STATISTIQUES A UNE VARIABLE

1) Introduction et vocabulaire

La statistique est la science qui consiste à réunir des données chiffrées, à les analyser, à les commenter et à les critiquer
Une étude statistique s'effectue sur un ensemble appelé **Population**, dont les éléments sont appelés **Individus**, et consiste à observer et étudier un même aspect sur chaque individu, appelé **Caractère.**

On distingue deux types de caractère :
- Les caractères **qualitatifs** : Ce sont les caractères dont les valeurs ne sont pas des nombres (profession, couleur des yeux)
- Les caractères **quantitatifs** : Ce sont les caractères qui prennent des valeurs numériques
  - Le caractère quantitatif est **discret** si les valeurs du caractère sont isolées (ex : nombre d'enfants). Ces valeurs sont appelées **modalités**
  - Le caractère est **continu** si les valeurs du caractère sont regroupées en intervalles, appelés **Classes** (ex : Taille $\in [170;175[$) La « largeur » de chaque intervalle s'appelle **l'amplitude**

2) Effectifs et fréquences

On appelle **effectif** d'une valeur (respectivement d'une classe, respectivement d'une modalité) le nombre d'individus possédant le caractère de cette valeur (respectivement d'une classe, respectivement d'une modalité)

On appelle **fréquence** d'une valeur (respectivement d'une classe, respectivement d'une modalité) le quotient de l'effectif de cette valeur par l'effectif total de la population

Les fréquences sont des nombres compris entre 0 et 1, souvent exprimées en pourcentage

$$\text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}} \underbrace{\times 100}_{\substack{\text{pour obtenir} \\ \text{un pourcentage}}}$$

Effectifs et fréquences cumulé(e)s croissant(e)s et/ou décroissant(e)s

Dans le cas d'une variable quantitative, on peut ordonner les différentes valeurs de la variable dans l'ordre croissant ou décroissant.

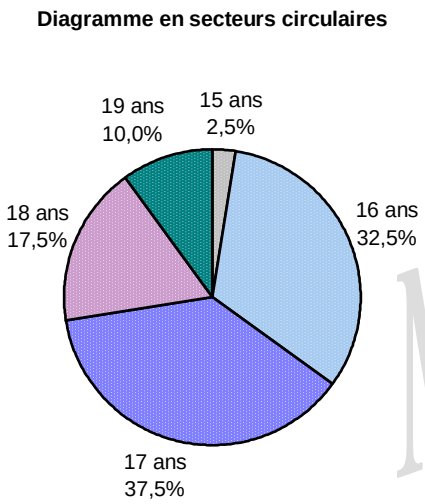
On peut ainsi déterminer "Quel effectif ou quelle fréquence de la population a une valeur du caractère au plus égale ou au moins égale à …."
Ce sont les notions **d'effectifs cumulés** croissants ou décroissants, ou de **fréquences cumulées** croissantes ou décroissantes

3) Les représentations graphiques

On peut visualiser la série statistique par le biais d'autres moyens, notamment :

Séries statistiques à caractère qualitatifs

On utilise souvent les **des diagrammes à secteurs :**



Les aires des secteurs sont proportionnelles aux effectifs ou aux fréquences
Les angles des secteurs sont proportionnels aux effectifs ou aux fréquences selon le tableau de proportionnalité :

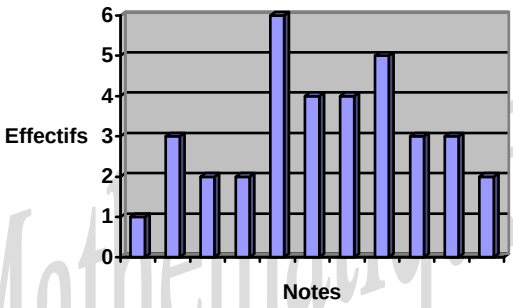
Effectif total	360 °
Effectif de la valeur	Angle du secteur

(Attention ! pour un diagramme semi-circulaire, l'effectif total correspond à un angle de 180°)

