (x_i - x) = (2j - n)x + \sum_{i=j+1}^{n} x_i - \sum_{i=1}^{j} x_i \quad \text{pour}$$

$j \in 1 ; n-1$

de plus cette expression reste vraie même si $x_j = x_{j+1}$ (mais s'applique alors sur un intervalle de mesure nulle)

ainsi $\varphi(x) = \sum_{i=1}^{n} |x_i - x| = a_j x + b_j$ pour $x \in \left[x_j ; x_{j+1} \right]$ $\forall\, j \in 1 ; n-1$

En résumé, φ est affine par morceaux et on peut en déduire les variations de φ en regardant le signe

des a_j. Chaque a_j est égal à la différence entre le nombre de termes qui s'écrivent $(x - x_i)$ et le nombre de termes qui s'écrivent $(x_i - x)$ sur l'intervalle considéré.

Minimiser φ revient à chercher la (ou les) valeur de j pour laquelle $a_j \leqslant 0$ et $a_{j+1} \geqslant 0$

▪ si $x \in]-\infty ; x_{k+1}[$ comme $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_k \leqslant x_{k+1} \leqslant \ldots \leqslant x_n$

en regardant la somme $|x_1 - x| + |x_2 - x| + \ldots + |x_k - x| + |\boldsymbol{x_{k+1}} - \boldsymbol{x}| + \boldsymbol{\ldots} + |\boldsymbol{x_n} - \boldsymbol{x}|$

on constate qu'il y a **au moins** $n-k$ termes qui s'écrivent $(\boldsymbol{x_i}-\boldsymbol{x})$, donc il y a **au plus** k termes qui s'écrivent $(x-x_i)$, ainsi $a_j \leqslant k-(n-k)=2k-n \quad \forall \quad j \leqslant k$

Ainsi il faut distinguer deux cas:

1er cas : $\boldsymbol{n=2k+1}$

$a_j \leqslant 2k-n=-1<0$ pour $1 \leqslant j \leqslant k$

donc φ est décroissante strictement sur $]-\infty;x_{k+1}[$

- $\varphi(x_{k+1})=\sum_{i=1}^{n}|x_i-x_{k+1}|=\sum_{i=k+2}^{n}x_i-\sum_{i=1}^{k}x_i$
- enfin pour $x \in]x_{k+1};+\infty[$

on regarde $|x_1-x|+|x_2-x|+...+|x_k-x|+|x_{k+1}-x|+|\boldsymbol{x_{k+2}}-\boldsymbol{x}|+...+|\boldsymbol{x_{2k+1}}-\boldsymbol{x}|$

il y a **au moins** $k+1$ termes qui s'écrivent $(x-x_i)$ (et donc k termes qui s'écrivent $(\boldsymbol{x_i}-\boldsymbol{x})$)

ainsi $a_j \geqslant 2(k+1)-n=1>0$ pour $k+1 \leqslant j \leqslant n-1$

donc φ est croissante strictement sur $]x_{k+1};+\infty[$

le minimum est donc atteint pour $\boldsymbol{x_{k+1}}$

2ème cas $\boldsymbol{n=2k}$

- si $x \in]-\infty;x_k[$ on a $|x_1-x|+|x_2-x|+...+|x_{k-1}-x|+|\boldsymbol{x_k}-\boldsymbol{x}|...+|\boldsymbol{x_{2k}}-\boldsymbol{x}|$

il y a **au moins** $k+1$ termes qui s'écrivent $(\boldsymbol{x_i}-\boldsymbol{x})$, donc **au plus** $k-1$ termes qui s'écrivent $(x-x_i)$,

ainsi $a_j \leqslant k-1-(k+1)=-2<0$ pour $1 \leqslant j \leqslant k-1$

donc φ est décroissante strictement sur $]-\infty;x_k[$

- $\varphi(x_k)=\sum_{i=1}^{n}|x_i-x_k|=\sum_{i=k+1}^{n}x_i-\sum_{i=1}^{k}x_i$ (ou encore $\sum_{i=k+2}^{n}x_i-\sum_{i=1}^{k-1}x_i$)
- si $x \in]x_k;+\infty[$ on a $|x_1-x|+|x_2-x|+...+|x_k-x|+|\boldsymbol{x_{k+1}}-\boldsymbol{x}|...+|\boldsymbol{x_{2k}}-\boldsymbol{x}|$

il y a **au moins** k termes qui s'écrivent $(x-x_i)$ et donc **au plus** k termes qui s'écrivent $(\boldsymbol{x_i}-\boldsymbol{x})$,

et ainsi $a_j \geqslant k-k=0$ pour $k \leqslant j \leqslant n-1$

et on doit à nouveau distinguer deux sous cas:

1er sous cas : $\boldsymbol{a_k = k-k=0}$ ce qui impose $x_k < x_{k+1}$

▪ pour $x \in [x_k ; x_{k+1}]$:

◦ si $x = x_k$

$$\varphi(x_k)=(x_k-x_1)+...+(x_k-x_{k-1})+(x_k-x_k)+\boldsymbol{(x_{k+1}-x_k)...+(x_{2k}-x_k)}=\sum_{i=k+1}^{n}x_i-\sum_{i=1}^{k}x_i$$

◦ si $x \in]x_k ; x_{k+1}[$

$$\varphi(x)=(x-x_1)+...+(x-x_k)+\boldsymbol{(x_{k+1}-x)...+(x_{2k}-x)}=\sum_{i=k+1}^{n}x_i-\sum_{i=1}^{k}x_i \quad \text{(puisque } a_k=0\text{)}$$

◦ si $x = x_{k+1}$

$$\varphi(x_{k+1})=(x_{k+1}-x_1)+...+(x_{k+1}-x_k)+\boldsymbol{(x_{k+1}-x_{k+1})+(x_{k+2}-x_{k+1})+...+(x_{2k}-x_{k+1})}$$

$$=\sum_{i=k+1}^{n}x_i-\sum_{i=1}^{k}x_i$$

ainsi φ est constante sur $[x_k ; x_{k+1}]$.

▪ pour $x \in]x_{k+1};+\infty[$ on a $|x_1-x|+|x_2-x|+...+|x_{k+1}-x|+\boldsymbol{|x_{k+2}-x|...+|x_{2k}-x|}$

il y a **au moins** $k+1$ termes qui s'écrivent $(x-x_i)$ et donc **au plus** $k-1$ termes qui s'écrivent $(\boldsymbol{x_i}-\boldsymbol{x})$,

ainsi $a_j \geqslant (k+1)-(k-1)=2>0$ pour $k+1 \leqslant j \leqslant n-1$

et φ est croissante strictement sur $]x_{k+1};+\infty[$.

le minimum est donc atteint pour toutes les valeurs de $[\boldsymbol{x_k ; x_{k+1}}]$,

2ème sous cas : $\boldsymbol{a_k > 0}$ ce qui impose que au moins k termes qui s'écrivent $(x-x_i)$

donc que $\boldsymbol{x_k = x_{k+1}}$, alors

▪ si $x \in]x_k ;+\infty[$ $a_j > 0$ pour $k \leqslant j \leqslant n-1$

donc φ est croissante strictement sur $]x_k ; +\infty[$

le minimum est donc atteint pour x_k **:** $\boldsymbol{x = x_k = x_{k+1}}$

Définition

la **médiane** m_e d'une série ordonnée $(x_i)_{i=1,..,n}$ est la valeur (ou l'une des valeurs) pour laquelle l'application φ atteint son minimum sur $\mathbb{R}$.

Dans le cas où *n* est impair ($\boldsymbol{n = 2k+1}$) le minimum est atteint pour $\boldsymbol{x_{k+1}}$.

Dans le cas où *n* est pair ($\boldsymbol{n = 2k}$) le minimum est atteint pour toutes les valeurs de $[x_k ; x_{k+1}]$, chacune peut être choisie pour être la médiane.

Une convention est de choisir le milieu du segment : $\boldsymbol{m_e = \frac{x_k + x_{k+1}}{2}}$, c'est celle du programme officiel !

Dans le cas où $\frac{x_k + x_{k+1}}{2}$ n'est pas un observable de la variable, certains auteurs préconisent de choisir pour la valeur de la médiane une valeur observable de $[x_k ; x_{k+1}]$.

En définitive, on obtient pour les classes de lycée :

Définition :

la médiane d'une série ordonnée $(x_i)_{i=1,..,n}$ **est** $\boldsymbol{m_e = x_{k+1}}$ **si** $\boldsymbol{n = 2k+1}$

$\boldsymbol{m_e = \frac{x_k + x_{k+1}}{2}}$ **si** $\boldsymbol{n = 2k}$

Enfin, le minimum de la fonction φ est $\sum_{i=1}^{n} |x_i - m_e|$.

Pour annuler l'effet de taille, on définit donc

l' écart moyen à la médiane : $\frac{1}{n}\sum_{i=1}^{n} |x_i - m_e|$ comme résumé de la dispersion de la série.

(ce n'est pas $\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{x}\right|$ *, un autre « écart moyen absolu » : moyenne des écarts à la moyenne).*

Conclusion :

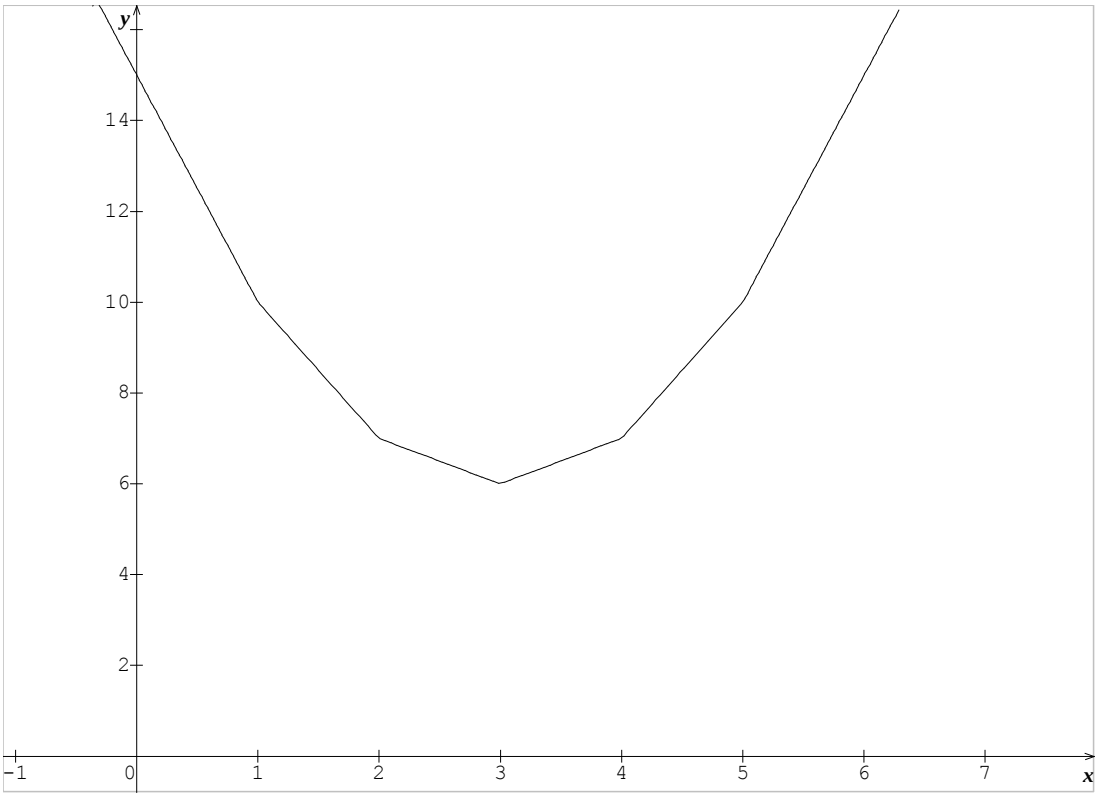
avec la distance d_1, le couple de résumés numériques associé à la série est le couple

(médiane, écart moyen à la médiane).

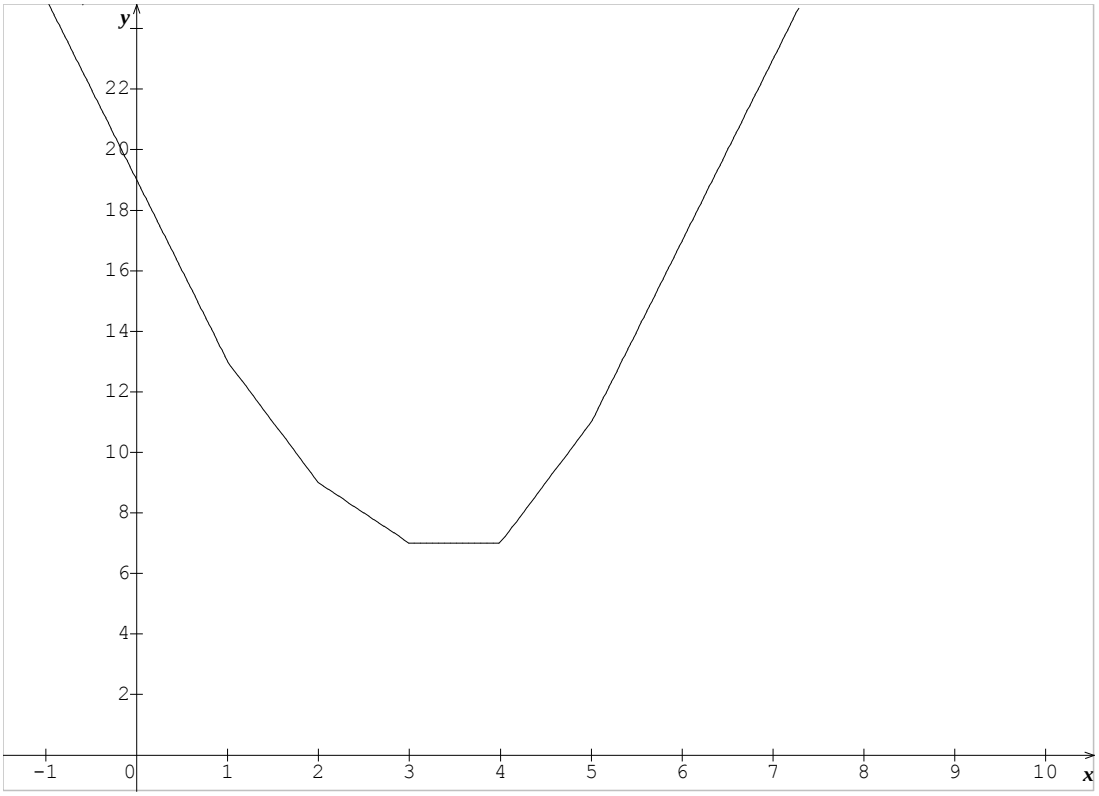
Attention : l'expression $\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\overline{x}\right|$ *n'est pas un indice statistique puisque la moyenne est associée à la distance* d_2.

Graphiquement, les choses sont plus simples :

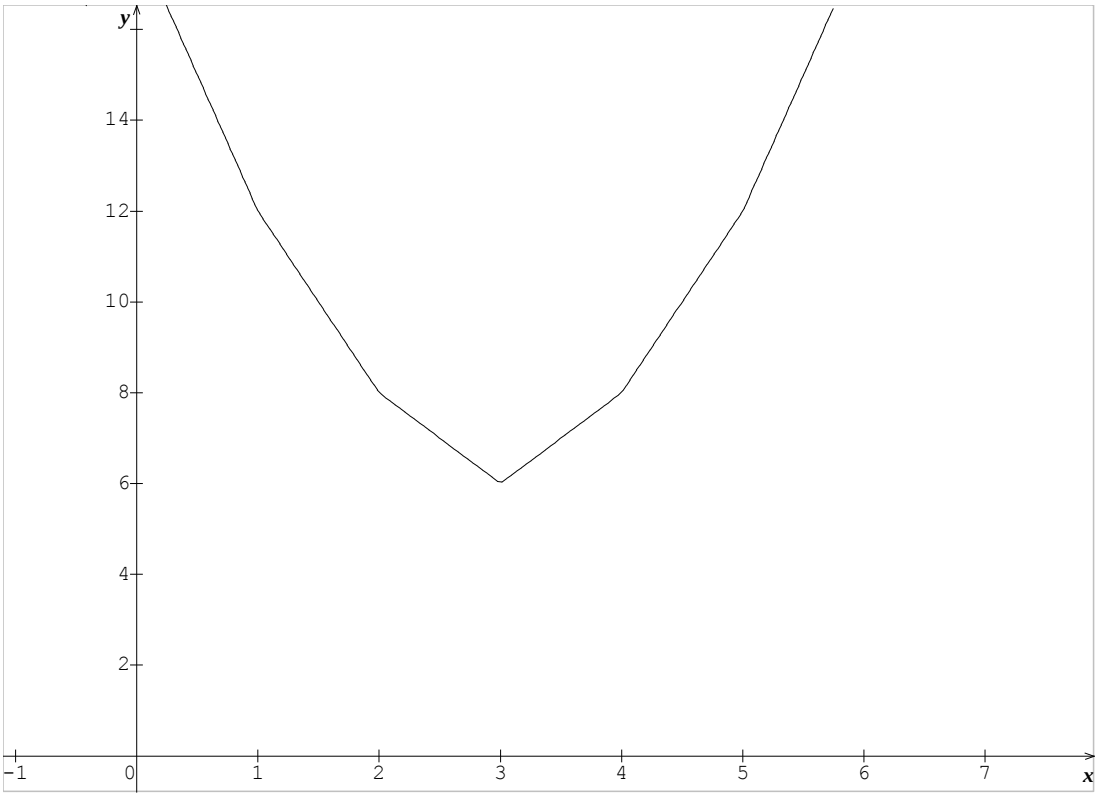
1er cas : $n=2k+1$: pour la série $\left(1\,;2\,;3\,;4\,;5\right)$ la courbe de φ est



2ème cas $n = 2k$ et $x_k < x_{k+1}$: pour la série $\left(1 ; 2 ; 3 ; 4 ; 4 ; 5 \right)$ la courbe de φ est



3ème cas $n = 2k$ et $x_k = x_{k+1}$: pour la série $\left(1 ; 2 ; 3 ; 3 ; 4 ; 5 \right)$ la courbe de φ est



II. Différentes façons de résumer une série

On dispose maintenant de nouveaux couples permettant de résumer une série statistique. On précise ici les avantages et les inconvénients de chacun afin de privilégier l'un ou l'autre de ces résumés suivant la situation considérée et les **objectifs du statisticien**.

1. Couples de résumés construits à l'aide de la même distance

Le couple (moyenne ; écart-type) permet de faire des calculs sur des regroupements et d'avoir une idée assez précise de la répartition de l'effectif pour une série symétrique. Par contre il est sensible aux valeurs extrêmes.

En pratique : ce couple est fondamental dans les nombreux domaines où les grandeurs sont modélisées par des variables gaussiennes (biologie, agronomie, industrie,…)

En théorie **:** ce couple intervient dans de nombreux théorèmes en Statistique inférentielle, (estimation, tests d'hypothèse, analyse de la variance, théorème central de la limite,…)
Le couple (moyenne ; écart-type) est « performant » pour les séries dont la distribution est proche d'une « distribution normale », dont l'une des caractéristique est la symétrie par rapport à la moyenne qui est aussi le mode, il l'est beaucoup moins pour des séries fortement asymétriques.

Le couple (milieu des extrêmes, demi-étendue) est peu employé semble-t-il.
On ne sait rien sur la distribution des observations : toutes les données sont dans l'intervalle , milieu des extrêmes $\pm$ demi- étendue c'est à dire $\left[x_1 ; x_n \right]$!

Le couple (médiane, écart moyen à la médiane) est moins utilisé, se prêtant mal aux calculs. Peu d'indications sur la distribution, mais l'écart moyen à la médiane peut compléter les indications données par les quartiles.

2. Autres types de résumés

Le couple (médiane ; étendue) est le plus simple. Il donne des indications « grossières » (notamment dans le cas où la série comporte des valeurs aberrantes : un seul individu peut modifier considérablement l'étendue !). Il est utilisé en général sur de petits échantillons, souvent pour faire du contrôle de qualité ou des tests « zéro défaut ».

Le couple (médiane ; intervalle interquartile) indique le « centre » de la distribution (médiane) et permet de calculer la longueur de l'intervalle contenant au moins la moitié des valeurs de la série (écart interquartile). Il a l'avantage, de par sa construction, d'être « robuste » en particulier par rapport aux valeurs extrêmes. Il permet de comparer de façon pertinente des distributions statistiques telles que salaires, durées de vie…

Le résumé (minimum, 1er quartile, médiane, 3^{e} quartile, maximum) permet de construire des diagrammes en boîtes et donc de mieux visualiser et comparer entre elles plusieurs séries
Mais l'inconvénient de ces paramètres est qu'ils ne se prêtent pas aux calculs (par ex. on ne peut pas calculer les indices du regroupement de 2 séries à partir des indices des 2 séries initiales).

III. Distributions des données

1. Distributions à partir du couple (moyenne, écart type)

Le couple (moyenne; écart-type) ne donne pas de renseignements "précis" sur la distribution des données autour de la moyenne.

Cependant, on peut affirmer qu'il y a au moins 75% des données dans l'intervalle $\left[\bar{x}-2s;\bar{x}+2s\right]$

En effet : s et $\bar{x}$ étant calculés ,

$\left|x_i-\bar{x}\right|>2s \Rightarrow \ x_i-\bar{x}\ ^2>4s^2$, et supposons qu'il y a p données vérifiant cette inégalité, alors les $n-p$ autres données vérifient $\left|x_i-\bar{x}\right|\leqslant 2s$ *donc* $x_i\in\left[\bar{x}-2s;\bar{x}+2s\right]$

et comme

$$ns^2=\sum_{p\ données}(x_i-\bar{x})^2+\sum_{n-p\ données}(x_i-\bar{x})^2$$

$$ns^2>4ps^2+\sum_{n-p\ données}(x_i-\bar{x})^2$$

$$\sum_{n-p\ données}(x_i-\bar{x})^2<ns^2-4ps^2$$

$$\sum_{n-p\ données}(x_i-\bar{x})^2<s^2(n-4p) \quad \text{(majoration assez grossière)}$$

donc $(n-4p)>0$ et $p<\frac{1}{4}n$

d'où $n-p\geqslant\frac{3}{4}n$ soit au moins 75% des données.

De la même façon on peut voir qu'il y a au moins $\frac{8}{9}$ des données dans l'intervalle $\left[\bar{x}-3s;\bar{x}+3s\right]$, soit 89% de l'effectif. ($n-9p>0$ donc $p<\frac{1}{9}n$).

Mais il se peut qu'il y ait une seule donnée dans l' intervalle $\left[\bar{x}-s;\bar{x}+s\right]$,

$n-p>0$ donc $p\leqslant n-1$. (construire un exemple).

Ce sont des pourcentages minimum correspondant à des types de séries particuliers, « les plus défavorables » mais pour les séries du type évoqué ci-avant, ces pourcentages sont plus importants .

Rappelons que pour une distribution normale, on a :

l'intervalle $\left[\overline{x}-s;\overline{x}+s\right]$ contient 68% de la population

l'intervalle $\left[\overline{x}-2s;\overline{x}+2s\right]$ contient 95% de la population

l'intervalle $\left[\overline{x}-3s;\overline{x}+3s\right]$ contient 99,7% de la population

Une utilisation pratique de l'intervalle de dispersion à 95% est que s sera de l'ordre du quart de l'étendue et en général inferieur.

2. Distributions à partir du couple (milieu des extrêmes, demi-étendue)

Aucun renseignement !

3. Distributions à partir du le couple (médiane, écart moyen à la médiane)

Ce couple ne donne pas non plus de renseignements "précis" sur la distribution des données autour de la médiane.

Cependant, on peut affirmer qu'il y a au moins 50% des données dans l'intervalle $\left[m_e-2e_m;m_e+2e_m\right]$

En effet : m_e et e_m étant calculés ,

supposons qu'il y a p données vérifiant $\left|x_i-m_e\right|>2e_m$, alors :

les $n-p$ autres données vérifient $\left|x_i-m_e\right|\leqslant 2e_m$ donc $x_i\in\left[m_e-2e_m;m_e+2e_m\right]$

et comme

$$e_m=\frac{1}{n}\sum_{k=1}^{n}\left|x_i-m_e\right|=\frac{1}{n}\sum_{p\ données}\left|x_i-m_e\right|+\frac{1}{n}\sum_{n-p\ données}\left|x_i-m_e\right|$$

$$ne_m=\sum_{p\ données}\left|x_i-m_e\right|+\sum_{n-p\ données}\left|x_i-m_e\right|$$

$$ne_m>2pe_m+\sum_{n-p\ données}\left|x_i-m_e\right|$$

$$ne_m-2pe_m>\sum_{n-p\ données}\left|x_i-m_e\right| \qquad \text{et} \qquad e_m(n-2p)>0$$

donc $n-2p>0$ et $p<\frac{n}{2}$

et $n-p \geqslant \frac{1}{2}n$ soit au moins 50% des données

De la même façon on peut voir qu'il y a au moins $\frac{2}{3}$ des données dans l'intervalle $\left[m_e-3e_m;m_e+3e_m\right]$, soit 67% de l'effectif.

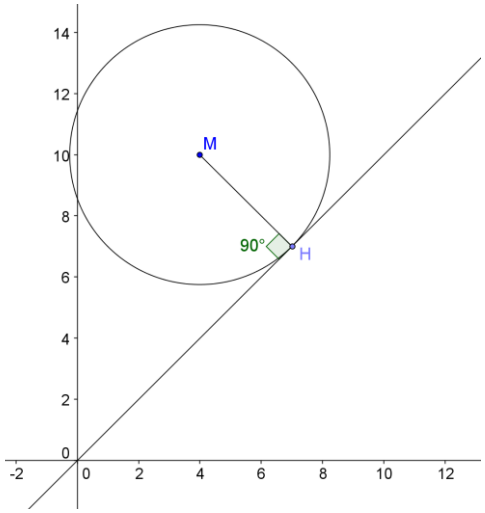
Mais il se peut qu'il y ait une seule donnée dans l' intervalle $\left[m_e-e_m;m_e+e_m\right]$,

$n-p>0$ donc $p \leqslant n-1$. (exemple : (1;1;1;3;5;5;5))

Terminons par un peu de géométrie récréative :

La distance du point *M* à la droite *x* = *y* (= *z*) est aussi le rayon de "la plus petite" boule de centre *M* qui intercepte la droite

1. **En dimension 2 : pour *M(4;10)***



avec la distance d_2

On voit que le rayon de la boule est $3\sqrt{2}$, (Pythagore)

la distance de ***M*** à la droite, est $3\sqrt{2}$

la variance est $\frac{1}{2}\left(3\sqrt{2}\right)^2 = 9$ et l'écart-type est 3

$H\left(7;7\right)$, la moyenne est $\overline{x} = 7$

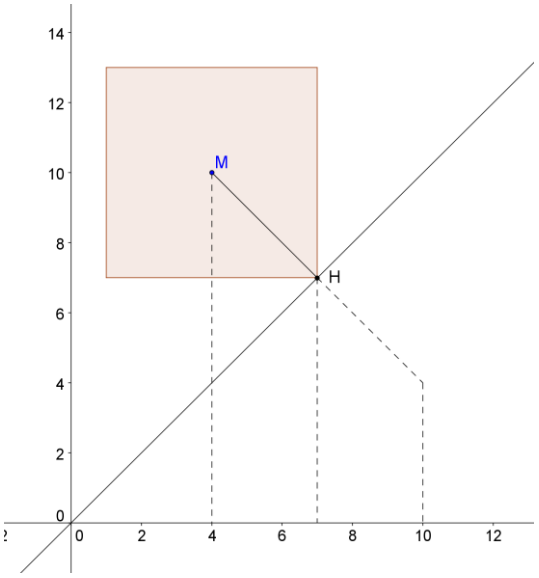
avec la distance d_∞ :

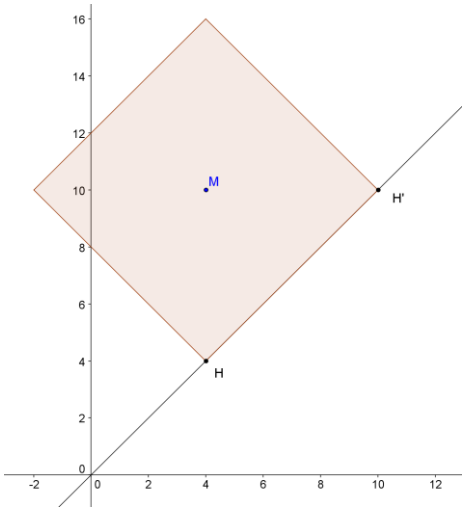
On voit que le rayon de la boule est 3

la distance de ***M*** à la droite est 3

la demi-étendue est $3/2$

$H\left(7;7\right)$, la demi-somme des extrêmes est 7.





avec la distance d_1 :

On voit que le rayon de la boule est 6

la distance de ***M*** à la droite est 6

l' écart moyen à la médiane est 6/2 = 3

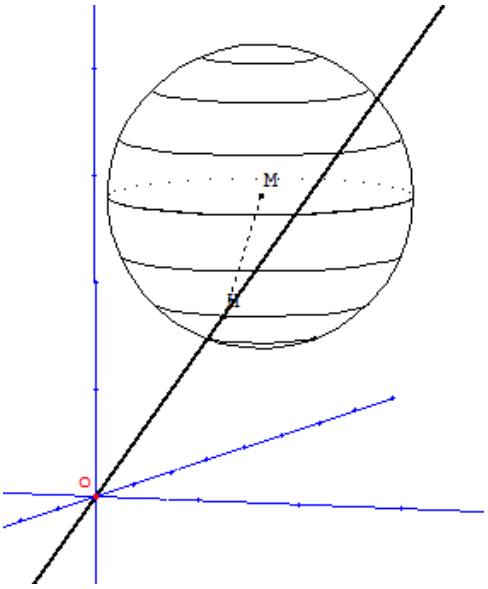
l'intersection de la droite et de la boule est le segment

[*H* , *H'*]

$m_e = 7$ ou toute valeur comprise entre 4 et 10,

(2 est pair).