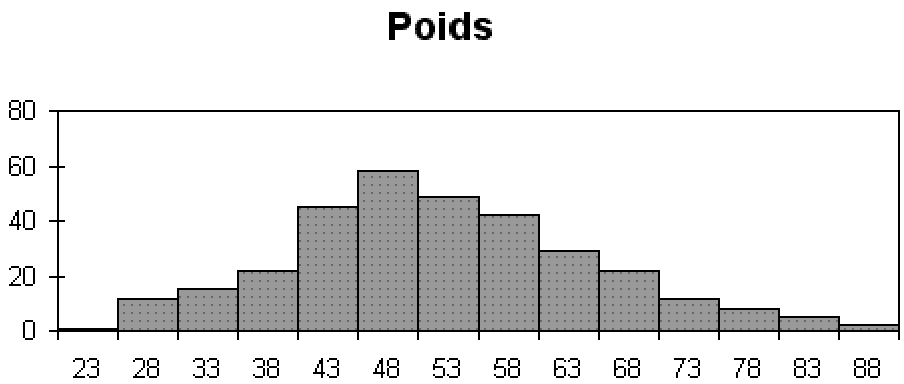
jgcuaz@hotmail.com

Séries statistiques à caractère quantitatifs
On utilise principalement deux types de représentations :

Pour les caractères discrets, on peut utiliser les diagrammes "en bâtons". Ainsi apparaît la discontinuité entre 2 valeurs de la variable ;



Pour les caractères continus, regroupés en intervalles, on peut utiliser un "histogramme".



Dans les deux types de représentation graphique, le caractère est porté en abscisses et l'effectif ou la fréquence sont portés en ordonnée.

Signalons un cas particulier : Histogramme à classe d'amplitudes inégales
Si les amplitudes des classes ne sont pas égales et alors ce sont les aires des rectangles qui doivent être proportionnelles aux effectifs des classes.
Sur l'axe des abscisses, on représente les classes. On ne doit pas représenter des classes d'amplitudes différentes avec une base identique. Si l'amplitude est double, la base doit être double. On se ramène à la plus petite amplitude appelée **amplitude élémentaire.**
En pratique, pour la construction de ces rectangles on procède de la manière suivante :
On cherche la classe d'amplitude élémentaire (ou on en choisit une si il y en a plusieurs) puis on choisit la hauteur du rectangle. Cette hauteur sert de base pour les hauteurs suivantes. Puis, pour les autres classes, la largeur du rectangle vaut l'amplitude de la classe (proportionnellement à l'amplitude de la classe choisie pour son amplitude élémentaire) et la hauteur du rectangle vaut:

$$\text{Effectif de la classe} \times \frac{\text{amplitude élémentaire}}{\text{amplitude de la classe}}$$

4) Etude des séries statistiques à une variable

Caractères de répartition

La vue d'un tableau ou d'un graphique ne permet pas forcément de connaître suffisamment des données pour pouvoir en analyser les répartitions, d'autant que la consultation de tableaux peut s'avérer très longue. On cherche alors à résumer celle-ci par une caractéristique de tendance centrale, c'est à dire par un seul nombre destiné à caractériser l'ensemble d'une façon objective et impersonnelle.

4-1) La moyenne arithmétique

La moyenne arithmétique d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par leur nombre. On la note $\bar{x}$

Exemple : Un élève qui a eu comme notes 4,5,7,9 et 12 a une moyenne égale à : $\bar{x} = \frac{4+5+7+9+12}{5} = 7{,}4$

Inconvénient Le calcul peut s'avérer très lourd lors de l'énumération d'un grand nombre de données.

jgcuaz@hotmail.com

4-2) La moyenne pondérée

Exemple :

Si, dans une classe, 4 élèves ont obtenu la note 8, 3 élèves ont obtenu la note 10 et 5 élèves ont obtenu la note12

On ne va pas calculer $\bar{x} = \frac{8+8+8+8+10+10+10+10.....}{12}$, mais on va effectuer $\boxed{\bar{x} = \frac{4\times 8+3\times 10+5\times 12}{12}}$

De manière générale :

Définition :

Si pour une population donnée, on a p valeurs du caractère $x_1,x_2,.....,x_p$ d'effectifs respectifs $n_1,n_2,.....,n_p$ alors la moyenne de cette série statistique est donnée par $\bar{x} = \frac{n_1\times x_1 + n_2\times x_2 + ... + n_p\times x_p}{n_1+n_2+...+n_p}$

4-3) Cas d'une variable continue

Pour calculer la moyenne d'une série statistique à caractère continu, **on remplace chaque classe par son milieu**, avec la part d'approximation que cela comporte.

4-4) Propriété de la moyenne

Propriété

Soit deux séries statistiques S_1 et S_2 d'effectifs totaux respectifs N_1 et N_2 et de moyennes respectives $\overline{x_1}$ et $\overline{x_2}$

Alors la moyenne de la série S obtenue en regroupant S_1 et S_2 est donnée par :

$\bar{x} = \frac{N_1\overline{x_1} + N_2\overline{x_2}}{N_1+N_2}$

Exemple :

Dans une classe de 22 élèves, il y a 4 filles et 18 garçons. Lors d'un devoir, les 4 filles obtiennent 13,6875 de moyenne et les 18 garçons 13,0833. La moyenne de la classe est donc $\bar{x} = \frac{4\times 13{,}6875 + 18\times 13{,}083}{4+18} = 13{,}19$

Définition : On dit que S_1 et S_2 sont des sous-séries statistiques (ou séries statistiques extraites) de S.