2. En dimension 3 : pour *M(1 ; 2 ; 3)*



avec la distance d_2

On peut calculer le rayon de la boule : $\sqrt{2}$

la distance de ***M*** à la droite est $\sqrt{2}$

la variance est $\frac{1}{3}\left(\sqrt{2}\right)^2 = \frac{2}{3}$

et l'écart-type est $\sqrt{\frac{2}{3}}$

$H(2;2;2)$, la moyenne est $\bar{x} = 2$

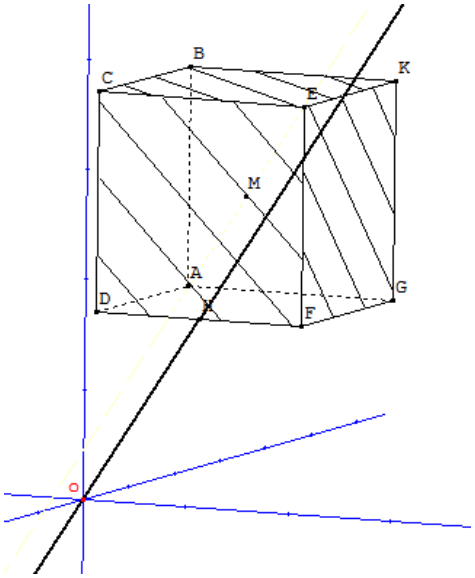
avec la distance d_∞ : ***(H est sur [D,F])***

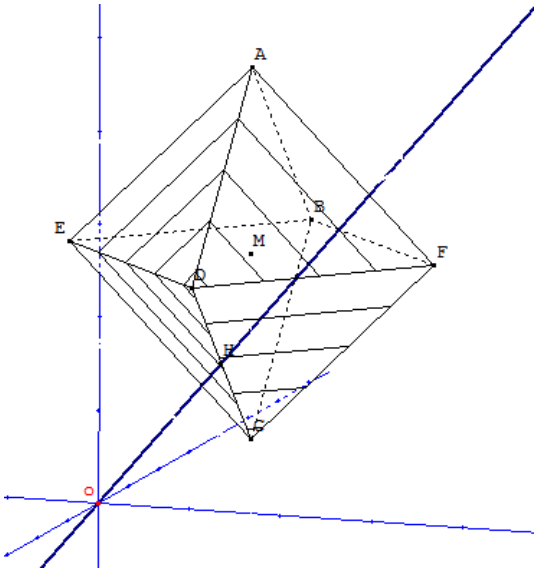
On peut calculer le rayon de la boule : 1

la distance de ***M*** à la droite est 1

la demi-étendue est $1/3$

$H(2;2;2)$, la demi-somme des extrêmes est 2





avec la distance d_1 : ***(H est sur[D,C])***

On peut calculer le rayon de la boule : 2

la distance de ***M*** à la droite est 2

l' écart moyen à la médiane est 2/3

$H(2;2;2)$, la médiane est $m_e = 7$

Partie B

PROBABILITES

LOI BINOMIALE

Ce document est une introduction à la loi binomiale et au calcul des coefficients binomiaux.

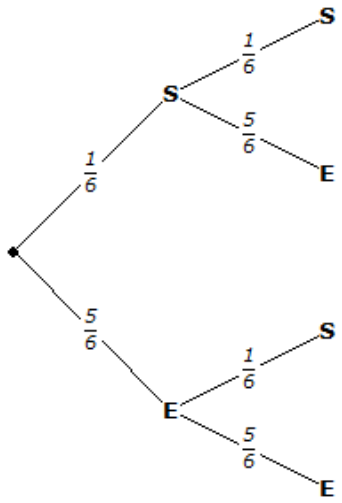
1) **On lance un dé cubique équilibré. On gagne si on obtient 4, et on perd sinon.**
   Soit S l'événement : « on obtient 4 » et E l'événement contraire.

   On définit ainsi une épreuve de Bernoulli. Les deux issues sont S (succès) et E (échec) avec $\boldsymbol{p(S) = \frac{1}{6}}$ $\quad$ $\boldsymbol{et\ p(E) = \frac{5}{6}}$

2) **On lance 2 fois un dé cubique équilibré. On appelle X la variable aléatoire donnant le nombre de succès.**

$Proba\ succès\ p = \frac{1}{6}$ $\quad$ $proba\ échec : q = \frac{5}{6}$ $\quad$ $nombre\ d'expériences\ n = 2$

Arbre pondéré :



Résultats possibles

$\boldsymbol{SS\ (X = 2)}$; $\boldsymbol{SE\ et\ ES\ (X = 1)}$; $\boldsymbol{EE\ (X = 0)}$

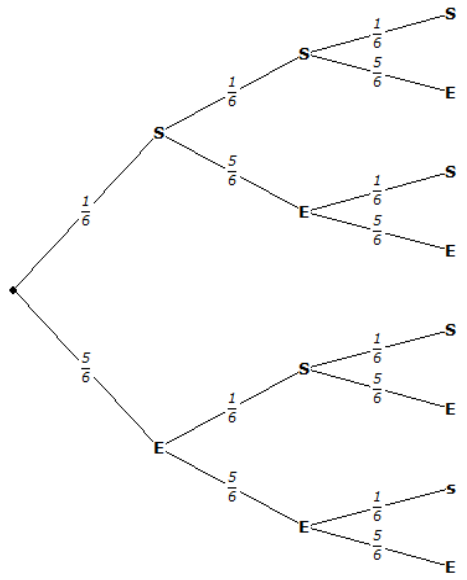
La loi de probabilité de X est :

k	0	1	2	
$P(X = k)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$	1

	Remarques
Espérance $E(X) = 0 \times \frac{25}{36} + 1 \times \frac{10}{36} + 2 \times \frac{1}{36} = \frac{12}{36} = \frac{1}{3}$	$np = 2 \times \frac{1}{6} = \frac{1}{3}$
Variance $\quad V(X) = E(X^2) - [E(X)]^2$ $V(X) = 0^2 \times \frac{25}{36} + 1^2 \times \frac{10}{36} + 2^2 \times \frac{1}{36} - \left(\frac{1}{3}\right)^2$ $V(X) = \frac{14}{36} - \frac{1}{9} = \frac{14}{36} - \frac{4}{36} = \frac{10}{36}$	$npq = 2 \times \frac{1}{6} \times \frac{5}{6} = \frac{10}{36}$
Ecart type $\quad \sigma_{X=} \sqrt{V(X)} \approx 0{,}527$	

3) On lance 3 fois un dé cubique équilibré. On appelle *X* la variable aléatoire donnant le nombre de succès.

$$Proba\ succès\ p = \frac{1}{6} \quad proba\ échec : q = \frac{5}{6} \quad nombre\ d'expériences\ n = 3$$

<table>
<tr>
<td>On complète l'arbre précédent
 
</td>
<td>Résultats possibles
$\boldsymbol{SSS}\ (\boldsymbol{X=3})$
$\boldsymbol{SSE}\ \ \boldsymbol{SES}\ \ \boldsymbol{ESS}\ \ (\boldsymbol{X=2})$
$\boldsymbol{SEE}\ \ \boldsymbol{ESE}\ \ \boldsymbol{EES}\ (\boldsymbol{X=1})$
$\boldsymbol{EEE}\ \ (\boldsymbol{X=0})$
La loi de probabilité de X est :
<table><tr><td>k</td><td>0</td><td>1</td><td>2</td><td>3</td><td></td></tr><tr><td>$P(X=k)$</td><td>$\frac{125}{216}$</td><td>$\frac{75}{216}$</td><td>$\frac{15}{216}$</td><td>$\frac{1}{216}$</td><td>1</td></tr></table></td>
</tr>
<tr>
<td>$P(X=0) = 1 \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$</td>
<td>1 est le nombre de chemins menant à 0 succès</td>
</tr>
<tr>
<td>$P(X=1) = 3 \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6}$</td>
<td>3 est le nombre de chemins menant à 1 succès. On obtient 1 succès en complétant SE et ES par E et en complétant EE par S. D'où 3 = 2+1</td>
</tr>
<tr>
<td>$P(X=2) = 3 \times \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6}$</td>
<td>3 est le nombre de chemins menant à 2 succès. On obtient 2 succès en complétant SE et ES par S et en complétant SS par E. D'où 3 = 2+1</td>
</tr>
<tr>
<td>$P(X=3) = 1 \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}$</td>
<td>1 est le nombre de chemins menant à 3 succès</td>
</tr>
</table>

Espérance	Remarque
$E(X) = 0 \times \frac{125}{216} + 1 \times \frac{75}{216} + 2 \times \frac{15}{216} + 3 \times \frac{1}{216} = \frac{108}{216} = 0{,}5$	$np = 3 \times \frac{1}{6} = 0{,}5$
Variance $V(X) = E(X^2) - [E(X)]^2$	
$V(X) = 0^2 \times \frac{125}{216} + 1^2 \times \frac{75}{216} + 2^2 \times \frac{15}{216} + 3^2 \times \frac{1}{216} - 0{,}5^2$	
$V(X) = \frac{144}{216} - \frac{1}{4} = \frac{144}{216} - \frac{54}{216} = \frac{90}{216} = \frac{5}{12}$	$npq = 3 \times \frac{1}{6} \times \frac{5}{6} = \frac{15}{36} = \frac{5}{12}$
$\sigma_{X=} \sqrt{V(X)} \approx 0{,}645$	

4) On lance 4 fois un dé cubique équilibré. Définir la variable aléatoire donnant le nombre de succès.

$Proba\ succès\ p = \frac{1}{6}$ $proba\ échec : q = \frac{5}{6}$ $nombre\ d'expériences\ n = 4$

$P(X=0) = \mathbf{1} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$	**1** est le nombre de chemins menant à 0 succès
$P(X=1) = \mathbf{4} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$	**4** est le nombre de chemins menant à 1 succès. On les obtient en complétant le résultat EEE avec S et aussi en complétant les résultats EES ESE SEE avec E ; d'où $4 = 1 + 3$
$P(X=2) = \mathbf{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6}$	**6** est le nombre de chemins menant à 2 succès. On les obtient en complétant EES ESE SEE par S et en complétant ESS SES SSE par E ; d'où $6 = 3 + 3$
$P(X=3) = \mathbf{4} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6}$	**4** est le nombre de chemins menant à 3 succès. On les obtient en complétant le résultat SSS avec E et aussi en complétant les résultats ESS SES SSE à 1 succès avec S ; d'où $4 = 1 + 3$
$P(X=4) = \mathbf{1} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}$	**1** est le nombre de chemins menant à 4 succès

D'où la loi de probabilité de X :

k	0	1	2	3	4	
$P(X=k)$	$\left(\frac{5}{6}\right)^4$	$4 \times \left(\frac{5}{6}\right)^3 \times \frac{1}{6}$	$6 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$	$4 \times \left(\frac{1}{6}\right)^3 \times \frac{5}{6}$	$\left(\frac{1}{6}\right)^4$	
$P(X=k)$	$\frac{625}{1296}$	$\frac{500}{1296}$	$\frac{150}{1296}$	$\frac{20}{1296}$	$\frac{1}{1296}$	1

	Remarques
Espérance $E(X) = 0 \times \frac{625}{1296} + 1 \times \frac{500}{1296} + 2 \times \frac{150}{1296} + 3 \times \frac{20}{1296} + 4 \times \frac{1}{1296}$ $= \frac{864}{1296}$ $E(X) = \frac{2}{3}$	$np = 4 \times \frac{1}{6} = \frac{2}{3}$
Variance $V(X) = E(X^2) - [E(X)]^2$ $V(X) = 0^2 \times \frac{625}{1296} + 1^2 \times \frac{500}{1296} + 2^2 \times \frac{150}{1296} + 3^2 \times \frac{20}{1296} + 4^2 \times \frac{1}{1296} - \left(\frac{2}{3}\right)^2$ $V(X) = 1 - \frac{4}{9} = \frac{5}{9}$	$npq = 4 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{9}$
Ecart type $\sigma_{X=} \sqrt{V(X)} \approx 0{,}746$	

5) On lance 5 fois un dé cubique équilibré. On appelle *X* la variable aléatoire donnant le nombre de succès.

$Proba\ succès\ p = \frac{1}{6}$ $proba\ échec : q = \frac{5}{6}$ $nombre\ d'expériences\ n = 5$

Résultats possibles :

k	0	1	2	3	4	5	
$P(X = k)$	$\mathbf{1} \times \left(\frac{5}{6}\right)^5$	$\mathbf{5} \times \frac{1}{6} \times \left(\frac{5}{6}\right)^4$	$10 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^3$	$10 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2$	$5 \times \left(\frac{1}{6}\right)^4 \times \frac{5}{6}$	$\left(\frac{1}{6}\right)^5$	
	$\frac{3125}{7776}$	$\frac{3125}{7776}$	$\frac{1250}{7776}$	$\frac{250}{7776}$	$\frac{25}{7776}$	$\frac{1}{7776}$	1

Le nombre de chemins menant à k succès est obtenu comme précédemment :

1 *chemin pour* $k = 0$

5 *chemins pour* $k = 1$ $(5 = 1 + 4)$

10 *chemins pour* $k = 2$ $(10 = 4 + 6)$

10 *chemins pour* $k = 3$ $(10 = 6 + 4)$

5 *chemins pour* $k = 4$ $(5 = 4 + 1)$

1 *chemin pour* $k = 5$

	Remarques
Espérance $E(X) = 0 \times \frac{3125}{7776} + 1 \times \frac{3125}{7776} + 2 \times \frac{1250}{7776} + 3 \times \frac{250}{7776} + 4 \times \frac{25}{7776} + 5 \times \frac{1}{7776}$ $E(X) = \frac{6480}{7776} = \frac{5}{6}$	$np = 5 \times \frac{1}{6} = \frac{5}{6}$
Variance $V(X) = E(X^2) - [E(X)]^2$ $V(X) = 0 \times \frac{3125}{7776} + 1 \times \frac{3125}{7776} + 2 \times \frac{1250}{7776} + 3 \times \frac{250}{7776} + 4 \times \frac{25}{7776} + 5 \times \frac{1}{7776} - (\frac{5}{6})^2$ $V(X) = \frac{10800}{7776} - \frac{25}{36} = \frac{25}{18} - \frac{25}{36} = \frac{25}{36}$	$npq = 5 \times \frac{1}{6} \times \frac{5}{6} = \frac{25}{36}$
$\sigma_{X=} \sqrt{V(X)} \approx 0{,}83$	

6) On constate sur chaque exemple : pour la loi $\mathcal{B}(n;p)$:

$$E(X) = np \quad V(X) = npq$$

a) **Comment calculer les coefficients binomiaux ?**

On note $\binom{n}{k}$ le nombre de chemins conduisant à k succès sur les n tirages

On constate :

$$\binom{n}{k} = \binom{n}{n-k}$$

Le nombre de chemins conduisant à k succès est le même que le nombre de chemins conduisant à $n-k$ échecs

On a trouvé sur les exemples

$n=2$	$\binom{2}{0}=1$	$\binom{2}{1}=2$	$\binom{2}{2}=1$			
$n=3$	$\binom{3}{0}=1$	$\binom{3}{1}=3$	$\binom{3}{2}=3$	$\binom{3}{3}=1$		
$n=4$	$\binom{4}{0}=1$	$\binom{4}{1}=4$	$\binom{4}{2}=6$	$\binom{4}{3}=4$	$\binom{4}{4}=1$	
$n=5$	$\binom{5}{0}=1$	$\binom{5}{1}=5$	$\binom{5}{2}=10$	$\binom{5}{3}=10$	$\binom{5}{4}=5$	$\binom{5}{5}=1$

C'est le triangle de Pascal (à compléter)

$n=1$	1											
$n=2$	1	2	1									
$n=3$	1	3	3	1								
$n=4$	1	4	6	4	1							
$n=5$	1	5	10	10	5	1						
$n=6$												
$n=7$												
$n=8$												
$n=9$												
$n=10$												

En général

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

On obtient k succès avec n tirages :
en prenant les $k-1$ succès de $n-1$ tirages , on ajoute S ;
en prenant les k succès de $n-1$ tirages, on ajoute E.

en particulier $\binom{n}{n} = 1 = \binom{n}{0}$ $\binom{n}{1} = \binom{n}{n-1} = n$

On admet :

Soit une expérience à deux issues , appelées Succès et Echec. La probabilité du succès est p , et la probabilité de l'échec est $q = 1 - p$

Cette expérience est une épreuve de Bernoulli

On répète n fois cette épreuve, et toutes les épreuves sont indépendantes.

La variable aléatoire qui prend pour valeurs le nombre de succès obtenus suit la loi Binomiale de paramètres n et p , notée $\mathcal{B}(n; p)$

Pour tout entier k entre 0 et n $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

L'espérance mathématique est $E(X) = np$

La variance est $V(X) = npq = np(1-p)$ et l'écart-type $\sigma = \sqrt{npq}$

Pour calculer les coefficients binomiaux :

On utilise le triangle de Pascal pour les petites valeurs.

On utilise la calculatrice pour les autres valeurs.

Remarque : Le programme ne donne pas la formule des coefficients binomiaux avec n!.

On peut définir n! dans le chapitre des suites numériques, et donner la formule avec les factorielles à cette occasion.

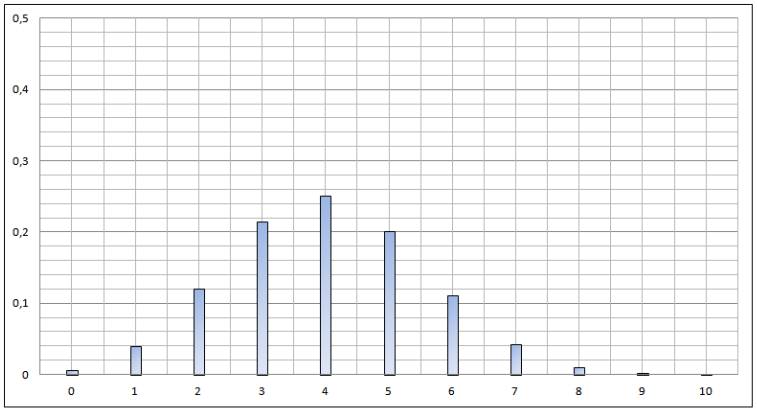
Introduction de la loi normale centrée réduite à partir de la loi binomiale

1. Mise en évidence de la courbe de la fonction définie sur $\mathbb{R}$ par $x \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

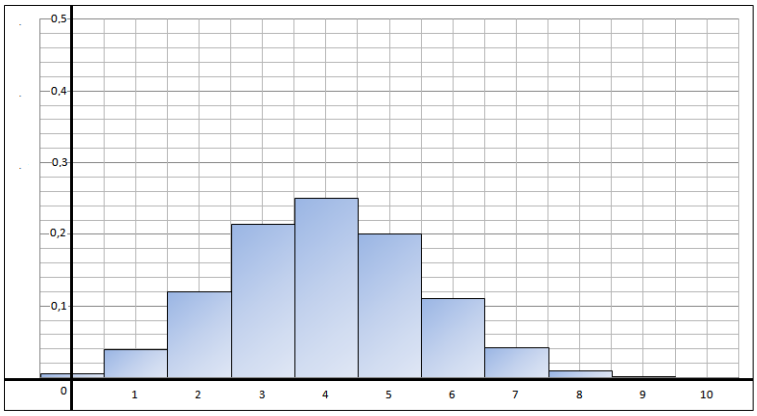
On considère une variable aléatoire X suivant une loi binomiale $B\ n\,;p$

$P\ X=k\ =\binom{n}{k}p^k q^{n-k}$ $\quad k\in\ 0,1,2,...,n$ $\quad E\ X\ =np$ et $\sigma\ X\ =\sqrt{npq}$

pour les figures on a choisi : $B\ 10\,;0{,}4$

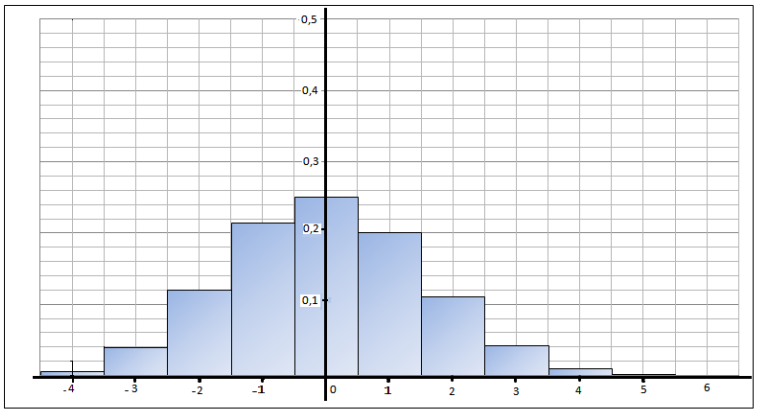


Puiqu'il s'agit d'une variable discrète, la représentation graphique est un diagramme en batons
les valeurs prises par la variable aléatoire X sont en nombre fini.
$P\ X=k$ est la hauteur des bâtons



Une autre représentation graphique est l'histogramme ci contre :

$P\ X=k$ est l'aire du rectangle dont l'un des côté est le segment $k-0.5;k+0.5$ de dimension 1, l'autre un segment de dimension $P\ X=k$



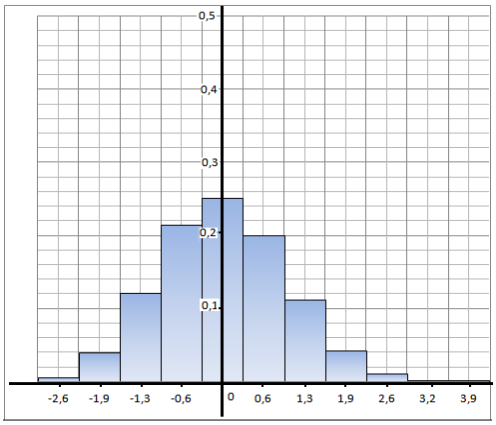
On considère maintenant la variable

$Y=X-np$ $\quad P\ Y=k'\ =P\ X=k'+np$

$k'\in\ -np,1-np,2-np,...,nq$

$E\ Y\ =0$ et $\sigma\ Y\ =\sqrt{npq}$

l'espérance ne dépend plus de p
la variable est ainsi centrée et la représentation est alors translatée



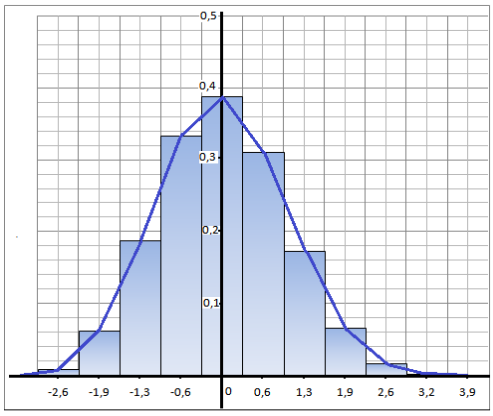
On considère maintenant la variable $Z = \dfrac{Y}{\sqrt{npq}} = \dfrac{X - np}{\sqrt{npq}}$

$$P(Z = a) = P(X = a\sqrt{npq} + np)$$

$$a \in \left\{ \frac{-np}{\sqrt{npq}}, \frac{1-np}{\sqrt{npq}}, \frac{2-np}{\sqrt{npq}}, \ldots, \frac{nq}{\sqrt{npq}} \right\}$$

$$E(Z) = 0 \quad \text{et} \quad \sigma(Z) = 1$$

ainsi l'écart-type, lui aussi, ne dépend plus de p, mais l'aire des rectangles est divisée par $\sqrt{npq}$, puisque l'une des dimension l'est par $\sqrt{npq}$

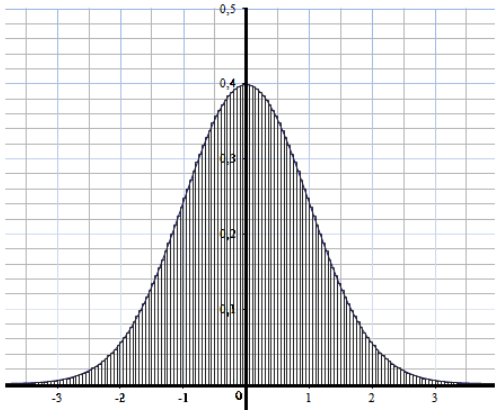


On multiplie donc l'autre dimension par $\sqrt{npq}$
et on construit la courbe passant par les milieux des côtés supérieurs des rectangles de coordonnées

$$\left(\frac{k-np}{\sqrt{npq}} \; ; \; \sqrt{npq}\, P(X = k) \right) \quad \text{soit} \quad \left(\frac{k-np}{\sqrt{npq}} \; ; \; \sqrt{npq} \binom{n}{k} p^k q^{n-k} \right)$$

et cette courbe quand n devient grand s'approche de celle de la fonction définie sur $\mathbb{R}$ par

$$x \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



pour n = 200, on obtient le dessin ci-contre, et
l'aire sous la courbe est proche de la somme des aires des rectangles, c'est à dire 1.

remarque 1

avec la formule de Stirling, on peut montrer (péniblement) que

$$\sqrt{npq}\, \frac{n!}{k_n!(n-k_n)!}\, p^{k_n} q^{n-k_n} \simeq \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad si \quad k_n = x\sqrt{npq} + np \quad \text{pour } n \text{ assez grand.}$$

Cette **fonction** est définie **continue, positive sur $\mathbb{R}$ et sommable sur $\mathbb{R}$ de somme 1**,

c'est à dire $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = 1$

on dit alors qu 'elle définit une **densité de probabilité** sur $\mathbb{R}$.

remarque 2

Pour calculer $\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx$, le plus simple est de considérer $\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy$

on utilise deux recouvrements de $\mathbb{R}^2$: d'une part

$$\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{D_R} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{D_R} e^{-r^2} \left|\det J(r,\theta)\right| drd\theta = \lim_{R\to+\infty} \iint_{D_R} e^{-r^2} rdrd\theta$$

$$= \lim_{R\to+\infty} \left(\int_0^R re^{-r^2} dr \right) \left(\int_0^{2\pi} d\theta \right)$$

$$= \lim_{R\to+\infty} \left[-\frac{1}{2} e^{-r^2} \right]_0^R \times \theta_0^{2\pi} = \lim_{R\to+\infty} \frac{1}{2} \left(1 - e^{-R^2}\right) \times 2\pi = \pi$$

$$car \quad \left|J(r,\theta)\right| = \left|\frac{D(x,y)}{D(r,\theta)}\right| = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

d'autre part avec le théorème de Fubini

$$\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{-R;R\times-R;R} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right) \left(\int_{-R}^R e^{-y^2} dy \right) = \lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right)^2$$

d'où $\lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right)^2 = \pi$ et $\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$ (l'intégrale de Gauss)

d'où enfin

$$\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx = \lim_{R\to+\infty} \int_{-R}^R e^{-\frac{x^2}{2}} dx = \lim_{R\to+\infty} \int_{-R/\sqrt{2}}^{R/\sqrt{2}} e^{-y^2} \sqrt{2} dy = \lim_{R\to+\infty} \sqrt{2} \int_{-R/\sqrt{2}}^{R/\sqrt{2}} e^{-y^2} dy = \sqrt{2\pi}$$

ainsi $\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx = 1$

2. La loi normale centrée réduite

la fonction définie sur $\mathbb{R}$ par $x \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ est la densité d'une loi continue :

la loi normale centrée réduite $N(0;1)$

Si une variable aléatoire continue X suit la loi $N(0;1)$

$P(a \leqslant X \leqslant b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}} dt$.

Sa fonction de répartition est définie sur $\mathbb{R}$ par**:**

$P(X \leqslant x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ **.** (aussi notée $\Pi(x)$)

et $P(a \leqslant X \leqslant b) = \Pi(b) - \Pi(a)$

on a : $E(X) = 0$ et $Var(X) = 1$ donc $\sigma(X) = 1$

3. Autres lois continues on généralise les notions rencontrées ci-dessus:

Définition 1
une **densité de probabilité** est une fonction f définie **continue, positive** sur I et **intégrable** sur I
de somme 1 : $\int_I f(x)dx = 1$

Définition 2
la loi de probabilité de densité f est l'application P qui, à tout $[a,b] \subset I$ associe le nombre $\int_a^b f(x)dx$

Définition 3
Si une variable aléatoire continue X suit la loi de densité f sur I , $\forall [a,b] \subset I$
la loi de probabilité de X est définie par $P(a \leqslant X \leqslant b) = \int_a^b f(x)dx$.

Les trois autres lois continues du programme :

la loi normale générale, la loi uniforme et la loi exponentielle

a. la loi normale de paramètres μ et σ : $N(\mu,\sigma^2)$

la fonction définie sur $\mathbb{R}$ par $x \longrightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ est la densité de la loi normale $N(\mu,\sigma^2)$

Si une variable aléatoire continue X suit la loi $N(\mu,\sigma^2)$

la loi de probabilité de X (P ou P_X) est définie par

$$P(a \leqslant X \leqslant b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

on a : $E(X) = \mu$ et $Var(X) = \sigma^2$ donc $\sigma(X) = \sigma$

b. la loi uniforme sur $[a,b]$ **,** notée $U\ \ a,b$ **,**

la fonction définie sur $[a,b]$ par $x \longrightarrow \frac{1}{b-a}$ est la densité de la loi $U\ \ a,b$

Si une variable aléatoire continue X suit la loi $U\ \ a,b$

la loi de probabilité de X (P ou P_X) est définie par

$$P(\alpha \leqslant X \leqslant \beta) = \int_\alpha^\beta \frac{1}{b-a} dx = \frac{\beta-\alpha}{b-a} \qquad \text{si} \quad a \leqslant \alpha \leqslant \beta \leqslant b$$

on a : $E(X) = \frac{a+b}{2}$ et $Var(X) = \frac{(b-a)^2}{12}$ donc $\sigma(X) = \frac{b-a}{2\sqrt{3}}$

c. la loi exponentielle de paramètre $\lambda \in \mathbb{R}_+^*$ **,** notée $\mathcal{E}(\lambda)$

la fonction définie sur $\mathbb{R}_+$ par $x \longrightarrow \lambda e^{-\lambda x}$ est la densité de la loi exponentielle $\mathcal{E}(\lambda)$

Si une variable aléatoire continue X suit la loi $\mathcal{E}(\lambda)$

la loi de probabilité de X (P ou P_X) est définie par

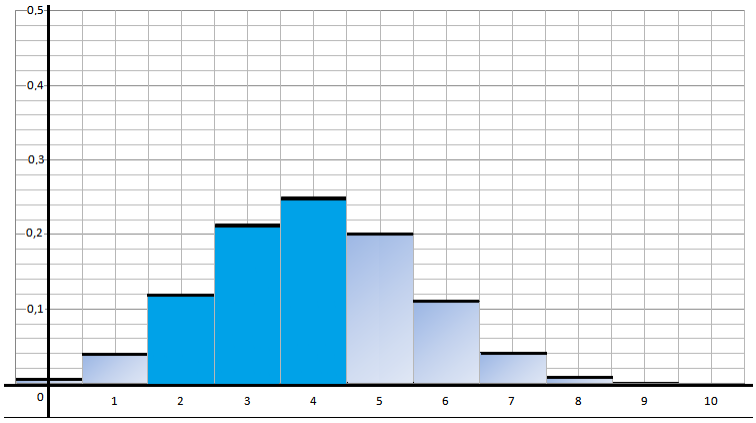
$$P(a \leqslant X \leqslant b) = \int_a^b \lambda e^{-\lambda x} dx = e^{-\lambda a} - e^{-\lambda b} \qquad \text{si} \quad x \geqslant 0$$

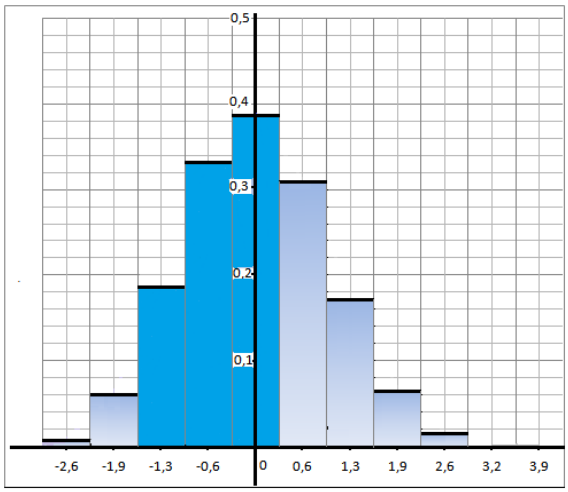
on a : $E(X)=\frac{1}{\lambda}$ et $Var(X)=\frac{1}{\lambda^2}$ donc $\sigma(X)=\frac{1}{\lambda}$

4. Approximation de la loi binomiale

si on regarde maintenant les fonctions en escalier "définie par les rectangles"

on a $P(X=k)=\int_{k-0,5}^{k+0,5} g_n(x)dx$





si on note g_n la fonction en escalier ci-dessus et f_n la fonction en escalier ci-dessus

on obtient :

$$P(k_1 \leqslant X \leqslant k_2)=\int_{k_1-0,5}^{k_2+0,5} g_n(x)dx$$

$$P(k_1 \leqslant X \leqslant k_2)=P\left(\frac{k_1-0,5-np}{\sqrt{npq}} \leqslant Z \leqslant \frac{k_2+0,5-np}{\sqrt{npq}}\right)$$

$$\int_{k_1-0,5}^{k_2+0,5} g_n(x)dx=\int_{\frac{k_1+0,5-np}{\sqrt{npq}}}^{\frac{k_2+0,5-np}{\sqrt{npq}}} \sqrt{npq}\, g_n\left(u\sqrt{npq}+np\right)du=\int_{\frac{k_1+0,5-np}{\sqrt{npq}}}^{\frac{k_2+0,5-np}{\sqrt{npq}}} f_n(u)du$$

(changement de variable licite puisque g_n est continue par morceaux
et le changement de variable affine)

et en passant à la limite

$$P(k_1 \leqslant X \leqslant k_2) \cong \int_{\frac{k_1+0,5-np}{\sqrt{npq}}}^{\frac{k_2+0,5-np}{\sqrt{npq}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}du \cong \int_{\frac{k_1-np}{\sqrt{npq}}}^{\frac{k_2-np}{\sqrt{npq}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}du$$

c'est le théorème de **Moivre-Laplace.**

remarque

la fonction définie sur $\mathbb{R}$ par $x \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ est la limite uniforme de la suite de fonction en escaliers f_n .

Le théorème de Moivre Laplace

Le théorème de Moivre Laplace

Si $(X_n)_n$ est une suite de variables aléatoires indépendantes suivant la même loi de Bernoulli de paramètre p , donc d'espérance p et de variance $\sigma^2 = pq$.

Alors si $F_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{Y_n}{n}$

$Z_n = \frac{F_n - E(F_n)}{\sigma(F_n)} = \frac{F_n - p}{\sqrt{pq/n}}$ ***converge en loi*** vers X où X suit la loi normale $N(0\ ;1)$

c'est-à-dire que pour tout $a < b$: $\lim_{n \to +\infty} P(a < Z_n < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

en particulier

$Y_n = X_1 + X_2 + ... + X_n$ ***suit approximativement*** la loi normale $N\left(np\ ;\sqrt{npq}\right)$

$F_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{Y_n}{n}$ ***suit approximativement*** la loi normale $N\left(p; \sqrt{pq/n}\right)$

Demonstration

soit $(a,b) \in \mathbb{R}^2$ et $a < b$.

$Z_n = \frac{F_n - E(F_n)}{\sigma(F)} = \frac{F_n - p}{\sqrt{pq/n}} = \frac{Y_n - np}{\sqrt{npq}}$ et on sait que Y suit une la loi binomiale $B\ n, p$

on a $P\ a < Z_n < b\ = P\ a\sqrt{npq} + np < Y_n < b\sqrt{npq} + np$

si on note $k_{n,1}, k_{n,2}, ...k_{n,j_n}$ les entiers appartenant à l'intervalle $\left] a\sqrt{npq} + np\ ;\ b\sqrt{npq} + np \right[$

$$P\ a < Z_n < b\ = \sum_{i=1}^{j_n} P\ Y_n = k_{n,i}$$

première étape

tout d'abord on regarde le comportement asymptotique de $P\ Y_n = k_{n,i}$ avec la formule de Stirling, laquelle fera apparaitre la fonction exponentielle.

On montre, en notant provisoirement k_m les entiers tels que $np+a\sqrt{npq}<k_m<np+b\sqrt{npq}$ **, que**

$$\forall k_m,\ \exists\,\varepsilon(n,k_m)\ \text{tel que}\quad P\ Y_n=k_m\ =\frac{1}{\sqrt{2\pi}\sqrt{npq}}\exp\left(-\frac{1}{2}\left(\frac{k_m-np}{\sqrt{npq}}\right)^2\right)1+\varepsilon(n,k_m)$$

$$\textbf{et}\ \ \exists\,K\in\mathbb{R}_+^*\ \text{tel que}\quad \left|\varepsilon(n,k_m)\right|\leqslant\frac{K}{\sqrt{n}}\qquad \forall k_m\in\left]\,a\sqrt{npq}+np\ ;\ b\sqrt{npq}+np\,\right[$$

preuve

Elle repose sur la formule de Stirling ($n!=n^n\,e^{-n}\,\sqrt{2\pi n}\,\left(1+\varepsilon(n)\right)$ où $\varepsilon(n)\sim\frac{1}{n}$)

en écriture logarithmique $\quad ln\left(n!\right)=\frac{1}{2}ln\left(2\pi\right)+\left(n+\frac{1}{2}\right)ln(n)-n+\varepsilon(n)$ où $\varepsilon(n)\sim\frac{1}{n}$

Pour tout k_m on a $\quad P\ Y=k_m\ =\frac{n!}{k_m!(n-k_m)!}\,p^{k_m}q^{n-k_m}$

comme $\quad k_m>np+a\sqrt{npq}>n\left(q+a\sqrt{\frac{pq}{n}}\right)$ et

$$n-k_m>n-np-b\sqrt{npq}=nq-b\sqrt{npq}=n\left(q-b\sqrt{\frac{pq}{n}}\right)$$

$$\lim_{n\to\infty}k_m=+\infty\qquad\text{et}\qquad\lim_{n\to\infty}(n-k_m)=+\infty$$

et on peut utiliser la formule de Stirling puis regrouper "astucieusement" les termes:

$$\ln\ P\ Y=k_m\ =\ln\ n!\ -\ln\ k_m!\ -\ln\ (n-k_m)!\ +k_m\ln\ p\ +(n-k_m)\ln\ q$$

$$\ln\ P\ Y=k_m\ =\frac{1}{2}\ln\ 2\pi\ +\left(n+\frac{1}{2}\right)\ln(n)-n+\varepsilon(n)-\frac{1}{2}\ln\ 2\pi\ -\left(k_m+\frac{1}{2}\right)\ln(k_m)+k_m+\varepsilon(k_m)$$
$$-\frac{1}{2}\ln\ 2\pi\ -\left(n-k_m+\frac{1}{2}\right)\ln(n-k_m)+(n-k_m)+\varepsilon(n-k_m)+k_m\ln\ p\ +(n-k_m)\ln\ q$$

$$\ln\ P\ Y=k_m\ =\left(n+\frac{1}{2}\right)\ln(n)-\left(k_m+\frac{1}{2}\right)\ln(k_m)+k_m\ln\ p\ -\left(n-k_m+\frac{1}{2}\right)\ln(n-k_m)+(n-k_m)\ln\ q$$
$$-\frac{1}{2}\ln\ 2\pi\ +\varepsilon(n)$$

$$\ln\ P\ Y=k_m\ =\left(n+\frac{1}{2}\right)\ln(n)-\left(k_m+\frac{1}{2}\right)\ln(k_m)-\ln\ p\ -\left(n-k_m+\frac{1}{2}\right)\ln(n-k_m)-\ln\ q$$
$$-\frac{1}{2}\ln\ 2\pi\ -\frac{1}{2}\ln\ p\ -\frac{1}{2}\ln\ q\ +\varepsilon(n)$$

$$\ln\ P\ Y=k_m\ =\left(n+\frac{1}{2}\right)\ln(n)-\left(k_m+\frac{1}{2}\right)\ln\left(\frac{k_m}{p}\right)-\left(n-k_m+\frac{1}{2}\right)\ln\left(\frac{n-k_m}{q}\right)$$
$$-\frac{1}{2}\ln\ 2\pi pq\ +\varepsilon(n)$$

$$\ln\ P\ Y=k_m\ =-\left(k_m+\frac{1}{2}\right)\left(\ln\left(\frac{k_m}{p}\right)-\ln(n)\right)-\left(n-k_m+\frac{1}{2}\right)\left(\ln\left(\frac{n-k_m}{q}\right)-\ln(n)\right)$$
$$-\frac{1}{2}\ln(n)-\frac{1}{2}\ln\ 2\pi pq\ +\varepsilon(n)$$

$$\ln\ P\ Y=k_m\ =-\left(k_m+\frac{1}{2}\right)\ln\left(\frac{k_m}{np}\right)-\left(n-k_m+\frac{1}{2}\right)\ln\left(\frac{n-k_m}{nq}\right)-\frac{1}{2}\ln\ 2\pi npq\ +\varepsilon(n)$$

$$\ln\ \sqrt{2\pi npq}P\ Y=k_m\ =-\left(k_m+\frac{1}{2}\right)\ln\left(\frac{k_m}{np}\right)-\left(n-k_m+\frac{1}{2}\right)\ln\left(\frac{n-k_m}{nq}\right)+\varepsilon(n)$$

on pose $z_{k_m}=\dfrac{k_m-np}{\sqrt{npq}}$ donc $k_m=np+z_{k_m}\sqrt{npq}$ et $\dfrac{k_m}{np}=1+z_{k_m}\dfrac{\sqrt{q/p}}{\sqrt{n}}$

et $n-k_m=n-np-z_{k_m}\sqrt{npq}$, $n-k_m=n(1-p)-z_{k_m}\sqrt{npq}$

donc $n-k_m=nq-z_{k_m}\sqrt{npq}$ et $\dfrac{n-k_m}{nq}=1-z_{k_m}\dfrac{\sqrt{p/q}}{\sqrt{n}}$

($np+a\sqrt{npq}<k_m<np+b\sqrt{npq}$ donc $a<z_{k_m}<b$)

et on peut faire un d.l. à l'ordre 3 de ln : $\ln(1+h)=h-\frac{1}{2}h^2+\frac{1}{3}h^3+\circ(h^3)=h-\frac{1}{2}h^2+O(h^3)$

puisque $\frac{1}{3}h^3+\circ(h^3)=\left(\frac{1}{3}+\circ(1)\right)h^3=O(h^3)$, on obtient :

$$\left(k_m+\frac{1}{2}\right)\ln\left(\frac{k_m}{np}\right)=\left(k_m+\frac{1}{2}\right)\ln\left(1+\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}\right)$$
$$=\left(k_m+\frac{1}{2}\right)\left(\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}-\frac{1}{2}\frac{q}{np}z_{k_m}^2+O\left(\frac{1}{n\sqrt{n}}\right)\right)$$

$$\left(n-k_m+\frac{1}{2}\right)\ln\left(\frac{n-k_m}{np}\right)=\left(n-k_m+\frac{1}{2}\right)\ln\left(1-\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}\right)$$
$$=\left(n-k_m+\frac{1}{2}\right)\left(-\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}-\frac{1}{2}\frac{p}{nq}z_{k_m}^2+O\left(\frac{1}{n\sqrt{n}}\right)\right)$$

d'où

$$\ln\ \sqrt{2\pi npq}P\ \ Y=k_m\ \ =-\left(np+z_{k_m}\sqrt{npq}+\frac{1}{2}\right)\left(\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}-\frac{1}{2}\frac{q}{np}z_{k_m}^2+O\left(\frac{1}{n\sqrt{n}}\right)\right)$$
$$-\left(nq-z_{k_m}\sqrt{npq}+\frac{1}{2}\right)\left(-\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}-\frac{1}{2}\frac{p}{nq}z_{k_m}^2+O\left(\frac{1}{n\sqrt{n}}\right)\right)+O\left(\frac{1}{n}\right)$$

$$\left(\text{car}\quad \varepsilon(n)\sim\frac{1}{n}=O\left(\frac{1}{n}\right)\right)$$

quand on développe, on constate que

$$\ln\ \sqrt{2\pi npq}P\ \ Y=k_m\ \ =-np\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}+\frac{1}{2}qz_{k_m}^2-qz_{k_m}^2-\frac{1}{2}\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}-\frac{1}{4}\frac{q}{np}z_{k_m}^2+\frac{1}{2}\frac{\sqrt{q^3/p}}{\sqrt{n}}z_{k_m}^3$$
$$+nq\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}+\frac{1}{2}pz_{k_m}^2-pz_{k_m}^2+\frac{1}{2}\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}+\frac{1}{4}\frac{p}{nq}z_{k_m}^2-\frac{1}{2}\frac{\sqrt{p^3/q}}{\sqrt{n}}z_{k_m}^3+O\left(\frac{1}{n}\right)$$
$$=-\frac{1}{2}z_{k_m}^2-\frac{1}{2\sqrt{n}}\ \sqrt{q/p}-\sqrt{p/q}\ \ z_{k_m}+\frac{1}{2\sqrt{n}}\ \sqrt{q^3/p}-\sqrt{p^3/q}\ \ z_{k_m}^3+O\left(\frac{1}{n}\right)$$
$$=-\frac{1}{2}z_{k_m}^2+O_1\left(\frac{1}{\sqrt{n}}\right)$$

où $O_1\left(\frac{1}{\sqrt{n}}\right)=-\frac{1}{2\sqrt{n}}\ \sqrt{q/p}-\sqrt{p/q}\ \ z_{k_m}+\frac{1}{2\sqrt{n}}\ \sqrt{q^3/p}-\sqrt{p^3/q}\ \ z_{k_m}^3+\circ\left(\frac{1}{\sqrt{n}}\right)$

car $\quad -np\frac{\sqrt{q/p}}{\sqrt{n}}z_{k_m}+nq\frac{\sqrt{p/q}}{\sqrt{n}}z_{k_m}=0\qquad$ et $\qquad z_{k_m}^2\frac{1}{2}q+z_{k_m}^2\frac{1}{2}p-z_{k_m}^2q-z_{k_m}^2p=-\frac{1}{2}z_{k_m}^2$

(c'est ce qui fait que cela marche!)

donc $\ln\ \sqrt{2\pi npq}P\ \ Y=k_m\ \ =-\frac{1}{2}z_{k_m}^2+O_1\left(\frac{1}{\sqrt{n}}\right)$

$$\sqrt{2\pi npq}P\ \ Y=k_m\ \ =exp\left(-\frac{1}{2}z_{k_m}^2+O_1\left(\frac{1}{\sqrt{n}}\right)\right)$$

$$\sqrt{2\pi npq}P\ \ Y=k_m\ \ =exp\left(-\frac{1}{2}z_{k_m}^2\right)exp\left(O_1\left(\frac{1}{\sqrt{n}}\right)\right)$$

avec le d.l. à l'ordre 1 de exp : $e^h=1+h+\circ(h)$ donc $e^{O_1(h)}=1+O_1(h)+\circ(h)$

$$\sqrt{2\pi npq}P\ \ Y=k_m\ \ =exp\left(-\frac{1}{2}z_{k_m}^2\right)\left(1+O_1\left(\frac{1}{\sqrt{n}}\right)+\circ\left(\frac{1}{\sqrt{n}}\right)\right)=exp\left(-\frac{1}{2}z_{k_m}^2\right)\left(1+O_2\left(\frac{1}{\sqrt{n}}\right)\right)$$

ainsi $\varepsilon(n,k_m)=O_2\left(\frac{1}{\sqrt{n}}\right)$.

De plus

$$O_2\left(\frac{1}{\sqrt{n}}\right)=-\frac{1}{2\sqrt{n}}\ \sqrt{q/p}-\sqrt{p/q}\ \ z_{k_m}+\frac{1}{2\sqrt{n}}\ \sqrt{q^3/p}-\sqrt{p^3/q}\ \ z_{k_m}^3+\circ\left(\frac{1}{\sqrt{n}}\right)$$

$$\left|O_2\left(\frac{1}{\sqrt{n}}\right)\right|\leqslant\frac{1}{2\sqrt{n}}\left|\ \sqrt{q/p}-\sqrt{p/q}\ \ z_{k_m}\right|+\frac{1}{2\sqrt{n}}\left|\ \sqrt{q^3/p}-\sqrt{p^3/q}\ \ z_{k_m}^3\right|+\left|\circ\left(\frac{1}{\sqrt{n}}\right)\right|$$

$$\leqslant\frac{1}{2\sqrt{n}}\left[\left|\sqrt{q/p}-\sqrt{p/q}\right||b|+\ \sqrt{q^3/p}-\sqrt{p^3/q}\ \ |b|+\left|\varepsilon_1\left(\frac{1}{\sqrt{n}}\right)\right|\right]\qquad\text{si}\quad |b|\geqslant|a|$$

où $\lim_{n\to\infty}\varepsilon_1\left(\frac{1}{\sqrt{n}}\right)=0$, et donc ε_1 est borné au voisinage de 0.

ainsi $\exists K\in\mathbb{R}_+\quad \left|O_2\left(\frac{1}{\sqrt{n}}\right)\right|\leqslant K\frac{1}{\sqrt{n}}$ pour tout k_n

d'où le résultat.

Deuxième étape

On pose $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ **,** on peut écrire :

$$P\ \ a<Z_n<b\ =\sum_{i=1}^{j_n}P\ \ Y_n=k_{n,i}\ =\sum_{i=1}^{j_n}\frac{1}{\sqrt{npq}}f\left(\frac{k_{n,i}-np}{\sqrt{npq}}\right)1+\varepsilon(n,k_{n,i})$$

soit

$$P\ \ a<Z_n<b\ =\sum_{i=1}^{j_n}P\ \ Y_n=k_{n,i}\ =\sum_{i=1}^{j_n}\frac{1}{\sqrt{npq}}f\left(\frac{k_{n,i}-np}{\sqrt{npq}}\right)+\varepsilon(n,k_{n,i})\sum_{i=1}^{j_n}\frac{1}{\sqrt{npq}}f\left(\frac{k_{n,i}-np}{\sqrt{npq}}\right)$$

d'une part :

- dans l'intervalle $\left]a\sqrt{npq}+np\ ;\ b\sqrt{npq}+np\right[$ il y a j_n entiers et $j_n\leqslant(b-a)\sqrt{npq}$,
- f continue, est bornée sur a,b par M
- $\left|\varepsilon(n,k_{n,i})\right|\leqslant\frac{K}{\sqrt{n}}$ pour tout $k_{n,i}$

donc le second terme est tel que

$$\left|\varepsilon(n,k_{n,i})\frac{1}{\sqrt{npq}}\sum_{i=1}^{j_n}\ f\left(\frac{k_{n,i}-np}{\sqrt{npq}}\right)\right|\leqslant\frac{j_nM.K}{\sqrt{npq}\sqrt{n}}\leqslant\frac{M.(b-a)\sqrt{npq}K}{\sqrt{npq}\sqrt{n}}=\frac{M.(b-a)K}{\sqrt{n}}$$

et sa limite est donc 0.

d'autre part

$\frac{1}{\sqrt{npq}}\sum_{k=1}^{j_n} f\left(\frac{k_{n,i}-np}{\sqrt{npq}}\right)$ est une somme de Riemann ($\sqrt{npq}$ est le pas de la subdivision) de la fonction f continue sur a,b qui est donc intégrable sur cet intervalle, or la limite de toute somme de Riemann est $\int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx$

et c'est fini

remarque

on peut montrer que $\left|P\left(a < Z_n < b\right) - \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx\right| < \frac{0,588}{\sqrt{npq}}$

voir le cours de Charles SUQUET (Université de Lille1)

Fonction caractéristique et Théorème Central Limite

Le théorème de Moivre-Laplace est un cas particulier d'un théorème plus général : le théorème central limite. La preuve repose sur l'utilisation de la transformée de Fourier et le théorème de Paul Lévy (qui sera admis), le cadre est celui de l'intégrale dite abstraite ou de Lebesgue.

1. Fonction caractéristique

Définition

Soit μ une mesure finie sur ($\mathbb{R}$,$Bor(\mathbb{R})$ tribu de Borel sur $\mathbb{R}$ **) ;** la tribu engendrée par les intervalles ouverts.

On appelle transformée de Fourier de μ la fonction $\hat{\mu} : \mathbb{R} \longrightarrow \mathbb{C}$ définie par :

$$\forall t \in \mathbb{R} \qquad \hat{\mu}(t) = \int_{\mathbb{R}} e^{itx} d\mu(x)$$

Lorsque μ est **la loi d'une variable aléatoire** X, $\hat{\mu}$ est la **fonction caractéristique** de X.

$$\text{si} \quad \mu = P_X \qquad \hat{\mu}(t) = \varphi_X(t) = E\, exp(itX)$$

exemple 1 pour une loi discrète

$$P_X\ A = \sum_{i \in \square} p_i \mathcal{X}_{a_i}\ A \qquad \text{où} \quad p_i \in \mathbb{R}_+^* \quad \text{et} \quad a_i \in \mathbb{R}$$

$$\varphi(t) = \int_{\square} e^{itx} dP_X(x) = \sum_{i \in \square} p_i e^{ita_i} = E\ e^{itX}$$

exemple 2 pour une loi à densité *f*

$$\begin{cases} Bor(\mathbb{R}) & \xrightarrow{P_X} \quad 0\,;+\infty \\ A & \longrightarrow \int \mathcal{X}_A\, f\, d\mu \end{cases}$$

$$\varphi(t) = \int_{\mathbb{R}} e^{itx} dP_X(x) = \int_{\mathbb{R}} e^{itx}\, d\ \ f.\mu\ (x) = \int_{\mathbb{R}} e^{itx} f(x)\, d\mu(x) = \int_{\mathbb{R}} e^{itx} f(x)\, dx = E\ e^{itX}$$

si μ est la mesure de Borel sur $\mathbb{R}$

et parce que l'intégrale de Borel et de Riemann coïncident si f est Riemann intégrable.