Propriété

Soit S une série statistique, de valeurs du caractère notées x_i affectées des coefficients ou effectifs n_i, et de moyenne $\bar{x}$

Soit a et b deux réels quelconques

Alors la série S', de valeurs du caractère $ax_i + b$ affectées des mêmes coefficients ou effectifs n_i, a pour moyenne $a\bar{x} + b$

Exemple :

Dans une classe, 4 élèves ont obtenu la note 8, 3 élèves ont obtenu la note 10 et 5 élèves ont obtenu la note12. La moyenne est donc $\bar{x} = \frac{4\times 8+3\times 10+5\times 12}{12}$

Si l'enseignant décide de transformer les notes sur 40, et de les augmenter de un point (sur 40), la moyenne de la nouvelle série statistique sera $2\bar{x}+1$

4-5) Le mode ou la valeur modale

Définition:

Le mode ou valeur modale est la valeur du carcatère que la variable statistique prend le plus fréquemment.

Si les données sont groupées en classes, on parle plutôt de **classe modale**

jgcuaz@hotmail.com

5) Médianes et quartiles

Définition:
La médiane d'une série statistique est la valeur du caractère qui partage l'effectif total en deux parties égales, c'est à dire telle qu'il y ait autant d'observations ayant une valeur supérieure ou égale à la médiane que d'observations ayant une valeur inférieure ou égale à la médiane.

Exemple :

Un groupe d'élève a obtenu les notes suivantes : 6,7,8,9 et 20 . Leur moyenne est donc $\bar{x} = \frac{6+7+8+9+20}{5} = 10$

Cette moyenne n'est pas très représentative de la répartition des notes, car tous les élèves sauf un, ont une note strictement inférieure à 10. La note médiane est égale à 8 :
Il y a autant d'élèves qui ont 8 ou plus que d'élèves qui ont 8 ou moins.

Cas où le nombre d'observations est pair :
Si le groupe obtient 2,3,14,14,18,18,20,20. Là encore $\bar{x} = 13{,}625$ n'est pas très représentatif

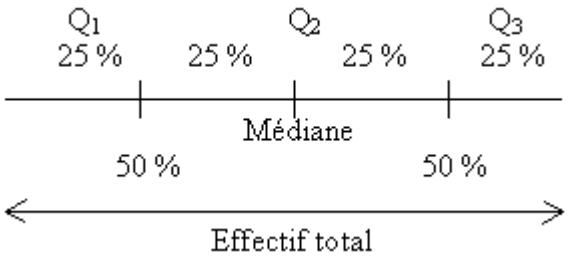
La note médiane est égale, par convention, à la moyenne arithmétique des $4^{ème}$ et $5^{ème}$ notes, soit $\frac{14+18}{2} = 16$ .Il y a autant d'élèves qui ont plus de 16 que d'élèves qui ont moins de 16.

Alors, si *n* est impair, $n = 2p+1$ alors la médiane correspond à la $p+1^{ème}$ valeur.
Si *n* est pair, $n=2p$ et la médiane correspond alors à la moyenne arithmétique entre la $p^{ème}$ et la $p+1^{ème}$ ème valeur

Cas d'une variable continue
Si le caractère est continu, on va déterminer la valeur du caractère correspondant à la fréquence cumulée 50% (ou à l'effectif cumulé de $\frac{n}{2}$), en utilisant le tableau ou l'histogramme des effectifs ou fréquences cumulé(e)s et en effectuant une interpolation linéaire

Les quartiles, déciles et centiles

Définition :
Les quartiles sont les valeurs du caractère qui partagent l'effectif total en 4 parties égales. Plus précisément :
Le quartile Q_1 est la plus petite valeur du caractère pour laquelle 25 % des valeurs de la série statistique lui sont inférieures ou égales. De même, le quartile Q_3 est la plus petite valeur du caractère pour laquelle 75 % des valeurs de la série statistique lui sont inférieures ou égales
Il y a donc trois quartiles, le $2^{ème}$ quartile correspondant à la médiane



Là encore, le procédé de calcul des quartiles est différent selon qu'il s'agit de variables discrètes en nombre pair ou impair ou de variables continu.