Propriétés

1. Effet d'une transformation affine : $\forall t \in \mathbb{R} \qquad \varphi_{aX_1+b}(t) = \varphi_X(at)e^{itb}$

2. Théorème d'injectivité
Deux mesures bornées sur $\mathbb{R}$ qui ont la même transformée de Fourier sont égales .

3. φ_X est uniformément continue sur $\mathbb{R}$

4. si X de carré intégrable, φ_X est deux fois dérivable et en particulier

$$\varphi'_X(0) = iE(X)$$
$$\varphi''_X(0) = -E(X^2)$$

5. si X_1 et X_2 sont deux variables aléatoires réelle **indépendantes**,

$$\forall t \in \mathbb{R} \qquad \varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t)$$

Exemples

1. la loi de Bernoulli de paramètre p

$$\varphi_X(t) = E\exp(itX) = qe^{it\times 0} + pe^{it\times 1}$$

$$\boxed{\varphi_X(t) = q + pe^{it}}$$

2. la loi binomiale $B(n,p)$

$$\varphi_X(t) = E\exp(itX) = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}e^{ikt} = \sum_{k=0}^{n}\binom{n}{k}(pe^{it})^k q^{n-k}$$

$$\boxed{\varphi_X(t) = \left(q + pe^{it}\right)^n}$$

3. la loi uniforme sur $[0;1]$

$$\varphi_X(t) = \int_0^1 e^{ixt}dx = \left[\frac{e^{ixt}}{it}\right]_0^1$$

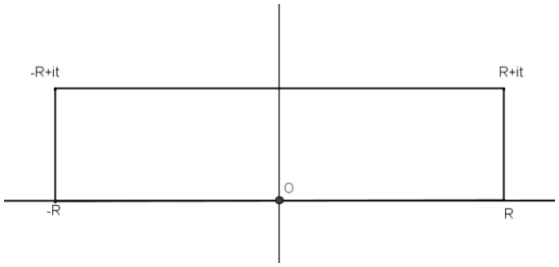
$$\boxed{\varphi_X(t) = \frac{e^{it}-1}{it}}$$

4. la loi normale centrée réduite *N(0,1)*

$$\varphi_X(t)=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{ixt}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{x^2-2itx}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{(x-it)^2+t^2}{2}}dx$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}}dx$$

pour calculer $\int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}}dx$, le plus simple est de considérer $\int_{\gamma} e^{-\frac{z^2}{2}}dz$ où γ est le chemin fermé rectangulaire ci-dessous , pout t fixé



puisque $z\longrightarrow e^{-\frac{z^2}{2}}$ est holomorphe sur $\mathbb{C}$ d'après le théorème de Cauchy

$$\int_{\gamma} e^{-\frac{z^2}{2}}dz=0$$

donc $$\int_{-R}^{R} e^{-\frac{z^2}{2}}dz+\int_{R}^{R+it} e^{-\frac{z^2}{2}}dz+\int_{R+it}^{-R+it} e^{-\frac{z^2}{2}}dz+\int_{-R+it}^{-R} e^{-\frac{z^2}{2}}dz=0$$

$$\int_{R+it}^{-R+it} e^{-\frac{z^2}{2}}dz=\int_{-R}^{R} e^{-\frac{(x-it)^2}{2}}dx=-\int_0^t ie^{-\frac{(R+iy)^2}{2}}dy+\int_{-R}^{R} e^{-\frac{x^2}{2}}dx-\int_t^o ie^{-\frac{(-R+iy)^2}{2}}dy$$

$$\begin{pmatrix} z=x-it \\ dz=dx \end{pmatrix} \qquad \begin{pmatrix} z=R+iy \\ dz=idx \end{pmatrix} \qquad \begin{pmatrix} z=x \\ dz=dx \end{pmatrix} \qquad \begin{pmatrix} z=-R+iy \\ dz=idx \end{pmatrix}$$

$$e^{-\frac{(R+iy)^2}{2}}=e^{-\frac{R^2}{2}-iRy+\frac{y^2}{2}} \qquad \text{donc} \qquad \left|e^{-\frac{(R+iy)^2}{2}}\right|=e^{-\frac{R^2}{2}}e^{\frac{y^2}{2}}\leqslant e^{-\frac{R^2}{2}}e^{\frac{t^2}{2}}$$

$$\left|\int_0^t ie^{-\frac{(R+iy)^2}{2}}dy\right|\leqslant\int_0^t\left|ie^{-\frac{(R+iy)^2}{2}}\right|dy\leqslant te^{-\frac{R^2}{2}}e^{\frac{t^2}{2}}$$

donc $$\lim_{R\to+\infty}\int_{R}^{R+it} e^{-\frac{z^2}{2}}dz=0 \quad \text{et de même} \quad \lim_{R\to+\infty}\int_{-R-it}^{-R} e^{-\frac{z^2}{2}}dz=0$$

d'où $$\lim_{R\to\infty}\int_{-R}^{R} e^{-\frac{(x-it)^2}{2}}dx=\lim_{R\to\infty}\int_{-R}^{R} e^{-\frac{x^2}{2}}dx=\sqrt{2\pi}$$

$$\varphi_X(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}}dx=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\times\sqrt{2\pi}$$

$$\boxed{\varphi_X(t)=e^{-\frac{t^2}{2}}}$$

Remarque

pour calculer $\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx$, le plus simple est de considérer $\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy$

on utilise deux recouvrements de $\mathbb{R}^2$

d'une part

$$\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{D_R} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{D_R} e^{-r^2} \left|\det J(r,\theta)\right| drd\theta = \lim_{R\to+\infty} \iint_{D_R} e^{-r^2} rdrd\theta$$

$$= \lim_{R\to+\infty} \left(\int_0^R re^{-r^2} dr \right) \left(\int_0^{2\pi} d\theta \right)$$

$$= \lim_{R\to+\infty} \left[-\frac{1}{2} e^{-r^2} \right]_0^R \times \theta \Big|_0^{2\pi} = \lim_{R\to+\infty} \frac{1}{2} \left(1 - e^{-R^2}\right) \times 2\pi = \pi$$

$$car \quad \left|J(r,\theta)\right| = \left|\frac{D(x,y)}{D(r,\theta)}\right| = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

d'autre part avec le théorème de Fubini

$$\iint_{\mathbb{R}^2} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \iint_{-R;R\times -R;R} e^{-x^2-y^2} dxdy = \lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right) \left(\int_{-R}^R e^{-y^2} dy \right) = \lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right)^2$$

d'où

$$\lim_{R\to+\infty} \left(\int_{-R}^R e^{-x^2} dx \right)^2 = \pi \qquad \text{et} \quad \int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi} \qquad \text{(l'intégrale de Gauss)}$$

d'où enfin $\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx = \lim_{R\to+\infty} \int_{-R}^R e^{-\frac{x^2}{2}} dx = \lim_{R\to+\infty} \int_{-R/\sqrt{2}}^{R/\sqrt{2}} e^{-y^2} \sqrt{2} dy = \lim_{R\to+\infty} \sqrt{2} \int_{-R/\sqrt{2}}^{R/\sqrt{2}} e^{-y^2} dy = \sqrt{2\pi}$

2. Théorème de Lévy

Définition convergence en loi

Soit, pour tout n , X_n une variable aléatoire réelle définie sur l'espace probabilisé $\Omega_n, \mathcal{T}_n, P_n$ et une variable aléatoire réelle X définie sur l'espace probabilisé $\Omega, \mathcal{T}, P$

la suite $\left(X_n\right)_{n\in\mathbb{N}}$ converge en loi vers X

si et seulement si $\forall x \in \mathbb{R} \qquad \lim_{n\to+\infty} P\left(X_n \leqslant x\right) = P\left(X \leqslant x\right)$

*ce qui est un abus de langage : c'est la **loi** de X_n qui converge vers celle de X*
ou de toute autre variable de même loi que

et nous admettons le théorème suivant :

Théorème de Lévy

Soient $X_{n\ n\in\mathbb{N}}$ une suite de variables aléatoires réelles et X une variable aléatoire réelle dont les fonctions caractéristiques respectives sont $\varphi_{X_{n}\ n\in\mathbb{N}}$ et φ_X ,

(1) si $X_{n\ n\in\mathbb{N}}$ converge en loi vers X ,
alors $\varphi_{X_{n}\ n\in\mathbb{N}}$ converge simplement vers φ_X
(et même uniformément sur tout compact de er $\mathbb{R}$ **)**

(2) si $\varphi_{X_{n}\ n\in\mathbb{N}}$ converge simplement vers une fonction φ continue en 0 ,
alors φ est la fonction caractéristique d'une variable aléatoire réelle X
et $X_{n\ n\in\mathbb{N}}$ converge en loi vers X

3. Théorème central limite

Théorème central limite

Si $(X_n)_n$ est une suite de variables aléatoires définies sur le même espace probabilisé $\Omega,\mathcal{T},P$ indépendantes et de même loi, d'espérance μ et de variance σ^2 finies, donc de carré intégrable

Alors la variable aléatoire $F_n=\dfrac{X_1+X_2+...+X_n}{n}$ ***converge en loi*** vers une variable aléatoire qui suit la loi normale $N\left(\mu;\dfrac{\sigma^2}{n}\right)$

ou bien $Z_n=\dfrac{F_n-E(F_n)}{\sigma(F_n)}=\dfrac{F_n-\mu}{\sigma/\sqrt{n}}$ ***converge en loi*** vers la loi normale $N(0\ ;1)$

Preuve

si la fonction caractéristique de $Y_i=\dfrac{X_i-\mu}{\sigma}$ est φ_Y

$\forall t\in\mathbb{R}\quad \varphi_{Y_1+...+Y_n}(t)=\ \varphi_Y(t)^{\ n}$

$Y_1+Y_2+...+Y_n=\dfrac{X_1+X_2+...+X_n-n\mu}{\sigma}=n\left(\dfrac{F_n-\mu}{\sigma}\right)=\dfrac{1}{\sqrt{n}}Z_n$

$\forall t\in\mathbb{R}\quad \varphi_{Z_n}(t)=\left(\varphi_Y\left(\dfrac{1}{\sqrt{n}}t\right)\right)^n$

comme Y_i est de carré intégrable, φ est dérivable deux fois et on peut appliquer le théorème de Taylor -Young à l'ordre 2 au voisinage de 0

$\forall u \in V(0) \qquad \varphi(u) = \varphi(0) + \varphi'(0)u + \varphi''(0)\frac{u^2}{2} + \circ(u^2)$

$\varphi'(0) = iE(Y) = i\left(\frac{E(X) - \mu}{\sigma}\right) = 0$

$\varphi''(0) = -E(Y^2) = -\ E(Y^2) - \ E(Y)^{\ 2}\ = -Var(Y) = -1$

donc pour t fixé,

$\forall \frac{t}{\sqrt{n}} \in V(0) \qquad \varphi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \circ\left(\frac{t^2}{n}\right)$

$\varphi(0) = 1$, donc par continuité, $\exists \varepsilon > 0$ tel que $|u| < \varepsilon \Rightarrow |\varphi(u)| > 0$ et $-\frac{\pi}{2} < Arg(\varphi(u) < \frac{\pi}{2}$

donc $log\left(\varphi\left(\frac{t}{\sqrt{n}}\right)\right)$ est bien défini pour *n* assez grand (le log complexe)

$\left|\frac{t}{\sqrt{n}}\right| < \varepsilon \qquad \varphi_{Z_n}(t) = exp\left(n\,log\left(\varphi\left(\frac{t}{\sqrt{n}}\right)\right)\right) = exp\left(n\left(-\frac{t^2}{2n} + \circ\left(\frac{t^2}{n}\right)\right)\right)^n = exp\left(-\frac{t^2}{2} + O\left(\frac{t^2}{\sqrt{n}}\right)\right)$

$\lim_{n \to +\infty} \varphi_{Z_n}(t) = exp\left(-\frac{t^2}{2}\right)$

et le théorème de Paul Lévy permet de conclure.

Annexes

Annexe 1 : Formule de Stirling

1. formule des trapèzes

$$\int_a^b f(x)dx = \frac{b-a}{2}\ f(b)-f(a)\ -\frac{1}{12}(b-a)f''(\xi) \quad \text{où} \quad \xi \in a,b$$

preuve

tout d'abord

$$\int_a^b f(x)dx = \frac{b-a}{2}\ f(b)-f(a)\ -\frac{1}{2}\int_a^b (x-a)(b-x)f''(x)dx$$

avec deux intégrations par parties

$$\int_a^b (x-a)(b-x)f''(x)dx = \left[(x-a)(b-x)f'(x)\right]_a^b - \int_a^b (a+b-2x)f'(x)dx$$

$$= 0\ -\ \left[(a+b-2x)f(x)\right]_a^b + \int_a^b 2f(x)dx\Big]$$

d'où le résultat

et avec le

premier théorème de la moyenne

Si f est continue sur $[a,b]$

g est de signe constant et intégrable au sens de Riemann sur $[a,b]$

Alors. $\exists c \in [a,b]$ tel que $\int_a^b f(t)g(t)dt = f(c)\int_a^b g(t)dt$

$$\int_a^b (x-a)(b-x)f''(x)dx = f''(\xi)\int_a^b (x-a)(b-x)dx = \frac{1}{6}(b-a)^3$$

d'où le résultat

2. formule de Moivre

pour $k \in \mathbb{N}^*$

$$\int_k^{k+1} \ln x dx = \frac{1}{2}(\ln k + \ln(k+1)) + \frac{1}{12}\frac{1}{\xi_k^2} \qquad \text{où} \qquad k < \xi_k < k+1$$

$$\int_1^n \ln x\ dx = \ln 2 + ... + \ln(n-1) + \frac{1}{2}\ln n + \frac{1}{12}\sum_{k=1}^{n-1}\frac{1}{\xi_k^2}$$

$$n\ln n - n + 1 = \ln\ n!\ - \frac{1}{2}\ln n + \frac{1}{12}\sum_{k=1}^{n-1}\frac{1}{\xi_k^2}$$

$$\left(n+\frac{1}{2}\right)\ln n - n - \ln\ n! \ = 1+\frac{1}{12}\sum_{k=1}^{n-1}\frac{1}{\xi_k^2}$$

$$\ln\left(\frac{n^{n+\frac{1}{2}}\,e^{-n}}{n!}\right) = 1+\frac{1}{12}\sum_{k=1}^{n-1}\frac{1}{\xi_k^2}$$ et comme la série à termes positifs est convergente

$$\lim_{n\to\infty}\ln\left(\frac{n^{n+\frac{1}{2}}\,e^{-n}}{n!}\right) = H$$ et donc $$\ln\left(\frac{n^{n+\frac{1}{2}}\,e^{-n}}{n!}\right) = H - \frac{1}{12}\sum_{k=n+1}^{+\infty}\frac{1}{\xi_k^2}$$

$$\frac{1}{k+1^{\ 2}} \leqslant \frac{1}{\xi_k^2} \leqslant \frac{1}{k^2} \qquad x\in\ k;k+1\ \ ,\ \ k>1 \quad \Rightarrow \quad \frac{1}{x+1^{\ 2}} \leqslant \frac{1}{k+1^{\ 2}} \leqslant \frac{1}{\xi_k^2} \leqslant \frac{1}{k^2} \leqslant \frac{1}{x-1^{\ 2}}$$

$$\int_k^{k+1} \frac{1}{x+1^{\ 2}}\,dx \leqslant \frac{1}{k+1^{\ 2}} \leqslant \frac{1}{\xi_k^2} \leqslant \frac{1}{k^2} \leqslant \int_k^{k+1} \frac{1}{x-1^{\ 2}}\,dx$$

d'où $$\int_{n+1}^{+\infty}\frac{1}{(x+1)^2}\,dx \leqslant \sum_{k=n+1}^{+\infty}\frac{1}{\xi_k^2} \leqslant \int_{n+1}^{+\infty}\frac{1}{(x-1)^2}\,dx$$

donc $$\frac{1}{n+2} \leqslant \sum_{k=n+1}^{+\infty}\frac{1}{\xi_k^2} \leqslant \frac{1}{n}$$, $$\sum_{k=n+1}^{+\infty}\frac{1}{\xi_k^2} = \frac{1}{n}+O\left(\frac{2}{n(n+1)}\right) = \frac{1}{n}+o\left(\frac{1}{n}\right)$$ et

$$\ln\left(\frac{n^{n+\frac{1}{2}}\,e^{-n}}{n!}\right) = H-\varepsilon_n$$ où $$\varepsilon_n = \frac{1}{12n}+o\left(\frac{1}{n}\right)$$

si $$e^{-(H-\varepsilon_n)} = e^{-H}e^{\varepsilon_n} = K\left(1+\varepsilon_n+o\left(\varepsilon_n\right)\right)$$, $$\frac{n!}{n^{n+\frac{1}{2}}\,e^{-n}} = K\ 1+\varepsilon_n$$

d'où enfin $$\boxed{n! = K\,n^{n+\frac{1}{2}}\,e^{-n}\ 1+\varepsilon_n}$$ avec $$\varepsilon_n = \frac{1}{12n}+o\left(\frac{1}{n}\right)$$

3. il reste à calculer K avec la **formule de Wallis**

on considère les intégrales $I_n = \int_0^{\frac{\pi}{2}} cos^n\,x.dx$ où $n\in\mathbb{N}$

$I_o = \frac{\pi}{2}$, $I_1 = 1$et avec deux intégrations par parties, on obtient $I_{n+2} = \frac{n-1}{n}I_n$ où $n\in$

$$I_{2n} = \frac{1.3....(2n-1)}{2.4....2n}I_0 = \frac{1.3....(2n-1)}{2.4....2n}\frac{\pi}{2}$$

$$I_{2n} == \frac{1.3....(2n-1)}{2.4....2n}\frac{2.4...2n}{2.4...2n}\times\ \frac{\pi}{2} = \frac{(2n)!}{2^{2n}(n!)^2}\times\ \frac{\pi}{2}$$

$$I_{2n} = \frac{3\times5\times\ldots\times(2n-1)}{2\times4\times\ldots\times2n}\times\frac{2\times4\times\ldots\times2n}{2\times4\times\ldots\times2n}\times\frac{\pi}{2} = \frac{(2n)!}{2^{2n}(n!)^2}\times\frac{\pi}{2}$$

$$I_{2n+1} = \frac{2\times4\times\ldots\times2n}{3\times5\times\ldots\times(2n+1)} = \frac{2\times4\times\ldots\times2n}{3\times5\times\ldots\times(2n+1)}\times\frac{2\times4\times\ldots\times2n}{2\times4\times\ldots\times2n} = \frac{2^{2n}(n!)^2}{(2n+1)!}$$

par ailleurs

$\cos^n x \leqslant \cos^{n+1} x \leqslant \cos^{n+2} x$ car $x \in \left[0\,;\frac{\pi}{2}\right]$

d'où $I_n \leqslant I_{n+1} \leqslant I_{n+2}$ et $1 \leqslant \frac{I_{n+1}}{I_n} \leqslant \frac{I_{n+2}}{I_n} = \frac{n+1}{n+2}$

donc $\lim\limits_{n\to+\infty} \frac{I_{n+1}}{I_n} = 1$

$$\lim_{n\to+\infty} \frac{I_{2n+1}}{I_{2n}} = \lim_{n\to+\infty} \frac{2^{2n}\ n!\ ^2}{2n+1\ !}\times\frac{2^{2n}\ n!\ ^2}{2n\ !}\times\frac{2}{\pi} = 1$$

donc $\lim\limits_{n\to+\infty} \frac{2^{2n}\ n!\ ^{2\ 2}}{2n\ !\ ^2\ 2n+1} = \frac{\pi}{2}$

et enfin $\boxed{\lim\limits_{n\to+\infty} \frac{2^{2n}\ n!\ ^2}{2n\ !\ \sqrt{2n+1}} = \sqrt{\frac{\pi}{2}}}$

comme $n! = K\,n^{n+\frac{1}{2}}\,e^{-n}(1+\frac{1}{12}\varepsilon_n)$

$$\lim_{n\to+\infty} \frac{2^{2n}\ n!\ ^2}{2n\ !\ \sqrt{2n+1}} = \sqrt{\frac{\pi}{2}} = \lim_{n\to+\infty} \frac{\left(2^{2n}\left(K\,n^{n+\frac{1}{2}}\,e^{-n}(1+\frac{1}{12}\varepsilon_n)\right)^2\right)}{\left(K\ \ 2n\ ^{2n+\frac{1}{2}}\,e^{-2n}(1+\frac{1}{12}\varepsilon_{2n})\right)\sqrt{2n+1}} = \frac{K}{2}$$

d'où $K = \sqrt{2\pi}$

formule de Stirling :

$$n! = \sqrt{2\pi}\ n^{n+\frac{1}{2}}\,e^{-n}\ 1+\varepsilon_n \qquad \text{avec} \qquad \varepsilon_n = \frac{1}{12n} + o\left(\frac{1}{n}\right)$$

Annexe 2 : Aperçu rapide de l'intégrale de Lebesgue

E un ensemble quelconque non vide et $\mathcal{P}(E)$ l'ensemble des parties de l'ensemble E.

1. Tribus , espaces mesurables et mesures

Définition algèbre de Boole ou clan

$\mathcal{A} \subset \mathcal{P}(E)$ est un clan sur E si

1. $\varnothing \in \mathcal{A}$ et $E \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3. $A \in \mathcal{A}$ et $B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$
4. $A \in \mathcal{A}$ et $B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$

(donc aussi pour toutes réunions et intersections finies)

Définition σ-algèbre ou tribu

$\mathcal{A} \subset \mathcal{P}(E)$ est une tribu sur E si

1. $\mathcal{A}$ est un clan sur E
2. toute intersection dénombrable d'éléments de $\mathcal{A}$ est un élément de $\mathcal{A}$
   (donc toute réunion dénombrable d'éléments de $\mathcal{A}$ est un élément de $\mathcal{A}$)

exemple 1 $\mathcal{P}(E)$

exemple 2 Si E est un espace topologique, on appelle tribu borélienne de E la tribu engendrée par les ouverts de E. C'est l'intersection de toutes les tribus qui contiennent les ouverts de E. On la note *Bor(E).*

(c'est cette tribu qui permet de définir l'intégrale de Lebesgue de façon qu'elle coïncide avec celle de Cauchy-Riemann))

Définition Espace mesurable

Un ensemble E muni d'une tribu $\mathcal{A} \subset \mathcal{P}(E)$ est appelé espace mesurable et noté $E, \mathcal{A}$

Définition Mesure positive

soit $E, \mathcal{A}$ un espace mesurable

On appelle mesure positive sur $E, \mathcal{A}$ toute application

$\mu : \mathcal{A} \longrightarrow 0 \; ; +\infty$ telle que :

1. $\mu \varnothing = 0$.
2. $\mu\left(\bigcup_{n\geqslant 1} A_n\right) = \sum_{n\geqslant 1} \mu A_n$ pour toute suite A_n d'éléments de $\mathcal{A}$ disjoints deux à deux.

(c'est la σ-additivité)

et si $A \in \mathcal{A}$, $\mu\ A$ est la mesure de A.

exemple 1 mesure de Lebesgue

C'est l'unique mesure invariante par translation, λ sur $\mathbb{R}, Bor(\mathbb{R})$ telle que

$\forall\ a,b$, intervalle borné , $\lambda\ \ a,b\ = \lambda\ \ a,b\ = b-a$

exemple 2 mesure de Probabilité (ou probabilité, ou loi de probabilité)

si $\Omega, \mathcal{A}$ est un espace mesurable, toute mesure telle que

$p\ : \mathcal{A} \longrightarrow 0\ ;1$

Définition espace mesuré

On appelle **espace mesuré** le triplet $E, \mathcal{A}, \mu$.

exemple : espace probabilisé $\Omega, \mathcal{A}, P$

2. Fonctions mesurables et fonctions étagées

Définition application mesurable

soient $E, \mathcal{A}$ et $F, \mathcal{B}$ deux espaces mesurables

l'application $f : E \longrightarrow F$ est mesurable si $\forall B \in \mathcal{B}$, $f^{-1}\ B\ \in \mathcal{A}$

Définition fonction étagée

soit $E, \mathcal{A}$ un espace mesurable, une fonction étagée est une fonction numérique de la forme

$$f = \sum_{k=1}^{n} x_k \chi_{A_k} \qquad \text{où} \quad x_k \in \mathbb{R} \quad \text{et} \quad A_k \in \mathcal{A} \qquad \forall k \in\ 0\ ;n$$

elle est donc mesurable.

où l'application χ_A est la **fonction indicatrice** ou caractéristique de A dans E

$$\chi_A\ \ x\ = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon} \end{cases}$$

et où $A_k\ _{0 \leqslant k \leqslant n}$ est une partition de E.

Théorème (fondamental d'approximation)

– Toute fonction mesurable positive est limite simple d'une suite croissante de fonctions étagée positives.

– Toute fonction mesurable réelle est limite simple d'une suite de fonctions étagée

3. Intégrales des fonctions étagées positives et Intégrale des fonctions mesurables positives

Définition Intégrales des fonctions étagées positives

soient $E, \mathcal{A}, \mu$ espace mesuré et f une fonction étagée positive

$$f = \sum_{k=1}^{n} \alpha_k \mathcal{X}_{A_k} \qquad \text{où} \quad x_k \in \mathbb{R}_+ \quad \text{et} \quad A_k \in \mathcal{A}$$

l'intégrale de f par rapport à μ est

$$\int_E f \, d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k)$$

on note aussi $\int_E f \, d\mu = \int_{x \in E} f(x) d\mu(x)$

Définition Intégrales des fonctions mesurables positives

soient $E, \mathcal{A}, \mu$ espace mesuré et f une fonction mesurable positive

l'intégrale de f par rapport à μ est

$$\int_E f \, d\mu = sup \left\{ \int_E h \, d\mu \ , \ h \in \{\text{fonctions étagées positives}\} \ , \ h \leqslant f \right\}$$

on note aussi $\int_E f \, d\mu = \int_{x \in E} f(x) d\mu(x)$

exemple 1 Intégration par rapport à une mesure discrète positive

Une mesure discrète positive est de la forme

$$\mu(A) = \sum_{i \in \square} p_i \mathcal{X}_{a_i}(A) \qquad \text{où} \quad p_i \in \mathbb{R}_+^* \quad \text{et} \quad a_i \in \mathrm{E}$$

Ainsi si f une fonction mesurable positive

$$\int_E f \, d\mu = \sum_{i \in \square} p_i f(a_i)$$

remarque pour une fonction étagée on a

$$\int_E f \, d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k) = \sum_{k=1}^{n} \alpha_k \sum_{i \in \square} p_i \mathcal{X}_{\{a_i\}}(A_k) = \sum_{i \in \square} p_i \sum_{k=1}^{n} \mathcal{X}_{\{a_i\}}(A_k) \alpha_k = \sum_{i \in \square} p_i f(a_i)$$

exemple 2 **Intégration par rapport à une mesure définie par une densité**

Proposition (et définition)

si f est une fonction mesurable positive
alors l'application

$$\begin{cases} \mathcal{A} & \xrightarrow{\nu} & 0\,;+\infty \\ A & \longrightarrow & \int \mathcal{X}_A \, f \, d\mu \end{cases}$$

est une mesure sur $E,\mathcal{A}$ appelée mesure de densité f par rapport à μ.
on la note $f.\mu$

théorème
si $g : F \longrightarrow \overline{\mathbb{R}}^+$ est mesurable et $\nu = f.\mu$

$$\int_E g \, d\nu = \int_E g \, d\left(f.\mu\right) = \int_E fg \, d\mu$$

remarque pour une fonction étagée on a

$$\int_E g \, d\nu = \sum_{k=1}^{n} \alpha_k \nu\left(A_k\right) = \sum_{k=1}^{n} \alpha_k \int_E \mathcal{X}_{A_k} \, f \, d\mu = \int_E \sum_{k=1}^{n} \alpha_k \mathcal{X}_{A_k} \, f \, d\mu = \int_E g \, f \, d\mu$$

4. Intégration des fonctions mesurables quelconques

définition Intégrale d'une fonction mesurable

Soit f une fonction mesurable, on pose $f^+ = sup(\, f\,;0\,)$ et $f^- = inf(\, f\,;0\,)$
alors $f = f^+ - f^-$ et f^+et f^- sont deux fonctions mesurables positives.

f est μ **intégrable** si $\int_E f^+ \, d\mu < +\infty$ et $\int_E f^- \, d\mu < +\infty$

et $\int_E f \, d\mu = \int_E f^+ \, d\mu - \int_E f^- \, d\mu \qquad (< +\infty)$

en d'autres termes, f est absolument μ intégrable

d'autre part on admet que

f admet une intégrale par rapport à μ si $\int_E f^+ \, d\mu < +\infty$ ou $\int_E f^- \, d\mu < +\infty$

(dans $\overline{\mathbb{R}}$, $0 \times \infty = 0$ a $\pm\infty = \pm\infty$ mais $\infty - \infty$ n'a pas de sens)

5. théorème de transfert

Il s'agit d'une formule de changement de variable

proposition mesure image

soient $E,\mathcal{A},\mu$ un espace mesuré , $F,\mathcal{B}$ un espace mesurables et

$\varphi : E \longrightarrow F$ une application mesurable

si $\forall B \in \mathcal{B}$, $\varphi^{-1}\ B\ \in \mathcal{A}$ alors la fonction d'ensemble

$$\begin{cases} \mathcal{B} & \xrightarrow{\ \nu\ } & 0\,;+\infty \\ B & \longrightarrow & \mu\ \varphi^{-1}\ B \end{cases}$$ est une mesure sur $F,\mathcal{B}$

c'est la mesure image de μ par φ (notée $\varphi_*\mu$)

théorème de transfert

si $g : F \longrightarrow \overline{\mathbb{R}}^+$ est mesurable et $\nu = \mu \circ \varphi^{-1}$

$\int_F g\,d\nu = \int_E (g \circ \varphi) d\mu$ ou $\int_F f(y) d\nu(y) = \int_E (f \circ \varphi)(x) d\mu(x)$

Partie C

STATISTIQUE INFÉRENTIELLE

Quelques notions de statistique inférentielle

1) Rappel sur la loi normale

Sa fonction de densité est $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ on notera $X \equiv N(\mu;\sigma)$

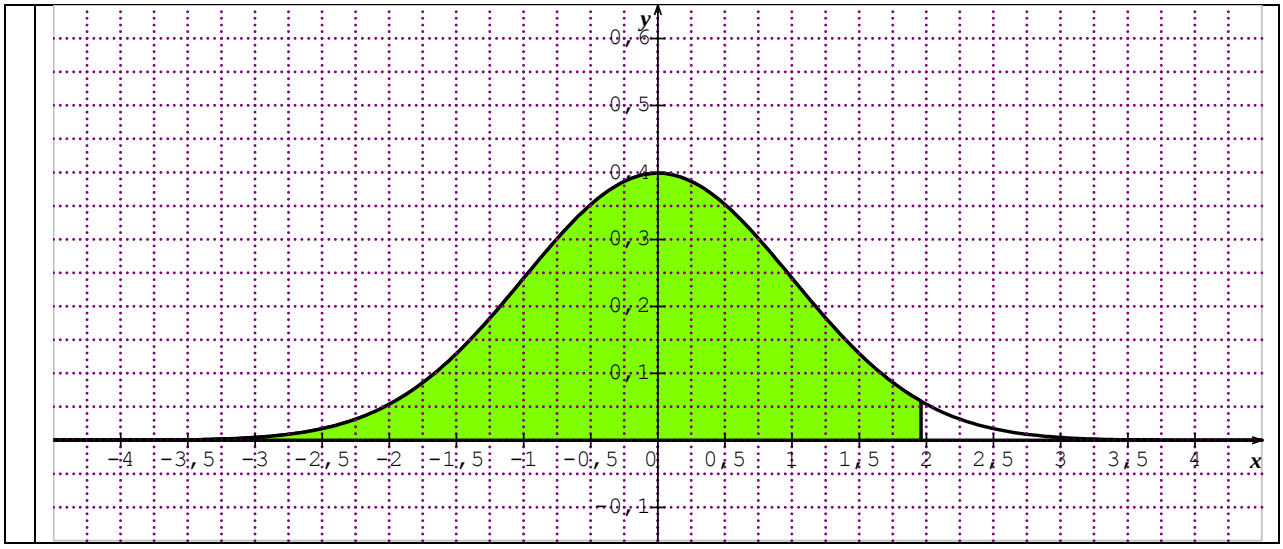
pour signifier que X suit la loi normale d'espérance μ et de variance σ,

autre notation $N(\mu;\sigma^2)$ *Si* $\sigma = 2$ *on écrit* $N(\mu;\ \sigma = 2)$ *pour préciser ou* $N(\mu;\ \sigma^2 = 4)$

On change de variable

On pose $Z = \frac{X-\mu}{\sigma}$ Z est la variable centrée réduite

La fonction de densité de Z est $f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$ *et* $Z \equiv N(0;1)$



L'aire sous la courbe est égale à 1. Donc $P(Z \geq a) = 1 - P(Z \leq a)$

La courbe est symétrique par rapport à l'axe des ordonnées. Donc

$P(Z \leq -a) = P(Z \geq a) = 1 - P(Z \leq a)$

Et aussi

$$P(-a \leq Z \leq a) = P(Z \leq a) - P(Z \leq -a) = P(Z \leq a) - \big(1 - P(Z \leq a)\big)$$
$$= 2P(Z \leq a) - 1$$

Si $X \equiv N(0;1)$ *on cherche a tel que* $P(-a \leq Z \leq a) = 0{,}95$

On cherche donc a tel que $2P(Z \leq a) - 1 = 0{,}95$ *soit* $P(Z \leq a) = 0{,}975$

On trouve $a = 1{,}96$ *donc* $P(-1{,}96 \leq Z \leq 1{,}96) = 0{,}95$

2) Calcul des valeurs associées à la loi normale

- On peut utiliser la table de la loi normale $N(0;1)$ pour calculer des probabilités

On trouve dans la table $P(Z \leq 1{,}96) = 0{,}975 = \pi(1{,}96)$.

Voir un exemple de table en annexe.

- Utilisation de Excel

Fonction	Paramètres	Syntaxe	Résultat
Loi.normale.n	(x, mu , sigma , 1 si cumul)	LOI.NORMALE.N(1,96;0;1;1)	0,9750
	(x, mu , sigma , 0 si densité)	LOI.NORMALE.N(1,96;0;1;0)	0,0584
Loi.normale.standard.n	(x, 1 si cumul)	LOI.NORMALE.STANDARD.N(1,96;1)	0,9750
	(x, 0 si densité)	LOI.NORMALE.STANDARD.N(1,96;0)	0,0584
Loi.normale.inverse	(p, mu , sigma)	LOI.NORMALE.INVERSE(0,975;0;1)	1,9600
Loi.normale.standard.inverse	(p)	LOI.NORMALE.STANDARD.INVERSE.N(0,975)	1,9600

3) Loi normale et fréquence d'un échantillon.

On considère une très grande population. Certains individus présentent un caractère particulier. La proportion d'individus qui présentent ce caractère dans la population est notée p.

Lorsqu'on prélève de façon aléatoire un échantillon de taille n ($n > 30$) dans la population, la fréquence d'apparition du caractère dans l'échantillon est notée f.

Comment se comporte la fréquence de l'échantillon ?

Pour chaque individu choisi, il y a deux possibilités : l'individu présente le caractère ou non.

Soit X_i la variable aléatoire qui prend la valeur 1 si l'individu i présente le caractère et 0 sinon.

Comme la population est très grande, le tirage est considéré comme un tirage avec remise.

La variable aléatoire $S_n = X_1 + X_2 + \cdots \ldots X_n$ prend pour valeurs le nombre d'individus présentant le caractère dans l'échantillon ; S_n suit la loi binomiale $B(n;p)$.

On a $E(S_n) = np \quad V(S_n) = npq$

On appelle F la variable aléatoire qui à chaque échantillon associe sa fréquence f :
(F est la moyenne des X_i)

$$F = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad E(F) = \frac{1}{n}\sum E(X_i) = \frac{1}{n} \times np = p$$

Les variables X_i sont indépendantes deux à deux, on a alors :

$$V(F) = V\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\left[n\,V(X_i)\right] = \frac{1}{n} \times pq \quad donc\ \sigma = \sqrt{\frac{pq}{n}}$$

Le théorème central limite dit que, approximativement :

$$F \equiv N(p; \sqrt{\frac{pq}{n}}) \qquad \text{dès que } n > 30 \text{ ; } np > 5 \text{ ; } nq > 5$$

4) Les trois situations possibles :

Situation 1	Situation 2	Situation 3
p est connue Dans quel intervalle peut-on s'attendre à trouver f ? Et avec quelle probabilité ?	p est inconnue la fréquence f de l'échantillon est connue. On cherche un intervalle de confiance de p	p est connue la fréquence f de l'échantillon est connue. Est-ce que l'échantillon fait partie de la population ?

a) <u>Situation 1</u>

Exemple : dans une population, on vote à 52% pour un candidat. $p = 0{,}52$

On prélève des échantillons de n bulletins de vote. On s'intéresse à la fréquence des bulletins favorables au candidat dans les échantillons choisis.

On cherche dans quel intervalle se trouvera f avec une probabilité de 0,95. Cet intervalle sera appelé **intervalle de fluctuation à 95% ou de risque α = 5% (le programme dit au seuil de 95%).**

On sait que $F \equiv N\left(p\,; \sqrt{\frac{pq}{n}}\right)$; alors $\frac{F-p}{\sqrt{\frac{pq}{n}}} \equiv N(0;1)$

On cherche les deux bornes a et –a telles que

$$P\left(-\text{a} \leq \frac{F-p}{\sqrt{\frac{pq}{n}}} \leq \text{a}\right) = 0{,}95$$

On a vu que a = 1,96 on a donc

$$P\left(-1{,}96 \leq \frac{F-p}{\sqrt{\frac{pq}{n}}} \leq 1{,}96\right) = 0{,}95$$

$$P\left(-1{,}96 \times \sqrt{\frac{pq}{n}} + p \leq F \leq 1{,}96 \times \sqrt{\frac{pq}{n}} + p\right) = 0{,}95$$

L'intervalle $\left[p - 1{,}96 \times \sqrt{\frac{pq}{n}}\ ; p + 1{,}96 \times \sqrt{\frac{pq}{n}}\ \right]$ est l'intervalle de fluctuation à 95%.

On choisit un échantillon au hasard, la probabilité que la fréquence f de cet échantillon soit dans l'intervalle est 0,95

Remarque : en classe de 2nde, on utilise l'intervalle de fluctuation $I = [p - \frac{1}{\sqrt{n}} \; ; \; p + \frac{1}{\sqrt{n}}]$ obtenu par approximation :

$$Comme \;\; pq = p(1-p) = p - p^2 \;\; avec\; p \; dans \; [0;1]$$

p	0	$\frac{1}{2}$	1
$-p^2 + p$	0 ↗	$\frac{1}{4}$ ↘	0

On a donc $pq \leq \frac{1}{4} \;\; et \sqrt{pq} \leq \frac{1}{2} \quad \frac{\sqrt{pq}}{\sqrt{n}} \leq \frac{1}{2\sqrt{n}} \quad et\; comme\; 1{,}96 \approx 2 \quad on\; a \quad 1{,}96 \times \sqrt{\frac{pq}{n}} \leq \frac{1}{\sqrt{n}}$ Et $\left[p - 1{,}96 \times \sqrt{\frac{pq}{n}} \; ; p + 1{,}96 \times \sqrt{\frac{pq}{n}}\right] \subset [\, p - \frac{1}{\sqrt{n}} \; ; \; p + \frac{1}{\sqrt{n}}]$

b) <u>Situation 2</u>

Exemple : lors d'une élection, on prélève à la fermeture des bureaux de vote un échantillon de n=1000 bulletins et on observe que la fréquence de votes pour le candidat A est $f = 0{,}56$. On veut une estimation par intervalle de la proportion p (inconnue au moment du sondage) de votes favorables pour le candidat A dans la population.

On peut estimer ponctuellement p par $\hat{p} = f$.

On appelle *F* la variable aléatoire qui à un échantillon associe sa fréquence.

On sait que

$$F \equiv N\left(p; \sqrt{\frac{pq}{n}}\right) \text{ donc } \frac{F-p}{\sqrt{\frac{pq}{n}}} \equiv N(0;1) \text{ et } P\left(-1{,}96 \leq \frac{F-p}{\sqrt{\frac{pq}{n}}} \leq 1{,}96\right) = 0{,}95$$

ainsi

$$P\left(F - 1{,}96\sqrt{\frac{pq}{n}} \leq p \leq F + 1{,}96\sqrt{\frac{pq}{n}}\right) = 0{,}95$$

On encadre p par des valeurs qu'on ne connaît pas.

Pour l'échantillon donné, F prend la valeur f, p inconnue estimée par $\hat{p} = f$

$I_c = [f - 1{,}96\sqrt{\frac{f(1-f)}{n}}\ ;\ f + 1{,}96\sqrt{\frac{f(1-f)}{n}}\ ]$

I_C est l'intervalle de confiance à 95% (le programme dit : au niveau de confiance 95%)

$I_c = [0{,}56 - 1{,}96\sqrt{\frac{0{,}56\times 0{,}44}{1000}}\ ;\ 0{,}56 + 1{,}96\sqrt{\frac{0{,}56\times 0{,}44}{1000}}]$ $\quad I = [\ 0{,}53\ ;0{,}59\ ]$

Pour diminuer l'amplitude de I, il faut augmenter n.

Attention : la phrase « Il y a 95% de chances que *p* soit dans *I* » n'a pas de sens.

L'échantillon a été choisi, l'intervalle de confiance associé est calculé. La proportion *p* est dans l'intervalle ou n'y est pas.

Chaque échantillon donne un intervalle de confiance. Il y a environ 95% des intervalles qui contiennent *p* et 5 % qui ne le contiennent pas.

La probabilité porte sur le choix de l'intervalle.

c) Situation3

Exemple : on sait que dans une forêt, il y a 61% de champignons vénéneux. Dans un échantillon de 400 champignons, on trouve 56% de vénéneux. Cet échantillon est-il issu de cette forêt ?

C'est ce que le programme désigne par « Prise de décision » (la notion de test statistique et le vocabulaire associé n'est pas au programme).

On appelle F la variable aléatoire qui à chaque échantillon associe sa fréquence. On sait que :

$$F \equiv N(p\ ;\sqrt{\frac{pq}{n}})$$

Deux possibilités :

$p = p_0$ connu (dans l'exemple $p_0 = 0{,}61$) et l'échantillon est considéré comme issu de cette population

ou

$p \neq p_0$ et l'échantillon est considéré comme issu d'une autre population

Classiquement ces 2 possibilités sont dénommées
hypothèse nulle et hypothèse alternative :

H_0 *hypothèse nulle* $p = p_0$

H_1 *hypothèse alternative* $p \neq p_0$

On se place sous H_0 $(p = p_0)$

et on construit un intervalle de fluctuation à 95 % on a alors

$$\frac{F - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \equiv N(0;1)$$

$$P\left(-1{,}96 \leq \frac{F - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \leq 1{,}96\right) = P\left(p_0 - 1{,}96\sqrt{\frac{p_0 q_0}{n}} \leq F \leq p_0 + 1{,}96\sqrt{\frac{p_0 q_0}{n}}\right) = 0{,}95$$

et, comme dit le programme on exploite cet intervalle pour rejeter ou non l'hypothèse sur la proportion p.

Pour un échantillon donné, F prend la valeur f ; et on regarde si f est dans l'intervalle :

$$I = \left[p_0 - 1{,}96\sqrt{\frac{p_0 q_0}{n}} \;;\; p_0 + 1{,}96\sqrt{\frac{p_0 q_0}{n}}\right]$$

Règle de décision :
Si $f \notin I$ on rejette l'hypothèse H_0
au risque α=5% de se tromper (appelé risque de première espèce)
Si $f \in I$ l'échantillon observé ne permet pas de rejeter H_0 ;
on ne connaît pas le risque d'erreur (appelé risque de deuxième espèce)

Remarque : α est la probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie.

Si H_0 est vraie, dans 95% des cas, f est dans I et dans 5% des cas f n'est pas dans I.

Si f est dans I et qu'on se trompe en disant que $p = p_0$ alors F suit une autre loi normale de paramètres inconnus. L'erreur est une fonction de ce paramètre p inconnu : cette erreur de 2e espèce est impossible à connaître.

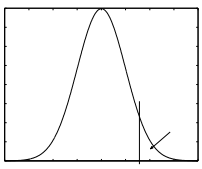
Dans l'exemple donné ; p_0 = 0,61 dans la population. Dans l'échantillon de taille 400, on a $f = 0{,}56$ et l'intervalle

$$I = \left[0{,}61 - 1{,}96\sqrt{\frac{0{,}61 \times 0{,}39}{400}} \;;\; 0{,}61 + 1{,}96\sqrt{\frac{0{,}61 \times 0{,}39}{400}}\right] = [0.5622 \;;\; 0.6578]$$

Ainsi $f = 0{,}56$ n'appartient pas à l'intervalle I ; on rejette l'hypothèse. On peut considérer que notre lot provient d'une autre forêt.

Table de la loi normale

P		+0.001	+0.002	+0.003	+0.004	+0.005	+0.006	+0.007	+0.008	+0.009
0	Inf	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656
0.0100	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0749
0.0200	2.0537	2.0335	2.0141	1.9954	1.9774	1.9600	1.9431	1.9268	1.9110	1.8957
0.0300	1.8808	1.8663	1.8522	1.8384	1.8250	1.8119	1.7991	1.7866	1.7744	1.7624
0.0400	1.7507	1.7392	1.7279	1.7169	1.7060	1.6954	1.6849	1.6747	1.6646	1.6546
0.0500	**1.6449**	1.6352	1.6258	1.6164	1.6072	1.5982	1.5893	1.5805	1.5718	1.5632
0.0600	1.5548	1.5464	1.5382	1.5301	1.5220	1.5141	1.5063	1.4985	1.4909	1.4833
0.0700	1.4758	1.4684	1.4611	1.4538	1.4466	1.4395	1.4325	1.4255	1.4187	1.4118
0.0800	1.4051	1.3984	1.3917	1.3852	1.3787	1.3722	1.3658	1.3595	1.3532	1.3469
0.0900	1.3408	1.3346	1.3285	1.3225	1.3165	1.3106	1.3047	1.2988	1.2930	1.2873
0.1000	1.2816	1.2759	1.2702	1.2646	1.2591	1.2536	1.2481	1.2426	1.2372	1.2319
0.1100	1.2265	1.2212	1.2160	1.2107	1.2055	1.2004	1.1952	1.1901	1.1850	1.1800
0.1200	1.1750	1.1700	1.1650	1.1601	1.1552	1.1503	1.1455	1.1407	1.1359	1.1311
0.1300	1.1264	1.1217	1.1170	1.1123	1.1077	1.1031	1.0985	1.0939	1.0893	1.0848
0.1400	1.0803	1.0758	1.0714	1.0669	1.0625	1.0581	1.0537	1.0494	1.0450	1.0407
0.1500	1.0364	1.0322	1.0279	1.0237	1.0194	1.0152	1.0110	1.0069	1.0027	0.9986
0.1600	0.9945	0.9904	0.9863	0.9822	0.9782	0.9741	0.9701	0.9661	0.9621	0.9581
0.1700	0.9542	0.9502	0.9463	0.9424	0.9385	0.9346	0.9307	0.9269	0.9230	0.9192
0.1800	0.9154	0.9116	0.9078	0.9040	0.9002	0.8965	0.8927	0.8890	0.8853	0.8816
0.1900	0.8779	0.8742	0.8705	0.8669	0.8633	0.8596	0.8560	0.8524	0.8488	0.8452
0.2000	0.8416	0.8381	0.8345	0.8310	0.8274	0.8239	0.8204	0.8169	0.8134	0.8099
0.2100	0.8064	0.8030	0.7995	0.7961	0.7926	0.7892	0.7858	0.7824	0.7790	0.7756
0.2200	0.7722	0.7688	0.7655	0.7621	0.7588	0.7554	0.7521	0.7488	0.7454	0.7421
0.2300	0.7388	0.7356	0.7323	0.7290	0.7257	0.7225	0.7192	0.7160	0.7128	0.7095
0.2400	0.7063	0.7031	0.6999	0.6967	0.6935	0.6903	0.6871	0.6840	0.6808	0.6776
0.2500	0.6745	0.6713	0.6682	0.6651	0.6620	0.6588	0.6557	0.6526	0.6495	0.6464
0.2600	0.6433	0.6403	0.6372	0.6341	0.6311	0.6280	0.6250	0.6219	0.6189	0.6158
0.2700	0.6128	0.6098	0.6068	0.6038	0.6008	0.5978	0.5948	0.5918	0.5888	0.5858
0.2800	0.5828	0.5799	0.5769	0.5740	0.5710	0.5681	0.5651	0.5622	0.5592	0.5563
0.2900	0.5534	0.5505	0.5476	0.5446	0.5417	0.5388	0.5359	0.5330	0.5302	0.5273
0.3000	0.5244	0.5215	0.5187	0.5158	0.5129	0.5101	0.5072	0.5044	0.5015	0.4987
0.3100	0.4959	0.4930	0.4902	0.4874	0.4845	0.4817	0.4789	0.4761	0.4733	0.4705
0.3200	0.4677	0.4649	0.4621	0.4593	0.4565	0.4538	0.4510	0.4482	0.4454	0.4427
0.3300	0.4399	0.4372	0.4344	0.4316	0.4289	0.4261	0.4234	0.4207	0.4179	0.4152
0.3400	0.4125	0.4097	0.4070	0.4043	0.4016	0.3989	0.3961	0.3934	0.3907	0.3880
0.3500	0.3853	0.3826	0.3799	0.3772	0.3745	0.3719	0.3692	0.3665	0.3638	0.3611
0.3600	0.3585	0.3558	0.3531	0.3505	0.3478	0.3451	0.3425	0.3398	0.3372	0.3345
0.3700	0.3319	0.3292	0.3266	0.3239	0.3213	0.3186	0.3160	0.3134	0.3107	0.3081
0.3800	0.3055	0.3029	0.3002	0.2976	0.2950	0.2924	0.2898	0.2871	0.2845	0.2819
0.3900	0.2793	0.2767	0.2741	0.2715	0.2689	0.2663	0.2637	0.2611	0.2585	0.2559
0.4000	0.2533	0.2508	0.2482	0.2456	0.2430	0.2404	0.2378	0.2353	0.2327	0.2301
0.4100	0.2275	0.2250	0.2224	0.2198	0.2173	0.2147	0.2121	0.2096	0.2070	0.2045
0.4200	0.2019	0.1993	0.1968	0.1942	0.1917	0.1891	0.1866	0.1840	0.1815	0.1789
0.4300	0.1764	0.1738	0.1713	0.1687	0.1662	0.1637	0.1611	0.1586	0.1560	0.1535
0.4400	0.1510	0.1484	0.1459	0.1434	0.1408	0.1383	0.1358	0.1332	0.1307	0.1282
0.4500	0.1257	0.1231	0.1206	0.1181	0.1156	0.1130	0.1105	0.1080	0.1055	0.1030
0.4600	0.1004	0.0979	0.0954	0.0929	0.0904	0.0878	0.0853	0.0828	0.0803	0.0778
0.4700	0.0753	0.0728	0.0702	0.0677	0.0652	0.0627	0.0602	0.0577	0.0552	0.0527
0.4800	0.0502	0.0476	0.0451	0.0426	0.0401	0.0376	0.0351	0.0326	0.0301	0.0276
0.4900	0.0251	0.0226	0.0201	0.0175	0.0150	0.0125	0.0100	0.0075	0.0050	0.0025



$P(Z > 1.6449) = .05$

Intervalle de fluctuation avec la loi binomiale

On s'intéresse à l'étude d'un caractère (quantitatif ou qualitatif) des individus d'une population.
Lorsque le caractère a deux modalités (avoir ou non une propriété donnée) , la proportion de l'une des modalités dans la population est notée *p*.
l'expérience est : observer le caractère d'un individu choisi au hasard dans la population.
il y a donc deux issues : ω_1 "avoir la propriété" et ω_2 "ne pas avoir la propriété".
on définit la variable aléatoire X par $X(\omega_1)=1$ et $X(\omega_2)=0$
donc ici $\Omega=\{\omega_1;\omega_2\}$ et $X(\Omega)=\{0;1\}$
X suit donc la loi de Bernoulli de paramètre *p:*
$P(X=1)=P(\omega_1)=p$
$P(X=0)=P(\omega_2)=1-p$

1. Echantillonnage

On procède à un échantillonnage, c'est à dire à un tirage aléatoire de n individus dans la population, sur lesquels on observe un caractère à deux modalités.
On obtient donc un échantillon aléatoire $X_1, X_2, ..., X_n$.

Lorsque les tirages ont lieu avec remise, (ou bien un tirage sans remise, mais sur une population de grande taille), les variables aléatoires X_i sont indépendantes et de même loi $B(p)$, ainsi la variable aléatoire

$Y_n = X_1 + X_2 + ... + X_n$ suit la loi binomiale $B(n, p)$

La fréquence d'un échantillon aléatoire est la variable aléatoire $F_n = \frac{X_1 + X_2 + ... + X_n}{n}$

la fréquence observée sur un échantillon $x_1, x_2, ..., x_n$ est $f = \frac{x_1 + x_2 + ... + x_n}{n}$

remarque 1

les fluctuations d'échantillonnage de F_n autour de *p* sont d'autant plus faibles que *n* est grand.

en effet, $E(Y_n) = \sum_{i=1}^{n} E(X_i) = np = E(nF_n) = nE(F_n)$ et $E(F_n) = p$

et les v.a. X_i étant indépendantes,

$V(Y_n) = \sum_{i=1}^{n} V(X_i) = npq = V(nF_n) = n^2 V(F_n)$ et $V(F_n) = \frac{pq}{n}$

Ainsi l'espérance de F_n est constante, mais sa variance diminue quand n augmente.

remarque 2

Quand la taille de l'échantillon, *n*, tend vers l'infini la fréquence observée f tend vers p.