Définition :
Les déciles et les centiles sont les valeurs du caractère qui partagent l'effectif total en respectivement 10 et 100 parties égales.
Plus précisément :
Le décile D_1 est la plus petite valeur du caractère pour laquelle 10 % des valeurs de la série statistique lui sont inférieures ou égales. On définit de même le décile D_9. On remarque que le $5^{ème}$ décile est égal à la médiane et que le $50^{ème}$ centile est égal à la médiane

jgcuaz@hotmail.com

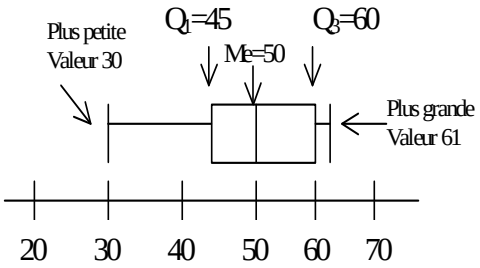
6) Diagrammes en boîte

Afin de représenter différentes caractéristiques d'une série statistique, on a recours, entres autres, aux représentations dites "diagrammes en boîte" ou "diagrammes à moustaches" ou "diagrammes à pattes".

Exemple :
Considérons la série statistique suivante :

Valeur du Caractère	50	45	30	60	61
Effectif	2	3	2	2	2

On vérifie facilement que Me=50 ; Q_1=45 et Q_3=60
D'autre part, la plus petite valeur de cette série est 30, et la plus grande, 61
On peut représenter graphiquement ces résultats de la manière suivante :



7) Caractères de dispersion

Intervalles interquartile

L'intervalle interquartile est une caractéristique de dispersion simple.
Par définition, il est égal à Q_3-Q_1. Il représente la zone centrale comprenant 50% des éléments, et est une mesure de dispersion qui élimine l'influence des valeurs extrêmes.

On utilise également le demi-interquartile (Q), encore appelé déviation partielle : $Q = \frac{Q_3 - Q_1}{2}$

Enfin, pour comparer la dispersion de deux séries dont les éléments sont mesurés avec des unités différentes, ou dont l'ordre de grandeur n'est pas le même, on emploie le rapport de l'interquartile à la médiane, appelé interquartile relatif, défini par $\frac{Q_3 - Q_1}{Q_2} = \frac{Q_3 - Q_1}{Me} = \frac{Q}{Me}$

Ecart absolu moyen (ou écart arithmétique)

Il est égal à la moyenne arithmétique des différences (en valeur absolue) existant entre les divers éléments et leur moyenne.

Exemple : Considérons une suite de salaires horaires : 55,58,62,63,65,69,71,77,83
Leur moyenne est de $\bar{x} = \frac{55+58+62+63+65+69+71+77+83}{9} = 67$
Les écarts des divers salaires et de leur moyenne sont donc :

55	58	62	63	65	69	71	77	83
-67	-67	-67	-67	-67	-67	-67	-67	-67
---	---	---	---	---	---	---	---	---
-12	-9	-5	-4	-2	+2	+4	+10	+16

jgcuaz@hotmail.com

Il est bien évident que la *somme algébrique* des écarts à la moyenne sera nulle (compte tenu de leurs signes) dans tous les cas et ne fournira, par suite, aucun renseignement sur la dispersion. Aussi additionne-t-on les *valeurs absolues*, l'écart moyen, ou écart arithmétique, e_a *étant égal, en définitive à*

$$\frac{12+9+5+4+2+2+4+10+16}{9}=\frac{64}{9}=7.11$$

Le calcul de l'écart moyen, quoique donnant une vue assez fidèle de la dispersion, est peu employé, car il se trouve compliqué par l'intervention des valeurs absolues, peu compatibles avec les calculs algébriques
D'où l'idée de considérer non plus les valeurs absolues des différences, mais leurs *carrés*, toujours positifs, et dont la somme, par conséquent, ne peut s'annuler.

Variance et écart-type

Définitions:
La variance V d'une série est la moyenne arithmétique des carrés des écarts à la moyenne.

$$V=\frac{\sum n_i\left(x_i-\overline{x}\right)^2}{\sum n_i}$$

L'écart-type d'une série est la moyenne quadratique des écarts à la moyenne, autrement dit, c'est la racine carrée de la variance.
On utilise souvent le symbole "sigma minuscule" σ

$$\sigma=\sqrt{\frac{\sum n_i\left(x_i-\overline{x}\right)^2}{\sum n_i}}=\sqrt{V}$$

Méthode de calcul - Théorème de KOENIG

Nous venons de calculer des écarts-types en nous référant à la définition. Cependant, ce calcul risque de devenir laborieux si la moyenne n'est pas un nombre entier : on a à traiter des "écarts à la moyenne" non entiers avec d'inévitables arrondis, d'où des calculs lourds et forcément peu précis. Pour alléger ces calculs, on se sert du théorème suivant:

Théorème de KOENIG:
Si la population est formée de groupes de n_i individus, chaque groupe correspondant à une valeur x_i, et si $n=\sum n_i$, alors

$$V=\frac{\sum n_i {x_i}^2}{n}-\left(\overline{x}\right)^2$$

Autrement dit, la variance est égale à la moyenne des carrés moins le carré de la moyenne.