C'est le théorème :

Théorème : Loi faible des grands nombres

Si $X_1,...,X_n$ sont *n* variables aléatoires **indépendantes** de même espérance μ et de même variance σ^2,

Alors $\overline{X} = F_n = \frac{X_1 + ... + X_n}{n}$ **converge en probabilité** vers μ,

c'est à dire : $\forall \varepsilon > 0, \quad \lim_{n \to +\infty} P\left(\left|\overline{X} - \mu\right) \right| > \varepsilon\right) = 0$

Démonstration : (cas des probabilités discrètes finies)

La démonstration repose sur l'inégalité de Bienaymé-Tchebychev :

Inégalité de Bienaymé-Tchebychev :

Si X est une variable aléatoire d'espérance $E\ X$ de variance $V\ X$

alors $\forall \varepsilon > 0, \quad P\left(\left|X - E(X)\right| > \varepsilon\right) \leqslant \frac{V(X)}{\varepsilon^2}$

qui elle même découle de l'inégalité de Markov

l'inégalité de Markov

Si X est une variable aléatoire

alors $\forall \varepsilon > 0, \quad P\left(\left|X\right| > \varepsilon\right) \leqslant \frac{E\left(\left|X\right|\right)}{\varepsilon}$

(*) **preuve de l'inégalité de Markov**

on considère $A = \left\{x \in X(\Omega) / |x| > \varepsilon\right\}$, alors A et $\bar{A}$ est une partition de $X\left(\Omega\right)$

$$E\ |X| = \sum_{x \in X(\Omega)} |x| P(X = x) = \sum_{x \in A} |x| P(X = x) + \sum_{x \in \bar{A}} |x| P(X = x)$$

donc $E\ |X| \geqslant \sum_{x \in A} |x| P(X = x) \geqslant \varepsilon \sum_{x \in A} P(X = x)$

et $E\ |X| \geqslant \varepsilon P\ X \in A$ soit $E\ |X| \geqslant \varepsilon P\ |X| > \varepsilon$

d'où le résultat.

(**) **preuve de l'inégalité de Bienaymé-Tchebychev**

par définition $E\left(\left(X-E(X)\right)^2\right)=V(X)$

et on applique l'inégalité de Markov à la variable aléatoire $\left(X-E(X)\right)^2$

comme $P\left(\left|X-E(X)\right|>\varepsilon\right)=P\left(\left(X-E(X)\right)^2>\varepsilon^2\right)$

alors $P\left(\left|X-E(X)\right|>\varepsilon\right)\leqslant\frac{V(X)}{\varepsilon^2}$

(***) **preuve loi faible des grands nombres**

$\forall\varepsilon>0,\ P\left(\left|X-E(X)\right|>\varepsilon\right)\leqslant\frac{V(X)}{\varepsilon^2}$

et en particulier pour la variable $\overline{X}$: $P\left(\left|\overline{X}-\mu\right|>\varepsilon\right)\leqslant\frac{V\left(\overline{X}\right)}{\varepsilon^2}$

comme $V\left(\overline{X}\right)=V\left(\frac{1}{n}\left(X_1+...+X_n\right)\right)=\frac{1}{n^2}\left(V(X_1)+...+V(X_2)\right)=\frac{1}{n^2}\times n\sigma^2$

(car $X_1,...,X_n$ sont *n* variables aléatoires indépendantes de même variance σ^2)

alors $P\left(\left|\overline{X}-\mu\right|>\varepsilon\right)\leqslant\frac{\sigma^2}{n\varepsilon^2}$ d'où le résultat : $\lim\limits_{n\to+\infty}P\left(\left|\overline{X}-\mu\right)\right|>\varepsilon\right)=0$.

Ainsi la convergence en probabilité résulte de la convergence en moyenne quadratique.

Cas particulier de la loi faible des grands nombres

Théorème de Bernoulli

Si $X_1,...,X_n$ sont *n* variables de Bernoulli de paramètre *p* **indépendantes** deux à deux (donc de même espérance *p* et de même variance *pq*)

Alors $\overline{X}$ **converge en probabilité** vers *p* (et ce uniformément par rapport à *p*).

c'est à dire : $\forall\varepsilon>0,\quad\lim\limits_{n\to+\infty}P\left(\left|\overline{X}-p\right)\right|>\varepsilon\right)=0$

Si on applique ce résultat à $\overline{X}=F_n$ et $\mu=p$

et puisque nF_n suit la loi binomiale $B\ n,p\ ,V\ F_n\ =\frac{pq}{n}$

on a $P\left(\left|F_n-p\right|>\varepsilon\right)\leqslant\frac{pq}{n\varepsilon^2}\leqslant\frac{1}{4n\varepsilon^2}$.

2. Intervalle de fluctuation à environ 95 % d'une fréquence

Dans le programme de Seconde, l'intervalle de fluctuation d'une fréquence *au seuil* de 95% $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$ est "donné" pour $n \geqslant 25$ *et* $0,2 < p < 0,8$.

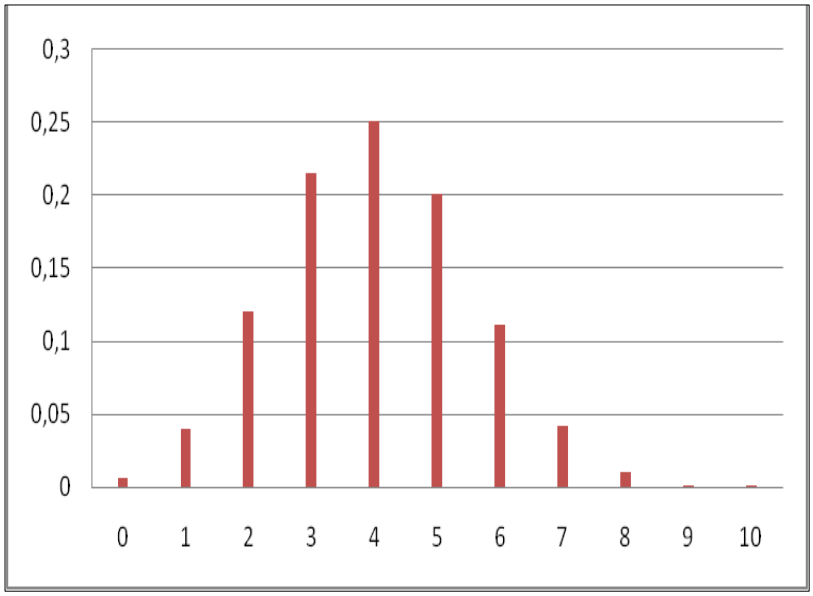
Dans le programme de Première, on utilise la loi binomiale :

$Y_n = X_1 + X_2 + ... + X_n = nF_n$ suit une loi binomiale $B\ n, p$, ce qui permet de déterminer a et b tels que $P(a \leqslant Y_n \leqslant b) \geqslant 0,95$ de la façon suivante :

Pour obtenir au moins 95% , on élimine de proche en proche les valeurs pour lesquelles la probabilité est la plus faible et de même ordre de grandeur aux extrémités de la distribution sans excéder 2,5% de chaque côté .
(ce qui est aisé avec le tableau excel)

exemple 1 $n = 10$ et $p = 0,4$ la distribution de la loi binomiale $B\ 10\ ;\ 0,4$ est

y	p cum	p	1-p cum	
0	0,00605	0,00605	1	
1	0,04636	0,04031	0,99395	**a**
2	0,16729	0,12093	0,95364	
3	0,38228	0,21499	0,83271	
4	0,6331	0,25082	0,61772	
5	0,83376	0,20066	0,3669	
6	0,94524	0,11148	0,16624	
7	0,98771	0,04247	0,05476	**b**
8	0,99832	0,01062	0,01229	
9	0,9999	0,00157	0,00168	
10	1	0,0001	0,0001	



$P(1 \leqslant Y_n \leqslant 7) = 0,040310784 + ... + 0,042467328 = 0,981658829$

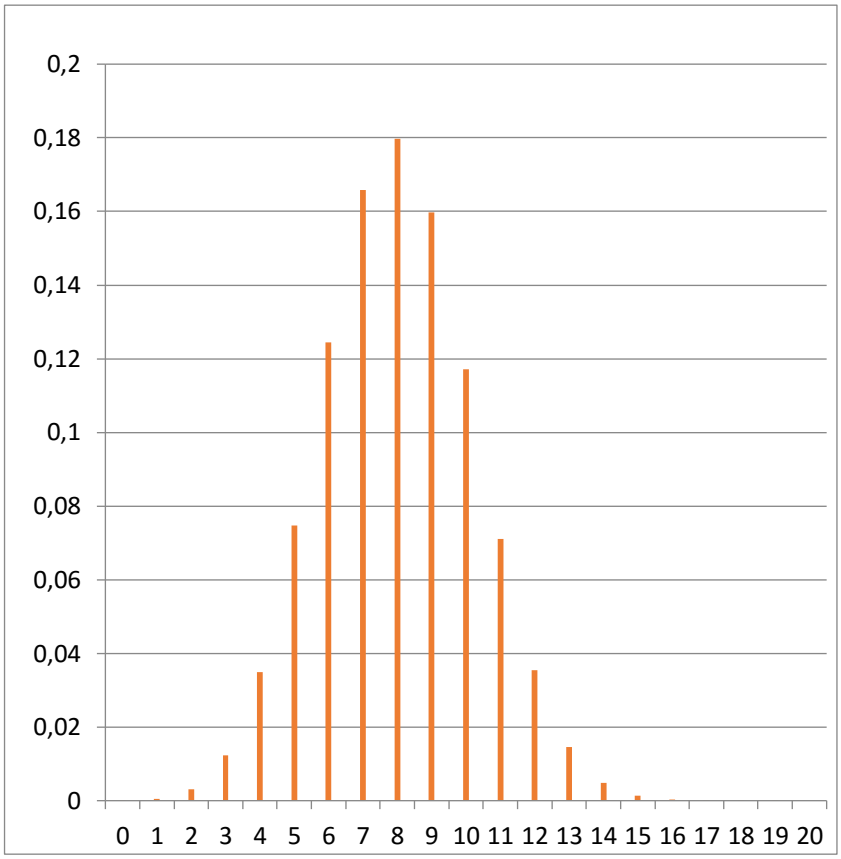
$P(2 \leqslant Y_n \leqslant 6) = 0,120932352 + ... + 0,111476736 = 0,898880717$

pratiquement , pour environ 98% des échantillons, la fréquence observée sera dans l'intervalle [0,1 ; 0,7] et pour environ 90% des échantillons, la fréquence observée sera dans l'intervalle [0,2 ; 0,6], ici il faudra prendre le premier intervalle

Dans ce cas, on ne comparera pas cet intervalle à celui rencontré en seconde, puisque la condition sur n n'est pas respectée.

Exemple 2 pour $n = 20$ et $p = 0,4$

y	p cum	p
0	3,6562E-05	3,7E-05
1	0,00052405	0,00049
2	0,00361147	0,00309
3	0,01596116	0,01235
4	0,05095195	0,03499
5	0,12559897	0,07465
6	0,25001067	0,12441
7	0,41589294	0,16588
8	0,59559873	0,17971
9	0,7553372	0,15974
10	0,87247875	0,11714
11	0,94347363	0,07099
12	0,97897107	0,0355
13	0,99353412	0,01456
14	0,99838848	0,00485
15	0,99968297	0,00129
16	0,99995266	0,00027
17	0,99999496	4,2E-05
18	0,99999966	4,7E-06
19	0,99999999	3,3E-07
20	1	1,1E-08



Ici, on lit $a = 4$ *et* $b = 12$

$P(4 \leqslant Y \leqslant 12) = 0,03499079 + ... + 0,03549744 = 0,96300991$

donc pour 96% des échantillons la fréquence observée est dans l'intervalle [0,2 ; 0,6]

Si on utilise l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ vu en seconde, on obtient $[0,18 \ ; \ 0,63]$, qui est déjà une bonne approximation alors que *n =20* seulement.

Remarque Dans l'accompagnement du programme de 1s, on peut lire :

L'intervalle de fluctuation au seuil de 95 % d'une fréquence *F*, correspondant à la réalisation, sur un échantillon aléatoire de taille *n*, de la variable aléatoire *X* égale à *nF* et de loi binomiale de paramètres *n* et *p*, est l'intervalle $\left[\frac{a}{n},\frac{b}{n}\right]$ défini par le système de conditions suivant :

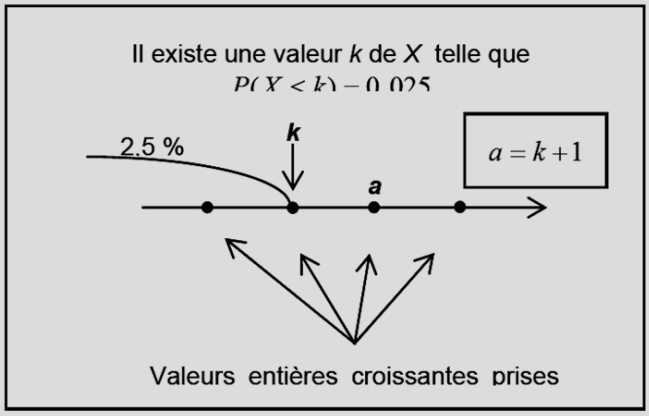
***a* est le plus grand entier tel que *P*(*X* < *a*) ≤ 0,025,**

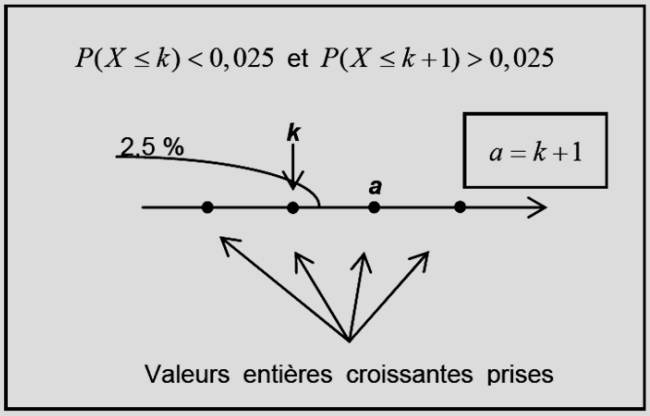
***b* est le plus petit entier tel que *P*(*X* > *b*) ≤ 0,025.**

ou encore par le système de conditions équivalent :

- *a* est le plus petit entier tel que $P(X \leq a) > 0{,}025$,
- *b* est le plus petit entier tel que $P(X \leq b) \geq 0{,}975$.

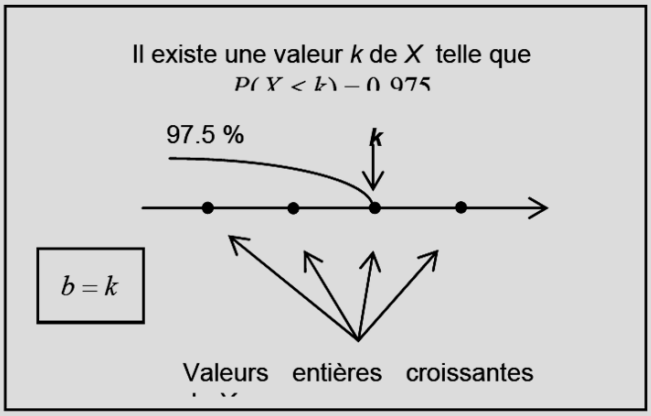
Détermination de *a*

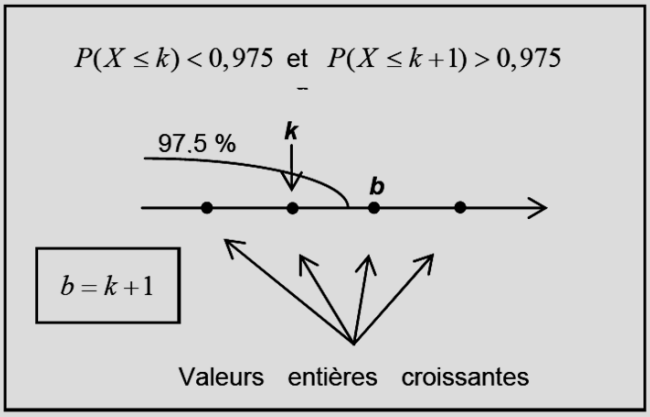




(il faut lire $P(X \leqslant k) = 0{,}025$ *)*

Détermination de *b*





(il faut lire $P(X \leqslant k) = 0{,}975$ *)*

La connaissance de la loi binomiale de la variable aléatoire *X* rend maintenant possible le calcul de la probabilité $P\left(\frac{a}{n} \le F \le \frac{b}{n}\right) = P(a \le X \le b)$.

On remarque que l'intervalle $\left[\frac{a}{n}, \frac{b}{n}\right]$ est quasiment centré sur *p* dès que *n* est « assez grand » et que l'intervalle $\left[\frac{a}{n}, \frac{b}{n}\right]$ est « quasiment » le même que l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ donné dans le programme de seconde pour les « grandes binomiales » ($n > 25$ et $0{,}2 < p < 0{,}8$ où *n* est la taille de l'échantillon prélevé et *p* est la proportion dans la population du caractère étudié, conditions énoncées dans le programme de seconde).

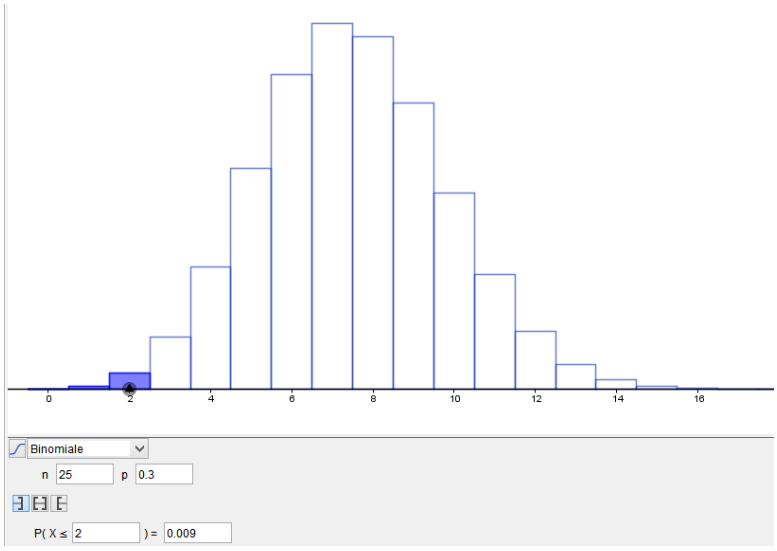
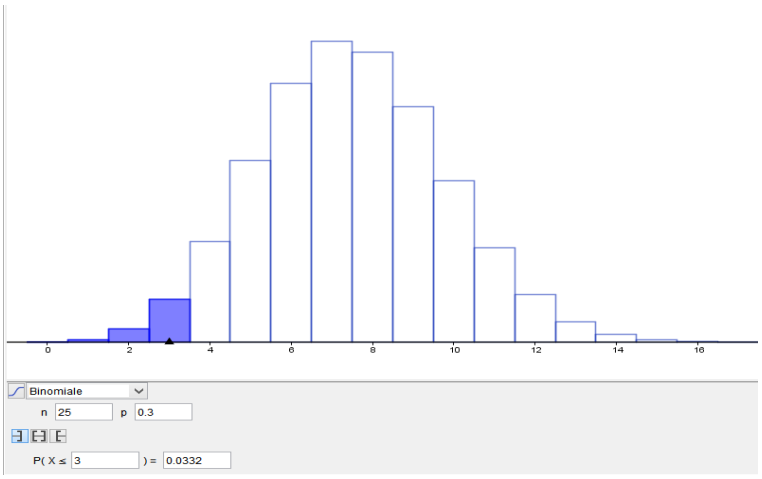
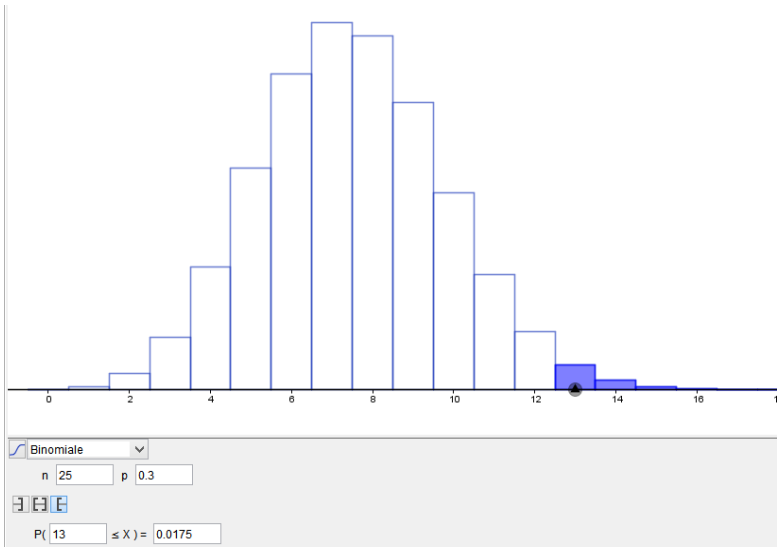
L'intérêt de l'intervalle $\left[\frac{a}{n}, \frac{b}{n}\right]$ (qu'il conviendrait de noter $\left[\frac{a_n}{n}, \frac{b_n}{n}\right]$ pour être précis), calculé à partir de la loi binomiale, est de fournir un intervalle convenable **pour toutes les valeurs de *n* et de *p***, alors que l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ **n'est pas adapté** pour les « petites binomiales ».

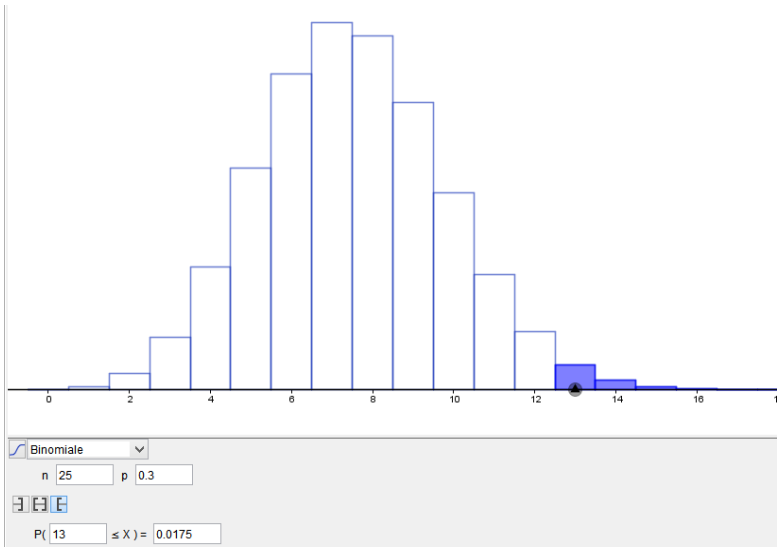
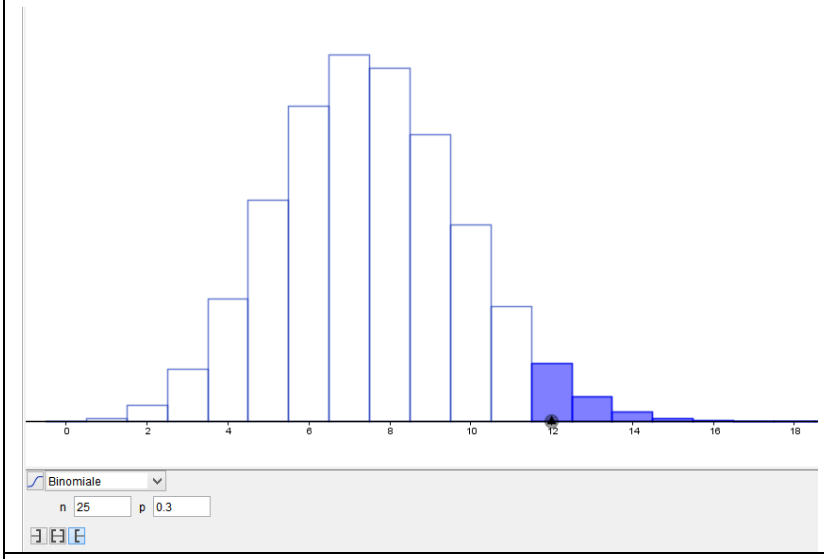
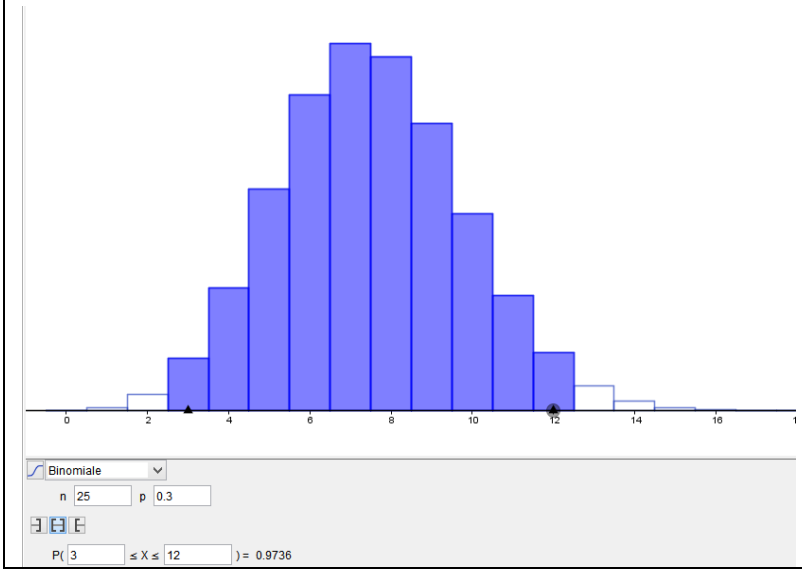
- on doit donc avoir $P(a \leqslant Y_n \leqslant b) \geqslant 0{,}95$
  et $P(a+1 \leqslant Y_n \leqslant b) < 0{,}95$
  et $P(a \leqslant Y_n \leqslant b-1) < 0{,}95$

- on obtient ainsi le "vrai" intervalle de fluctuation de F_n puisque l'on utilise la vraie loi de nF_n et non une approximation de celle-ci par une loi normale, avec laquelle on obtient l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

- remarquons que l'intervalle $[\mu - 2\sigma \ ; \mu + 2\sigma \ ] = \left[p - 2\sqrt{\frac{pq}{n}} ; p + 2\sqrt{\frac{pq}{n}}\right]$ est un intervalle de fluctuation, à **environ** 95 % , construit avec une approximation par la loi normale et peut donc permettre de localiser plus rapidement les valeurs a et b.

- on pourra utiliser Géogebra pour obtenir facilement les valeurs de *a* et *b*.
  dans le module "calculs de probabilités"

On obtient pour $n = 25$ et $p = 0{,}3$	
	
	on choisit donc $a = 3$
	

	
	on choisit donc b = 12
	donc on a un intervalle à 97,36 %

3. Exploiter cet intervalle pour rejeter ou non une hypothèse sur une proportion : prise de décision

C'est formulé ainsi dans les programmes officiels.

On suppose que l'on connaît p , mais que l'on doute a priori de sa valeur.

• Sous l'hypothèse que p prend une valeur donnée p_0 , (hypothèse H_0) à l'aide de la loi binomiale $B\ n\ ;\ p_0$ on détermine un intervalle de fluctuation à environ 95% $\left[p_1\ ;\ p_2\right]$.

Ainsi sous l'hypothèse H_0, la probabilité pour que la fréquence d'un échantillon appartienne à $\left[p_1\ ;\ p_2\right]$ est environ 0,95.

• la **règle de décision est la** suivante :

pour un échantillon dont la fréquence observée est *f*, alors

si $f \notin \left[p_1\ ;\ p_2\right]$, on rejette l'hypothèse $p = p_0$.

si $f \in \left[p_1\ ;\ p_2\right]$ on ne rejette pas l' hypothèse.

Cette règle de décision présente deux types d'erreur :

1. Si l'hypothèse H_0 est vraie, la probabilité que la fréquence observée de l'échantillon n'appartienne pas à l'intervalle est 0,05. La probabilité de rejeter à tort l'hypothèse H_0 est 0,05.

C'est le **risque de première espèce** α choisi a priori : ici $\alpha = P_{p=p_0}\ F \notin\ p_1\ ;\ p_2\ \ \simeq 0,05$

en d'autres termes, si l'échantillon fait partie des 5% dont la fréquence sort de l'intervalle, on commettra une erreur en rejetant l'hypothèse H_0 .

2. Si on ne rejette pas l'hypothèse, l'erreur est appelée **"risque de deuxième espèce** β **".**

Cette probabilité de **ne pas rejeter à tort** est la fonction définie sur $\left[\ 0\ ;1\right]$ par

$$\beta\ x\ = P_{p=x}\ F \in\ p_1\ ;\ p_2$$

Remarque : lorsque H_0 est vraie, la probabilité de ne pas **rejeter** H_0 est

$P_{p=\ p_0}\ F \in\ p_1\ ;\ p_2\ \ = 1-\alpha = 95\%$, qui donc est importante.

Avec l'exemple 2 :

1. si on obtient pour un échantillon $f = 0,65$

et si l'hypothèse H_0 est $p_0 = 0,4$

puisque $f \notin \left[0,2\ ;\ 0,6\right]$ on rejette l'hypothèse $p_0 = 0,4$.

Remarque : $\left[0,2\ ;\ 0,6\right]$ est un intervalle de fluctuation à 96,3%, le risque de rejeter à tort est donc 3,7%.

2. si on obtient pour un échantillon $f = 0{,}48$

et si l'hypothèse H_0 est $p_0 = 0{,}4$

puisque $f \in [0{,}2\,;0{,}6]$ on ne rejette pas l'hypothèse .

Exemple 3 pour $n = 100$ et $p = 0,4$

k	k/100	P cum
29	0,29	0,01477532
30	0,3	0,02478282
31	0,31	0,03984788
32	0,32	0,06150391
33	0,33	0,0912536
34	0,34	0,13033653
35	0,35	0,17946935
36	0,36	0,23861071
37	0,37	0,30680976
38	0,38	0,38218766
39	0,39	0,46207534
40	0,4	0,54329449
41	0,41	0,62253268
42	0,42	0,69673987
43	0,43	0,76346882
44	0,44	0,82109837
45	0,45	0,86890955
46	0,46	0,90701991
47	0,47	0,93621082
48	0,48	0,95769858
49	0,49	0,9729008
50	0,5	0,98323831
51	0,51	0,98999486
52	0,52	0,99423935
53	0,53	0,99680207

on a : $a = 31$ et $b = 50$

$$P(31 \leqslant Y \leqslant 50) = P(Y \leqslant 50) - P(Y \leqslant 30)$$
$$= 0{,}983238313 - 0{,}024782823 = 0{,}95845549$$

donc pour environ 95% des échantillons la fréquence observée est dans l'intervalle [0,31 ; 0,50] .

avec la "formule" $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$, on obtient $[0{,}30\ ;\ 0{,}50]$

qui est donc une bonne approximation.

($n \geqslant 30$, $np \geqslant 5$ et $nq \geqslant 5$)

Si on obtient pour un échantillon $f = 0{,}48$, alors $f \in [0{,}31\,;0{,}50]$

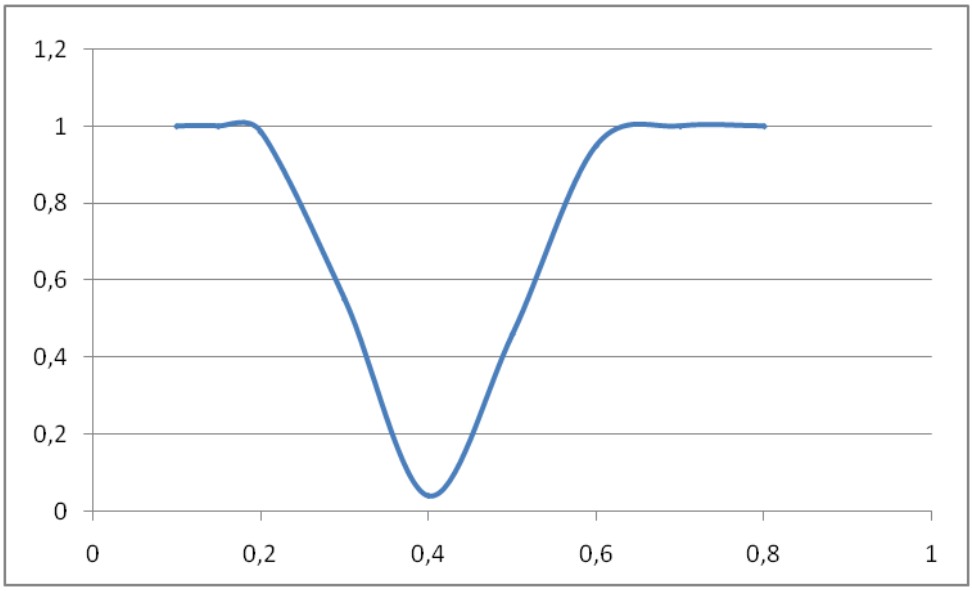
et on ne rejette pas l' hypothèse $p = 0{,}4$.

De plus, on peut déterminer $\beta\ x = P_{p=x}\ F \in 0{,}31\,;0{,}50$ pour quelques valeurs de *x*, qui est la probabilité de ne pas rejeter $p = p_0$ alors que $p = x$.

À l'aide d'un autre tableau Excel où $p = x$ et en faisant varier *x*, on obtient:

x	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
B(x)	6.10^{-9}	0,006	0,45	0,96	0,54	0,027	2.10^{-5}	2.10^{-11}	6.10^{-24}

Et on peut illustrer cela avec la courbe de β, ou celle de $1-\beta$ appelée courbe de puissance du test comme ci-dessous:



Remarque

Avec l'exemple 2 , on a $\beta(x) = P_{p=x}$ $F \in 0,2\,;0,6$ et on obtient:

P_1	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$B(x)$	0,13	0,59	0,89	0,96	0,87	0,58	0,23	0,03	0,0004

ce qui permet de constater que β augmente quand n diminue.

Intervalles de fluctuation et intervalles de confiance avec la loi normale

1. Intervalle de fluctuation en seconde et en première

L'intervalle de fluctuation $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ est l'intervalle de fluctuation de F_n lorsque la valeur du paramètre p est connu.

Il signifie que la probabilité que la fréquence F_n d'un échantillon aléatoire de taille n appartienne à l'intervalle $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ est" proche" de 0,95.

D'où vient-il ?

- la probabilité 0,95 est choisie à priori .

- Lorsqu'on choisit un individu au hasard dans la population, on peut définir la variable aléatoire X qui prend les deux valeurs , 1 et 0, selon que l'individu possède ou non la dite propriété. X suit alors la loi de Bernoulli de paramètre p.

- Si on répète n fois cette expérience, donc si on procède à n tirages d'un individu, et si les tirages ont lieu **avec remise**, (ou bien un tirage sans remise, mais sur une population de grande taille) les variables aléatoires X_i sont indépendantes (ou considérées comme telles) et de même loi $B\ p$.
la variable aléatoire $Y_n = X_1 + X_2 + ... + X_n = nF_n$ suit alors la loi binomiale $B\ n, p$
l'espérance de cette loi est np et l'écart type $\sqrt{np(1-p)} = \sqrt{npq}$.
La proportion de la modalité d'un l'échantillon aléatoire est la variable aléatoire
$F_n = \frac{X_1 + X_2 + ... + X_n}{n}$
la proportion calculée à partir d'une réalisation de l'échantillon $x_1, x_2, ..., x_n$ est f

- la variable aléatoire F_n correspondant à la fréquence observée sur un échantillon de taille n suit une loi d'espérance $E(F_n) = \frac{1}{n} \times np = p$ et d'écart type $\sigma(F_n) = \frac{1}{n}\sqrt{npq} = \sqrt{\frac{pq}{n}}$.

● Le théorème central limite , ou plus précisément son cas particulier pour des variables de Bernoulli : le théorème de Moivre-Laplace, permet d'approcher cette loi par la loi normale $N\left(p\,;\frac{pq}{n}\right)$

Les conditions d'utilisation sont, en classe de seconde : $n \geqslant 25$ et $0{,}2 < p < 0{,}8$
et en classe de première : $n \geqslant 30$, $np \geqslant 5$ et $nq \geqslant 5$.

Théorème de Moivre-Laplace

Si $(X_n)_n$ est une suite de variables aléatoires indépendantes de même loi,

d'espérance μ et de variance σ^2 finies et si $F_n = \frac{X_1 + X_2 + ... + X_n}{n}$,

$Z_n = \frac{F_n - E(F_n)}{\sigma(F_n)} = \frac{F_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ ***converge en loi*** vers Z où Z suit la loi normale $N(0\,;1)$

c'est-à-dire que pour tout $a < b$: $\lim\limits_{n\to+\infty} P(a < Z_n < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

● On détermine alors a tel que $P(p - a \leqslant F_n \leqslant p + a) = 1 - 0{,}05 = 0{,}95$.

Puisque $Z_n = \frac{F_n - p}{\sqrt{\frac{pq}{n}}}$ suit approximativement la loi $N(0;1)$, à l'aide de la table de $N(0;1)$ ou de la calculatrice, on détermine u_α tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 0{,}95$:

$$P(-u_\alpha \leqslant Z \leqslant u_\alpha) = \Pi(u_\alpha) - \Pi(-u_\alpha) = \Pi(u_\alpha) - (1 - \Pi(u_\alpha)) = 2\Pi(u_\alpha) - 1$$

$2\Pi(u_\alpha) - 1 = 0{,}95$ d'où $\Pi(u_\alpha) = 0{,}975$

on lit $u_\alpha = 1{,}96$ et donc $a = 1{,}96\sqrt{\frac{pq}{n}}$.

● Ainsi sensiblement 95 % des échantillons aléatoires de taille n fournissent une fréquence f appartenant à l'intervalle $\left[p - 1{,}96\sqrt{\frac{pq}{n}}\,;\, p + 1{,}96\sqrt{\frac{pq}{n}}\right]$ puisque

$$P\left(p - 1{,}96\sqrt{\frac{pq}{n}} \leqslant F_n \leqslant p + 1{,}96\sqrt{\frac{pq}{n}}\right) \simeq 0{,}95$$

● et comme le maximum de la fonction $p \longrightarrow p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2$ sur $[0;1]$ est $\frac{1}{4}$

$$\left[p-1{,}96\sqrt{\frac{pq}{n}} ; p+1{,}96\sqrt{\frac{pq}{n}} \right] \subset \left[p-\sqrt{\frac{1}{n}} \ ; \ p+\sqrt{\frac{1}{n}} \right]$$

Alors $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ **contiendra sensiblement 95 % des fréquences observées** sur les échantillons de taille ***n*.**

soit $P\left(F_n \in \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right] \right) \simeq 0{,}95$

remarque

On peut montrer que (programme de terminale S)

$\exists n_0 \in \mathbb{N}$ tel que $n \geqslant n_0 \Rightarrow P\left(p-1{,}96\sqrt{\frac{pq}{n}} \leqslant F_n \leqslant p+1{,}96\sqrt{\frac{pq}{n}} \right) \geqslant 0{,}95$

Mais cela ne signifie pas que 95% des *m* fréquences observées seront nécessairement dans l'intervalle même si cela est fortement probable pour *m* assez grand (loi des grands nombres)

2. Intervalle de confiance en seconde et en première

L'intervalle de confiance $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}} \right]$ 'est l'intervalle de confiance du paramètre p lorsque ce paramètre est **inconnu.**

Il signifie que comme $P\left(p \in \left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}} \right] \right) \simeq 0{,}95$**,** (ici F_n est une variable aléatoire) quand on prélève un échantillon aléatoire de taille n , et que l'on calcule la fréquence f , on "estime" que p appartient à $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}} \right]$ avec un **niveau de confiance** de 0,95.(c'est une estimation par intervalle).

On dit aussi que : $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}} \right]$ est un intervalle de confiance de p au **niveau** de 0,95.

d'où vient-il ?

On sait que la probabilité que la fréquence F_n d'un échantillon aléatoire de taille n appartienne à l'intervalle $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$ est proche de 0,95 pour une valeur de p donnée.

***algébriquement, (formellement) on a* :**

$$F_n \in \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right] \Leftrightarrow p-\frac{1}{\sqrt{n}} \leqslant F_n \leqslant p+\frac{1}{\sqrt{n}} \Leftrightarrow F_n-\frac{1}{\sqrt{n}} \leqslant p \leqslant F_n+\frac{1}{\sqrt{n}} \Leftrightarrow$$

$$p \in \left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}} \right]$$

Ainsi $P\left(p \in \left[F_n-\frac{1}{\sqrt{n}} ; F_n+\frac{1}{\sqrt{n}} \right] \right) = P\left(F_n \in \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right] \right) \simeq 0{,}95$

On en déduit que pour 95 % des échantillons aléatoires de taille n possibles, le paramètre inconnu p appartient à l'intervalle $\left[f-\frac{1}{\sqrt{n}} ; f+\frac{1}{\sqrt{n}} \right]$.

Cet intervalle est déterminé par l'échantillon, il y a donc autant d'intervalles que d'échantillons.

C'est à dire que si l'on avait un grand nombre d'échantillons, donc d'intervalles $I_i = \left[f_i - \frac{1}{\sqrt{n}} ; f_i + \frac{1}{\sqrt{n}} \right]$, et si l'on choisissait au hasard l'un d'eux, pour cette nouvelle expérience aléatoire, épreuve de Bernoulli, on aurait $P(p \in I) = 0{,}95$ et $P(p \notin I) = 0{,}05$.
(I est un intervalle aléatoire)

Mais on ne peut pas dire que :

p a 95 % de chances d'appartenir à l' intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$.

(erreur souvent commise dans la littérature, ou disons abus de langage toléré)

En effet, $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ n'est pas un intervalle aléatoire, l'expérience est réalisée et il n'y a plus de probabilité. Cependant pour exprimer qu'avant réalisation la probabilité d'obtenir un intervalle qui contienne *p* était 0,95,

on dit seulement que :

$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ est un intervalle de confiance de *p* au **niveau** de 0,95.

Remarque
On peut noter aussi qu'il y a un risque de confusion (implicite) entre "inconnue" et "variable aléatoire"

3. L'intervalle de confiance en classe de terminale STI2D et STL

L'intervalle $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\ ;\ f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)} \right]$ est l'intervalle de confiance au niveau 0,95 de *p*, qui est au programme de terminale STL et STI2D.

D'où vient-il ?

Le point de départ n'est plus $P\left(p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}} \right] \right) \simeq 0{,}95$, (p 86)

$$\text{mais} \quad P\left(p - 1{,}96\sqrt{\frac{pq}{n}} \leqslant F_n \leqslant p + 1{,}96\sqrt{\frac{pq}{n}} \right) \simeq 0{,}95 .$$

Donc, sous la condition $n \geqslant 30$, $np \geqslant 5$ et $nq \geqslant 5$, la probabilité que la fréquence F_n d'un échantillon aléatoire de taille *n* appartienne à l'intervalle $\left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} ; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right]$ est proche de 0,95 pour une valeur de *p* donnée.

algébriquement, on a aussi **(méthode du trinôme)**

$$F_n \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} ; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right] \Leftrightarrow p - 1{,}96\frac{1}{\sqrt{n}}\sqrt{p(1-p)} \leqslant F_n \leqslant p + 1{,}96\frac{1}{\sqrt{n}}\sqrt{p(1-p)}$$

$$\Leftrightarrow -1{,}96\frac{1}{\sqrt{n}}\sqrt{p(1-p)} \leqslant F_n - p \leqslant 1{,}96\frac{1}{\sqrt{n}}\sqrt{p(1-p)} \Leftrightarrow |F_n - p| \leqslant 1{,}96\frac{1}{\sqrt{n}}\sqrt{p(1-p)}$$

si on pose $u = 1{,}96$,

$$\Rightarrow p^2 - 2F_n p + {F_n}^2 \leqslant \frac{u^2}{n} p(1-p) \Rightarrow \left(1 + \frac{u^2}{n}\right) p^2 - \left(2F_n + \frac{u^2}{n}\right) p + {F_n}^2 \leqslant 0$$

$$\Delta = \left(2F_n + \frac{u^2}{n}\right)^2 - 4\left(1 + \frac{u^2}{n}\right){F_n}^2 = \frac{u^4}{n^2} + 4\frac{u^2}{n} F_n(1-F_n)$$

ainsi $p \in \left[\frac{F_n + \frac{u^2}{2n} - \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + F_n(1-F_n)}}{1 + \frac{u^2}{n}} ; \frac{F_n + \frac{u^2}{2n} + \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + F_n(1-F_n)}}{1 + \frac{u^2}{n}} \right]$

remarque

l'intervalle ci dessus permet d'obtenir un autre intervalle de confiance, intervalle qui n'est pas au programme.

Comme $\frac{u^2}{2n} = o\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{n}\right)$ et $\frac{u^2}{n} = o\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{n}\right)$, on peut écrire

$$p\in\left[\frac{F_n-\frac{u}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)+o\left(\frac{1}{\sqrt{n}}\right)}+o\left(\frac{1}{\sqrt{n}}\right)}{1+o\left(\frac{1}{\sqrt{n}}\right)};\frac{F_n+\frac{u}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)+o\left(\frac{1}{\sqrt{n}}\right)}+o\left(\frac{1}{\sqrt{n}}\right)}{1+o\left(\frac{1}{\sqrt{n}}\right)}\right]$$

et avec deux développements limités à l'ordre 1 on a

$$p\in\left[F_n-\frac{u}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)}+o\left(\frac{1}{\sqrt{n}}\right);F_n+\frac{u}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)}+o\left(\frac{1}{\sqrt{n}}\right)\right]$$

et comme $P\left(a<Z_n<b\right)=\int_a^b\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx+O\left(\frac{1}{\sqrt{n}}\right)$ où $Z_n=\frac{F_n-E\left(F_n\right)}{\sigma\left(F_n\right)}=\frac{F_n-p}{\sqrt{pq/n}}$

il semble donc légitime de considérer que

$$P\left(p\in\left[F_n-\frac{1,96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\ ;\ F_n+\frac{1,96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\right]\right)\simeq 0,95$$

voyons cela :

En échangeant le rôle de F_n et p , on obtient

$$p\in\left[F_n-\frac{1,96}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)}\ ;\ F_n+\frac{1,96}{\sqrt{n}}\sqrt{F_n\left(1-F_n\right)}\right]\quad\Leftrightarrow$$

$$F_n\in\left[p-1,96\sqrt{\frac{p(1-p)}{n}}+o\left(\frac{1}{\sqrt{n}}\right);p+1,96\sqrt{\frac{p(1-p)}{n}}+o\left(\frac{1}{\sqrt{n}}\right)\right]$$

(où $o\left(\frac{1}{\sqrt{n}}\right)\simeq\frac{2}{n}-\frac{4p}{n}\pm\frac{1}{n\sqrt{n}}\left(8\sqrt{pq}-\frac{1}{\sqrt{pq}}\right)$, en fait, on peut vérifier que $\left|o\left(\frac{1}{\sqrt{n}}\right)\right|\leqslant\frac{3}{n}$)

rappelons que :

$P\ \left(p-1,96\sqrt{pq/n}\leqslant F_n\leqslant p+1,96\sqrt{pq/n}\right)=P\ \left(-1,96\leqslant Z_n\leqslant 1,96\right)$ et que, d'après le théorème de

Moivre Laplace on considère que $Z_n=\frac{F_n-p}{\sqrt{pq/n}}$ suit "approximativement" la loi $N\left(0\,;1\right)$

par conséquent :

$$P\left(F_n\in\left[p-1,96\sqrt{\frac{p(1-p)}{n}}+o_1\left(\frac{1}{\sqrt{n}}\right);p+1,96\sqrt{\frac{p(1-p)}{n}}+o_2\left(\frac{1}{\sqrt{n}}\right)\right]\right)\simeq\int_{borne\ inf}^{borne\ sup}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

où $borne\,inf=\frac{p-1,96\sqrt{\frac{p(1-p)}{n}}+o_1\left(\frac{1}{\sqrt{n}}\right)-p}{\sqrt{pq/n}}$ $et\ borne\ sup=\frac{p+1,96\sqrt{\frac{p(1-p)}{n}}+o_2\left(\frac{1}{\sqrt{n}}\right)-p}{\sqrt{pq/n}}$

$$P\left(F_n\in\left[p-1,96\sqrt{\frac{p(1-p)}{n}}+o_1\left(\frac{1}{\sqrt{n}}\right);p+1,96\sqrt{\frac{p(1-p)}{n}}+o_2\left(\frac{1}{\sqrt{n}}\right)\right]\right)\simeq\int_{-1,96+O_1\left(\frac{1}{\sqrt{n}}\right)}^{1,96+O_2\left(\frac{1}{\sqrt{n}}\right)}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

mais

$$\left|\int_{-1{,}96+O_1\left(\frac{1}{\sqrt{n}}\right)}^{1{,}96+O_2\left(\frac{1}{\sqrt{n}}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du - \int_{-1{,}96}^{1{,}96} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du\right|$$

$$= \left| -\int_{-1{,}96}^{-1{,}96+O_1\left(\frac{1}{\sqrt{n}}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du + \int_{1{,}96}^{1{,}96+O_2\left(\frac{1}{\sqrt{n}}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \right| = O\left(\frac{1}{\sqrt{n}}\right)$$

Or l'erreur commise quand on approxime la loi binomiale par la loi normale est un $O\left(\frac{1}{\sqrt{n}}\right)$

on peut donc écrire à bon droit :

$$P\left(p \in \left[F_n - \frac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\ ;\ F_n + \frac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\right]\right)$$

$$= P\left(F_n \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} + o_1\left(\frac{1}{\sqrt{n}}\right); p + 1{,}96\sqrt{\frac{p(1-p)}{n}} + o_2\left(\frac{1}{\sqrt{n}}\right)\right]\right)$$

$$\simeq 0{,}95$$

Et si $P\left(p \in \left[F_n - \frac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\ ;\ F_n + \frac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\right]\right) \simeq 0{,}95$,

on en déduit que pour 95 % des échantillons aléatoires de taille *n* possibles, le paramètre inconnu *p* appartient à l'intervalle $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\ ;\ f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right]$.

On traduit cela en disant que $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\ ;\ f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right]$ est un intervalle de confiance de *p* au niveau 95%.

Et on obtient ainsi un intervalle moins grossier que $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$, l'intervalle du programme de seconde.

Mais encore une fois $P\left(p \in \left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\ ;\ f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right]\right)$ **n'existe pas.**

De façon plus théorique :

1. D'après le théorème de Bernoulli, (c.à.d. l'inégalité de Bienaymé-Tchebychev) on sait que :

   F_n converge en probabilité vers p,

   on peut alors montrer que $\sqrt{F_n(1-F_n)}$ converge en probabilité vers $\sqrt{p(1-p)}$.

2. D'après le théorème de Moivre Laplace on sait que :

   $Z_n = \dfrac{F_n - E(F_n)}{\sigma(F_n)} = \sqrt{n}\,\dfrac{F_n - p}{\sqrt{p(1-p)}}$ converge en loi vers X où X suit la loi $N(0\,;1)$

3. Avec un lemme de Slutsky, on peut montrer que :

   $\sqrt{n}\,\dfrac{F_n - p}{\sqrt{p(1-p)}} \times \dfrac{\sqrt{p(1-p)}}{\sqrt{F_n(1-F_n)}}$ converge en loi vers X où X suit la loi $N(0\,;1)$

4. donc $Z_n = \sqrt{n}\,\dfrac{F_n - p}{\sqrt{F_n(1-F_n)}}$ converge en loi vers X où X suit la loi $N(0\,;1)$

   et enfin $\lim\limits_{n\to\infty} P\left(p \in \left[F_n - \dfrac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\ ;\ F_n - \dfrac{1{,}96}{\sqrt{n}}\sqrt{F_n\ 1-F_n}\right]\right) = 0{,}95$

En fait l'intervalle $\left[f - \dfrac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\ f - \dfrac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right]$ est une réalisation de l'intervalle de confiance $\left[F_n - \dfrac{1{,}96}{\sqrt{n}}\sqrt{F_n(1-F_n)}\,;\ F_n + \dfrac{1{,}96}{\sqrt{n}}\sqrt{F_n(1-F_n)}\right]$.

exemples

1. $f = 0{,}2$, $n = 50$

   - avec la méthode du trinôme : $\left(1 + \dfrac{1{,}96^2}{50}\right)p^2 - \left(2\times 0{,}2 + \dfrac{1{,}96^2}{50}\right)p + 0{,}2^2 \leqslant 0$

on obtient l'intervalle de confiance $[0{,}112\,;0{,}331]$

   - $\left[f - \dfrac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\ f + \dfrac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right] = [0{,}089\,;0{,}315]$
   - $\left[f - \dfrac{1}{\sqrt{n}}\,;\ f + \dfrac{1}{\sqrt{n}}\right] = [0{,}059\,;\ 0{,}341]$

Mais ici la condition $np \geqslant 5$ n'est pas vérifiée par toutes les valeurs possibles de p.

2. $f = 0{,}2$, $n = 100$

- avec la méthode du trinôme : $\left(1+\frac{1{,}96^2}{100}\right)p^2 - \left(2\times 0{,}2+\frac{1{,}96^2}{100}\right)p + 0{,}2^2 \leqslant 0$

on obtient l'intervalle de confiance $[0{,}133\,;\,0{,}290]$

- $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\; f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right] = [0{,}122\,;0{,}2784]$
- $\left[f - \frac{1}{\sqrt{n}}\,;\, f + \frac{1}{\sqrt{n}}\right] = [0{,}1;0{,}3]$

3. $f = 0{,}5$, $n = 100$

- avec la méthode du trinôme : $\left(1+\frac{1{,}96^2}{100}\right)p^2 - \left(2\times 0{,}5+\frac{1{,}96^2}{100}\right)p + 0{,}5^2 \leqslant 0$

on obtient l'intervalle de confiance $[0{,}404\,;\,0{,}596]$

- $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\; f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right] = [0{,}402\,;0{,}598]$
- $\left[f - \frac{1}{\sqrt{n}}\,;\, f + \frac{1}{\sqrt{n}}\right] = [0{,}4;0{,}6]$

4. $f = 0{,}2$, $n = 1000$

- avec la méthode du trinôme : $\left(1+\frac{1{,}96^2}{1000}\right)p^2 - \left(2\times 0{,}2+\frac{1{,}96^2}{1000}\right)p + 0{,}2^2 \leqslant 0$

on obtient l'intervalle de confiance $[0{,}176\,;\,0{,}226]$

- $\left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\; f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right] = [0{,}175;0{,}225]$
- $\left[f - \frac{1}{\sqrt{n}}\,;\, f + \frac{1}{\sqrt{n}}\right] = [0{,}168;0{,}231]$

On vérifie donc que ces intervalles sont d'autant plus proches que *n* est grand et/ou f proche de 0,5.

4. Intervalle de fluctuation asymptotique en classe de terminale S

On construit ici l'intervalle de fluctuation asymptotique au seuil 1−α de la variable aléatoire fréquence $\boldsymbol{F_n = \frac{X_1 + X_2 + ... + X_n}{n}}$ qui, à tout échantillon aléatoire $\boldsymbol{X_1, X_2, ..., X_n}$ de taille *n*, associe la fréquence obtenue.

Théorème de Moivre Laplace

Si $(X_n)_n$ est une suite de variables aléatoires indépendantes suivant la même loi de Bernoulli de paramètre p, donc d'espérance p et de variance $\sigma^2 = pq$.

Alors si $F_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{Y_n}{n}$

$Z_n = \frac{F_n - E(F_n)}{\sigma(F_n)} = \frac{F_n - p}{\sqrt{pq/n}}$ ***converge en loi*** vers X où X suit la loi normale $N(0\,;1)$

c'est-à-dire que pour tout $a < b$: $\lim\limits_{n \to +\infty} P(a < Z_n < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$

Pour un risque α donné, $\alpha \in]0;1[$, on détermine w_α tel que $P(p - w_\alpha \leqslant F_n \leqslant p + w_\alpha) = 1 - \alpha$

comme pour $\boldsymbol{n \geqslant 30}$, $\boldsymbol{np \geqslant 5}$ **et** $\boldsymbol{nq \geqslant 5}$,on considère que $Z_n = \frac{F_n - p}{\sqrt{pq/n}}$ suit approximativement la loi $N(0;1)$, et que

$p - w_\alpha \leqslant F_n \leqslant p + w_\alpha \Leftrightarrow \frac{p + w_\alpha - p}{\sqrt{pq/n}} \leqslant \frac{F_n - p}{\sqrt{pq/n}} \leqslant \frac{p + w_\alpha - p}{\sqrt{pq/n}} \Leftrightarrow \frac{-w_\alpha}{\sqrt{pq/n}} \leqslant Z_n \leqslant \frac{w_\alpha}{\sqrt{pq/n}} \Leftrightarrow$

$-u_\alpha \leqslant Z_n \leqslant u_\alpha$

si on pose $u_\alpha = \frac{w_\alpha}{\sqrt{pq/n}}$ **,** on montre que $\exists !\ u_\alpha$ tel que $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1 - \alpha$,

à l'aide de la calculatrice , on détermine u_α tel que $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1 - \alpha$ et $w_\alpha = u_\alpha \sqrt{\frac{pq}{n}}$

remarque

Les élèves doivent connaitre $u_{0,05} = 1{,}96$ et $u_{0,01} = 2{,}58$

Et puisque $P\left(p - u_\alpha \sqrt{pq/n} \leqslant F_n \leqslant p + u_\alpha \sqrt{pq/n}\right) = P\left(F_n \in \left[p - u_\alpha \sqrt{pq/n}\ ;\ p + u_\alpha \sqrt{pq/n}\right]\right) = 1 - \alpha$ **,**

on peut énoncer :

définition si $n \geqslant 30$, $np \geqslant 5$ et $nq \geqslant 5$

$\left[p - u_\alpha \sqrt{pq/n} \; ; p + u_\alpha \sqrt{pq/n} \right]$ est l'intervalle de fluctuation asymptotique au seuil $1-\alpha$ de F_n

exemple

$\left[p - 1{,}96\sqrt{pq/n} ; p + 1{,}96\sqrt{pq/n} \right]$ est l'intervalle de fluctuation au seuil 0,95 de F_n

signification

"sensiblement" 95 % des échantillons aléatoires de taille *n* fournissent une fréquence f appartenant à l'intervalle : $\left[p - 1{,}96\sqrt{\frac{pq}{n}} ; p + 1{,}96\sqrt{\frac{pq}{n}} \right]$

Dans le cours de terminale S , on démontre d'abord le théorème 1 :

Théorème 1
Si X **est une variable aléatoire suivant la loi normale** N **(0,1)**
alors, $\forall \alpha \in]0;1[$ **, il existe un unique réel** u_α **tel que** $P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1-\alpha$

Démonstration (faisant partie des exigibles en terminale S).

Tout d'abord $\alpha \in]0;1[\Rightarrow 1-\alpha \in]0;1[$

D'après la symétrie de la courbe de f la fonction de densité de la loi normale , on a pour tout réel *u* positif,

$$P(-u_\alpha \leqslant X \leqslant u_\alpha) = 2P(0 \leqslant X \leqslant u_\alpha) = 2\int_0^{u_\alpha} f(x)dx = 2F(u_\alpha)$$

où F est la primitive de f sur $]0;+\infty[$, qui existe, **puisque** f est continue sur $\mathbb{R}$ **,** qui s'annule en 0.

1. La fonction $2F$ est donc continue , **puisque** dérivable ,
2. La fonction $2F$ strictement croissante sur $]0;+\infty[$, **puisque**
   $F'(x) = f(x) > 0 \quad \forall x \in]0;+\infty[$,
3. $1-\alpha \in \left] 2F(0); \lim\limits_{x\to\infty} 2F(x) \right[=]0;1[$ **car** $F(0) = 0$ et $\lim\limits_{x\to\infty} F(x) = \frac{1}{2}$ puisque c'est la moitié de l'aire sous la courbe de f

donc d'après le corollaire du théorème des valeurs intermédiaires, l'équation $2F(x) = 1-\alpha$ à une solution unique u_α dans $]0;+\infty[$,

et on a $2F(u_\alpha) = P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1-\alpha$.

puis le théorème 2 :

Théorème 2

Si une variable aléatoire Y_n suit la loi B (n, p), avec p dans l'intervalle]0, 1[,

alors $\forall \alpha \in]0;1[$, $\lim\limits_{n\to\infty} P(F_n \in I_n) = 1-\alpha$ **si** $F_n = \dfrac{Y_n}{n}$ **et**

$I_n = \left[p - u_\alpha \sqrt{pq/n} \; ; p + u_\alpha \sqrt{pq/n} \right]$

et où u_α est l'unique réel tel que $P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1-\alpha$ où X suit la loi normale $N(0\,;1)$

Démonstration (exigible en terminale S **)**

D'après le théorème de Moivre-Laplace, si $Z_n = \dfrac{F_n - p}{\sqrt{pq/n}} = \dfrac{Y_n - np}{\sqrt{npq}}$,

$$\lim_{n\to\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P(-u_\alpha \leqslant Z \leqslant u_\alpha)$$

or $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P\left(p - u_\alpha \sqrt{pq/n} \leqslant \dfrac{Y_n}{n} \leqslant p + u_\alpha \sqrt{pq/n} \right) = P\left(np - u_\alpha \sqrt{npq} \leqslant Y_n \leqslant np + u_\alpha \sqrt{npq} \right)$

signalons que la suite $P\ F_n \in I_n$ n'est pas monotone.

Et on peut alors définir l'intervalle de fluctuation asymptotique au seuil $1-\alpha$**.**

Remarque

Dans le cas $\alpha = 0{,}05$, donc $u_\alpha = 1{,}96$, on retrouve l'intervalle de fluctuation du programme de seconde en majorant *pq* par 1/4 et 1,96 par 2, en effet:

le maximum de la fonction $p \longrightarrow p(1-p) = \dfrac{1}{4} - \left(p - \dfrac{1}{2} \right)^2$ sur $[0;1]$ est $\dfrac{1}{4}$

$$\left[p - 1{,}96\sqrt{\frac{pq}{n}} ; p + 1{,}96\sqrt{\frac{pq}{n}} \right] = \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} ; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right]$$

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

ainsi $\left[p - \dfrac{1}{\sqrt{n}} ; p + \dfrac{1}{\sqrt{n}} \right]$ contiendra a fortiori sensiblement 95 % des fréquences observées des échantillons de taille n **;** propriété que l'on peut préciser par le théorème :

Théorème

Si la variable aléatoire suit la loi B (n, p) alors, pour tout p dans]0, 1[,

il existe un entier n_0 tel que $n \geqslant n_0 \Rightarrow P\left(p - \dfrac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \dfrac{1}{\sqrt{n}} \right) \geqslant 0{,}95$

Ce qui ne signifie pas que 95% des fréquences observées seront nécessairement dans l'intervalle, même si cela est fortement" probable" pour un nombre d'échantillons assez grand (loi des grands nombres).

Démonstration (non exigible mais mentionnée comme "intéressante" dans le programme)

$$P\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right)\geqslant P\left(p-2\sqrt{pq/n}\leqslant F_n\leqslant p+2\sqrt{pq/n}\right)$$

on sait que $\lim\limits_{n\to\infty} P\left(p-u_\alpha\sqrt{pq/n}\leqslant F_n\leqslant p+u_\alpha\sqrt{pq/n}\right)=1-\alpha$

or si $u_\alpha=2$ alors $\alpha=0{,}0456$ car $P(-2\leqslant X\leqslant 2)=0{,}9544$ où X suit la loi normale $N(0\,;1)$

donc $\lim\limits_{n\to\infty} P\left(p-2\sqrt{pq/n}\leqslant F_n\leqslant p+2\sqrt{pq/n}\right)=0{,}9544$

et par définition de la limite, l'intervalle ouvert $]0{,}9504\,;\,0{,}9584[$ contenant la limite 0,9544, contient tous les termes de la suite $\left(P\left(p-2\sqrt{pq/n}\leqslant F_n\leqslant p+2\sqrt{pq/n}\right)\right)$ à partir d'un certain rang n_0 ,

donc comme $\left[p-1{,}96\sqrt{pq/n}\,;p+1{,}96\sqrt{pq/n}\,\right]\subset\left[p-\sqrt{1/n}\;\;;\;\;p+\sqrt{1/n}\,\right]$

$n\geqslant n_0\;\Rightarrow\; P\left(p-\sqrt{1/n}\leqslant F_n\leqslant p+\sqrt{1/n}\right)\geqslant P\left(p-2\sqrt{pq/n}\leqslant F_n\leqslant p+2\sqrt{pq/n}\right)\geqslant 0{,}9504$.

Dans l'accompagnement du programme, on précise :

la valeur de n_0 varie avec la valeur de p
Il est difficile de déterminer cette valeur de n_0. On peut cependant donner des valeurs de n_0 grâce à un algorithme.

P	0,35	0,36	0,37	0,38	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5
n_0	31	30	36	64	56	81	90	120	143	209	271	288	304	399	399	529

On peut remarquer que la plus grande valeur de n_0 est atteinte pour p =1/2 . C'est effectivement pour cette valeur que la fluctuation est la plus importante puisque la variance est maximale et donc la dispersion.

Remarque : Taille minimale de l'échantillon pour avoir une précision donnée

De manière générale, au seuil α donné , si a est l'amplitude de l'intervalle de fluctuation on résout $2u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant a$ donc $n \geqslant \dfrac{4u_\alpha^2 p(1-p)}{a^2}$.

Dans le programme, la capacité attendue est: "déterminer une taille d'échantillon suffisante pour obtenir, avec une précision donnée, une estimation d'une proportion au niveau de confiance 0,95" et ne concerne donc que l'intervalle de confiance $\left[f - \dfrac{1}{\sqrt{n}} ; f + \dfrac{1}{\sqrt{n}} \right]$.

au niveau de confiance $0{,}95$, $2\dfrac{1}{\sqrt{n}} \leqslant a \Rightarrow n \geqslant \dfrac{4}{a^2}$.

Remarque

On sait que si Y suit la loi normale $N\left(\mu ; \sigma^2\right)$, alors

$P\ \mu - \sigma \leqslant Y \leqslant \mu + \sigma \simeq 0{,}683$ à 10^{-3} prés
$P\ \mu - 2\sigma \leqslant Y \leqslant \mu + 2\sigma \simeq 0{,}954$ à 10^{-3} prés
$P\ \mu - 3\sigma \leqslant Y \leqslant \mu + 3\sigma \simeq 0{,}997$ à 10^{-3} prés

comme $Y_n = X_1 + X_2 + ... + X_n$ ***suit approximativement*** la loi normale $N\left(np ; \sqrt{npq}\right)$, on a donc de façon immédiate 3 intervalles de fluctuation :

$\mu - \sigma, \mu + \sigma$ au seuil $\alpha = 1 - 0{,}683 = 0{,}317$

(peut intéressant, car le risque est alors trop grand))

$\mu - 2\sigma, \mu + 2\sigma$ au seuil $\alpha = 1 - 0{,}954 = 0{,}046$

$\mu - 3\sigma, \mu + 3\sigma$ au seuil $\alpha = 1 - 0{,}997 = 0{,}003$

5. Intervalle de confiance à partir de la loi binomiale

Quoique hors programme (heureusement), ce qui suit peut aider à comprendre la notion d'intervalle de confiance.

Rappelons comment on obtient l'intervalle de fluctuation de F_n à partir de la loi binomiale :

On sait que $Y_n = X_1 + X_2 + ... + X_n = nF_n$ suit la loi binomiale $B\ n, p$,

on détermine a et b tels que : $P(a \leqslant Y_n \leqslant b) \geqslant 0{,}95$ où

a est le plus grand entier tel que $P(Y_n \leqslant a) > 0{,}025$

b est le plus grand entier tel que $P(Y_n \leqslant b) \geqslant 0{,}975$

donc aussi $P(a+1 \leqslant Y_n \leqslant b) < 0{,}95$ et $P(a \leqslant Y_n \leqslant b-1) < 0{,}95$

et ainsi $P\ Y_n \in\ a\,;b\ \simeq 0{,}95$ (en fait $\geqslant 0{,}95$)

que l'on pourrait écrire plus précisément $P\ Y_n \in \left[\ a_n\ \ p\ ;b_n\ \ p\ \right] \simeq 0{,}95$

et enfin l'intervalle $\left[\ \frac{a}{n}\ ;\ \frac{b}{n}\ \right]$ est l'intervalle de fluctuation à (au moins) 0,95 de F_n (à partir de la loi binomiale, qui est la **vraie loi** de Y_n **).** soit $P\left(F_n \in \left[\ \frac{a}{n}\ ;\ \frac{b}{n}\ \right]\right) \simeq 0{,}95$

La démarche pour obtenir un intervalle de confiance est la suivante :

Partant de la fréquence observée f d'un échantillon, on recherche ce que l'on peut en déduire ou induire sur p qui est donc ici une inconnue.

Une possibilité est de chercher l'ensemble des valeurs de p pour lesquelles $nf \in \left[\ a_n(p)\,;b_n(p)\ \right]$ **,** puisqu'alors, avant tirage, nous serons dans le cas où $P\ Y_n \in \left[\ a_n\ \ p\ ;b_n\ \ p\ \right] \geqslant 0{,}95$.

Cet ensemble est un intervalle $\left[\ p_1\,;p_2\ \right]$ puisque p est un réel et que les fonctions $a_n(p)$ et $b_n(p)$ sont croissantes .

La détermination n'est pas aisée, car les relations fonctionnelles ne sont pas explicites (il faudrait écrire un programme informatique)

Voyons sur un exemple :

Supposons que pour un échantillon de taille $n = 50$, la fréquence observée f est 0,20, alors $nf = 10$. On cherche les valeurs de p avec trois chiffres significatifs.

Avec un tableur, on obtient les distributions des lois binomiales $B\ 50\,;\ p$,

puis on localise les valeurs de p , la plus petite et la plus grande, qui conviennent,

c'est à dire telles que

$$10 \geqslant a_{50}(p) \text{ et } 10 \leqslant b_{50}(p)$$

($a_{50}(p)$ et $b_{50}(p)$ sont croissantes **)**

k\p	0,100	0,101	0,102
0	0,00515378	0,00487511	0,00461123
1	0,03378586	0,03226033	0,03079973
2	0,11172876	0,10763824	0,10367843
3	0,25029391	0,24313399	0,23612613
4	0,43119841	0,42199895	0,41289513
5	0,61612301	0,60687249	0,59761677
6	0,77022684	0,76264747	0,75497986
7	0,87785492	0,87265288	0,867332
8	0,94213279	0,93908134	0,9359256
9	0,97546206	0,97390886	0,9722848
10	0,9906454	0,98995122	0,98921733
11	0,99678008	0,99650508	0,99621113
12	0,99899538	0,99889808	0,99879292

pour $p = 0{,}100$
$P(Y_n \leqslant 9) = 0{,}9754...$ et $P(Y_n \leqslant 8) = 0{,}942...$
donc $b = 9$

pour $p = 0{,}101$
$P(Y_n \leqslant 9) = 0{,}9739...$ et $P(Y_n \leqslant 10) = 0{,}989...$
donc $b = 10$

et pour $p = 0{,}102$
$P(Y_n \leqslant 9) = 0{,}972...$ et $P(Y_n \leqslant 10) = 0{,}989...$
donc $b = 10$

on choisit donc $p = 0{,}101$

k\p	0,314	0,315	0,316
0	6,5495E-09	6,0888E-09	5,6599E-09
1	1,5644E-07	1,4609E-07	1,364E-07
2	1,8374E-06	1,7234E-06	1,6162E-06
3	1,4148E-05	1,3328E-05	1,2555E-05
4	8,0358E-05	7,6033E-05	7,1932E-05
5	0,00035917	0,00034132	0,0003243
6	0,00131633	0,00125625	0,00119875
7	0,00407019	0,00390089	0,00373809
8	0,01084545	0,01043768	0,01004372
9	0,02531778	0,02446554	0,02363834
10	0,05247767	0,05091368	0,0493886
11	0,09768415	0,09514017	0,09264795
12	0,16493373	0,16123779	0,15760021

pour $p = 0{,}314$
$P(Y_n \leqslant 9) = 0{,}0253...$ et $P(Y_n \leqslant 8) = 0{,}010...$
donc $a = 9$

pour $p = 0{,}315$
$P(Y_n \leqslant 9) = 0{,}0244...$ et $P(Y_n \leqslant 10) = 0{,}0509...$
donc $a = 10$

pour $p = 0{,}316$
$P(Y_n \leqslant 9) = 0{,}0236...$ et $P(Y_n \leqslant 8) = 0{,}049...$
donc $a = 10$

on choisit donc $p = 0{,}315$

l'intervalle de confiance est donc $[0{,}101; 0{,}315]$ à un niveau supérieur à 0,95

remarquons que:

- pour $p = 0{,}101$, l'intervalle de fluctuation est $\left[\frac{1}{50}; \frac{10}{50}\right] = [0{,}02; 0{,}2]$ au niveau 0,99

ou $\left[\frac{2}{50};\frac{9}{50}\right]=[0{,}02\,;\,0{,}18]$ au niveau 0,942

(la borne inférieure de l'intervalle peut donc être sujette à discussion)

- pour $p=0{,}315$, l'intervalle de fluctuation est $\left[\frac{10}{50};\frac{21}{50}\right]=[0{,}2\,;\,0{,}42]$ au niveau 0,954
- l'intervalle de confiance du programme est $\left[0{,}2-\frac{1}{\sqrt{50}}\,;0{,}2+\frac{1}{\sqrt{50}}\right]=[0{,}0586\,;0{,}342\,]$ au niveau 0,95 .

Interprétation

C'est un intervalle de confiance au niveau 0,95 car il est obtenu à partir d'intervalles de fluctuation au niveau (supérieur à) 0,95 :

$\forall p\in\ 0\,;1$ et $\forall n\in\mathbb{N}^*$, $a_n\ \ p$ et $b_n\ \ p$ sont tels que $P\left(F_n\in\left[\frac{a_n\ \ p}{n};\frac{b_n\ \ p}{n}\right]\right)\simeq 0{,}95$

p_1 et p_2 sont tels que $\forall p\in[p_1\,;p_2\,]\quad f\in\left[\frac{a_n(p)}{n};\frac{b_n(p)}{n}\right]$.

mais encore une fois , $P\left(f\in\left[\frac{a_n\ \ p}{n};\frac{b_n\ \ p}{n}\right]\right)\simeq 0{,}95$ **n'a pas de sens!**

il n'y a pas de variable aléatoire dans les parenthèses, donc ici P n'existe pas.

On sait seulement que, avant tirage, $\forall p\in[p_1\,;p_2\,]$, $P\left(F_n\in\left[\frac{a_n\ \ p}{n};\frac{b_n\ \ p}{n}\right]\right)\simeq 0{,}95$.

remarque

Revenons sur le paragraphe 2 et en adoptant la même démarche que ci-dessus, on cherche donc les valeurs de p pour lesquelles $f\in\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$, au niveau de confiance 0,95. La réponse est alors immédiate : $p\in\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$, puisque

$$P\left(p\in\left[F_n-\frac{1}{\sqrt{n}}\,;\,F_n+\frac{1}{\sqrt{n}}\right]\right)=P\left(F_n\in\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]\right)\simeq 0{,}95.$$

.

Comparaison de deux fréquences

On souhaite comparer les proportions p_1 et p_2 d'un même caractère, dans deux populations distinctes, à partir des fréquences f_1 et f_2 observées sur un échantillon de chacune des deux populations.

Si on a fait l'hypothèse que les proportions sont égales, la question posée est de savoir si la différence $f_1 - f_2$ est significative, ou due aux fluctuations d'échantillonnage.

1. Pour y répondre, on fait selon le programme la chose suivante en terminale:

à partir de chaque échantillon, si $n \geqslant 30$, $nf \geqslant 5$ et $n(1-f) \geqslant 5$, on détermine un intervalle de confiance,

$$\left[f_1 - \frac{1{,}96}{\sqrt{n_1}}\sqrt{f_1(1-f_1)} \; ; \; f_1 + \frac{1{,}96}{\sqrt{n_1}}\sqrt{f_1(1-f_1)} \right] \quad \text{et} \quad \left[f_2 - \frac{1{,}96}{\sqrt{n_2}}\sqrt{f_2(1-f_2)} \; ; \; f_2 + \frac{1{,}96}{\sqrt{n_2}}\sqrt{f_2(1-f_2)} \right].$$

S'ils sont disjoints , on a alors $|f_1 - f_1| > 1{,}96\left(\sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}}\right)$, on conclut que la différence $f_1 - f_2$ est significative, en d'autres termes, on rejette l'hypothèse $p_1 = p_2$.

Sinon on ne peut rien dire.

2. Ce qui justifie cette règle de décision (et qui est bien sûr hors programme)

Construction classique du test (G. Saporta, p 348):

Notons F_1 et F_2 les fréquences aléatoires des deux échantillons aléatoires indépendants de tailles n_1 et n_2.

la variable aléatoire $F_1 - F_2$ suit alors approximativement la loi normale

$$N\left(\mu = p_1 - p_2 \; ; \sigma^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

l'intervalle de fluctuation de la variable $F_1 - F_2$ au seuil de 5% est

$$\left[p_1 - p_2 - 1{,}96\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \; ; \; p_1 - p_2 + 1{,}96\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \right]$$

sous l'hypothèse $p_1 = p_2 = p$, il devient

$$\left[-1{,}96\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} \; ; \; 1{,}96\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} \right]$$

sous l'hypothèse $p_1 = p_2 = p$, on estime p par $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$,

et on rejette donc l'hypothèse si $|f_1 - f_2| > 1{,}96\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$

On peut montrer que $\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} \leqslant \sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}}$ mais sous certaines conditions .

Le critère du programme est donc plus "sévère" que celui construit avec la méthode ci-dessus lorsqu'il s'agit de rejeter l'hypothèse $p_1 = p_2$.

car $|f_1 - f_2| > 1{,}96\left(\sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}} \right) \Rightarrow |f_1 - f_2| > 1{,}96\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$

L'inégalité est vraie dés que les fréquences observées ne sont pas trop éloignées et que la taille des deux échantillons est du même ordre de grandeur .

Remarques

1. Si on utilise les intervalles de confiance du programme de seconde,

comme $1{,}96\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} \leqslant 1{,}96\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}} \leqslant \frac{1{,}96}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leqslant \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leqslant \sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_1}}$

alors $|f_1 - f_2| > \sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_2}} \quad \Rightarrow |f_1 - f_2| > 1{,}96\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$.

Ainsi, si les intervalles de confiance du programme de seconde sont disjoints, il est donc légitime de rejeter l'hypothèse $p_1 = p_2$.

2. Dans l'accompagnement du programme de terminale on construit un autre test :

La variable aléatoire F_1 suit approximativement la loi normale $N\left(\mu = p_1 ; \sigma^2 = \frac{p_1 q_1}{n_1}\right)$

et on estime σ par $\sqrt{\frac{f_1(1-f_1)}{n_1}}$.

La variable aléatoire F_2 suit approximativement la loi normale $N\left(\mu = p_2 ; \sigma^2 = \frac{p_2 q_2}{n_2}\right)$

et on estime σ par $\sqrt{\frac{f_2(1-f_2)}{n_2}}$.

Sous l'hypothèse $p_1 = p_2$, la variable aléatoire $F_1 - F_2$ suit alors approximativement la loi normale $N\left(\mu = 0 ; \sigma^2 = \frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}\right)$

il est clair alors par concavité de la fonction racine que

$$|f_1 - f_2| > 1{,}96\left(\sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}}\right) \Rightarrow |f_1 - f_2| > 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}\ .$$

Ainsi, si les intervalles de confiance du programme de terminale STL et STI2D sont disjoints, il est donc légitime de rejeter l'hypothèse $p_1 = p_2$.

Exemple

$f_1 = 0{,}16$, $n_1 = 200$, $f_2 = 0{,}25$ et $n_2 = 150$

Les intervalles de confiance du programme de terminale STL et STI2D,

$$\left[f_1 - \frac{1{,}96}{\sqrt{n_1}}\sqrt{f_1(1-f_1)}\ ;\ f_1 + \frac{1{,}96}{\sqrt{n_1}}\sqrt{f_1(1-f_1)}\right] \text{ et } \left[f_2 - \frac{1{,}96}{\sqrt{n_2}}\sqrt{f_2(1-f_2)}\ ;\ f_2 + \frac{1{,}96}{\sqrt{n_2}}\sqrt{f_2(1-f_2)}\right]$$

sont $[0{,}1091\,;\,0{,}2108]$ et $[0{,}1807\,;\,0{,}3193]$, donc on ne rejette pas l'hypothèse.

Les intervalles de confiance du programme de seconde utilisés dans les autres terminales,

$\left[f_1 - \frac{1}{\sqrt{n_1}} \,;\, f_1 + \frac{1}{\sqrt{n_1}} \right]$ et $\left[f_2 - \frac{1}{\sqrt{n_2}} \,;\, f_2 + \frac{1}{\sqrt{n_2}} \right]$ sont $[0{,}0893 \,;\, 0{,}2307]$ et $[0{,}1684 \,;\, 0{,}3316]$

donc on ne rejette pas l'hypothèse.

Si on utilise la construction classique du test

L'intervalle de fluctuation de la variable $F_1 - F_2$, sous l'hypothèse $p_1 = p_2$ au seuil de 5% est

$\left[-1{,}96\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} \,;\, 1{,}96\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} \right]$, soit $[-0{,}0844 \,;\, 0{,}0844]$.

Or $f_1 - f_2 = -0{,}09$, donc on rejette l'hypothèse.

L'intervalle de fluctuation de la variable $F_1 - F_2$ tel que construit dans l'accompagnement est

$\left[-1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}} \,;\, 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}} \right]$, soit $[-0{,}0859 \,;\, 0{,}0859]$.

Or $f_1 - f_2 = -0{,}09$, on rejette encore l'hypothèse.

Conclusion

Le test du programme est fait pour rejeter. Donc, quand on fabrique un exercice, il est préférable de s'assurer que les intervalles seront disjoints.

Bibliographie

Les documents d'accompagnement des programmes du Lycée.

Jean Bass , cours de mathématiques, Editions Masson 1968.

Gilbert Saporta, Probabilités, Analyses des données et Statistique, Editions Technip 1990.

Jean-Yves Ouvrard, Probabilité tomes 1 et 2, Edition Cassini 1998

Paul Deheuvels, l'intégrale, Puf 1980

Charles Suquet, université de Lille1 : http://math.univ-lille1.fr/~suquet/

Cours Jussieu : http://www.proba.jussieu.fr/supports.php

Titre : **Probabilités et Statistique au Lycée**

Auteurs : **Noël BASCOU, Daniel BRESSON, Françoise DELATOUR, Christian LAVERGNE, Jean-Marie SCHADECK**

Niveau : **Lycée**

Date : **Janvier 2016**

Public concerné : Professeurs de mathématiques de Lycée

Mots clés : **Statistique descriptive, probabilités, statistique inférentielle, loi binomiale, loi normale, théorème central limite.**

Résumé : Ce document a été réalisé par le groupe IREM Probabilités et Statistique de Montpellier.

Il est destiné aux enseignants de lycée général, technologique et professionnel. Pour sa conception, nous avons analysé les programmes et leurs différents documents d'accompagnement. Pour répondre aux questionnements de nos collègues, nous proposons ici quelques compléments.

À partir des notions abordées dans les programmes officiels, nous reprenons les fondamentaux de la statistique descriptive, des probabilités et de la statistique inférentielle.

Le document est composé de trois sections.

La première aborde les indices de la statistique univariée, à savoir les quantiles, et les diverses notions de dispersion et de résumés statistiques associés.

La deuxième section aborde la loi binomiale, la loi normale, le théorème de Moivre Laplace, le théorème central limite avec en annexe une preuve de la formule de Stirling et un aperçu sur l'intégrale de Lebesgue.

La troisième section reprend les bases de la statistique inférentielle utiles pour l'enseignement de la statistique au lycée.