Ce résultat simplifie considérablement les calculs nécessaires pour obtenir la variance et l'écart-type.

Exemple :
Le tableau suivant nous donne les notes obtenues par deux élèves à 4 contrôles coefficientés :

Notes de l'élève *A*	8	10	8	10
Notes de l'élève *B*	5	13	5	13
Coefficients	1	2	2	1

La moyenne de l'élève *A* est de $\overline{x}_A=\frac{1\times8+2\times10+2\times8+1\times10}{1+2+2+1}=9.$

La variance de l'élève *A* est :

$$V_A=\frac{1\times8^2+2\times10^2+2\times8^2+1\times10^2}{1+2+2+1}-\overline{x_A}=82-81=1 \text{ d'où } \sigma_A=\sqrt{V_A}=1$$

Calculer l'écart type de l'élève *B*. Quel est l'élève le plus régulier , c'est à dire celui qui a le plus petit écart type ?

jgcuaz@hotmail.com

8) Expérience aléatoire, simulations

Définition :
On appelle expérience aléatoire toute expérience réalisée suivant un protocole expérimental précis et reproductible à l'identique, dont les résultats sont liés au hasard, mais dont on peut dresser la liste des résultats possibles.

Exemples :
1) Jet d'un dé. L'ensemble des résultats possibles est {1;2;3;4;5;6}
2) Jet d'une pièce L'ensemble des résultats possibles est {PILE;FACE}

Définition : On appelle événement toute partie de l'ensemble des résultats d'une expérience aléatoire.

Exemple : Jet d'un dé. L'événement "obtenir un nombre pair" est le sous-ensemble {2;4;6}

Définition : On appelle fréquence d'apparition d'un événement le rapport entre le nombre de réalisations de cet événement et le nombre de répétitions de l'expérience aléatoire.

Exemple : Si, au cours de 10 lancer de dès, le numéro 5 apparaît 3 fois, alors la fréquence d'apparition de l'événement « le n°3 apparaît » est $\frac{3}{10}$

Propriétés : La fréquence d'un événement est la somme des fréquences des valeurs constituant cet événement.

Exemples :
Jet d'un dé. L'ensemble des résultats possibles est {1;2;3;4;5;6. Les fréquences de chacune de ces valeurs sont données par

Nombre	1	2	3	4	5	6
Fréquence	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

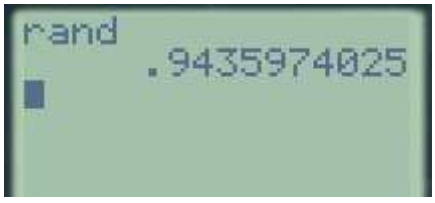
La fréquence l'événement "obtenir un nombre pair" est égale à $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}=\frac{1}{2}$

Simulation

Définition : Simuler une expérience aléatoire, c'est remplacer cette expérience par une autre, plus rapide et plus facile à exécuter, à condition que les fréquences d'apparition de tous les événements possibles soient identiques pour les deux expériences

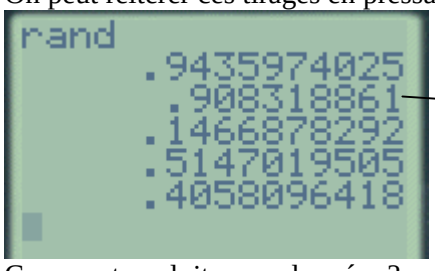
L'instruction RANDOM de la calculatrice

Les calculatrices possèdent une instruction permettant de simuler le tirage aléatoire d'un nombre décimal appartenant à l'intervalle [0 ;1[grâce à l'instruction Rand ou Ran#



CASIO	TI
Menu MATH+PRB ou OPTN+PRB	Menu MATH+PRB
Instruction Ran #	Instruction Rand

On peut réitérer ces tirages en pressant plusieurs fois sur la touche ENTER (ou EXE)



La dernière décimale est un 0, qui n'est donc pas affiché

Comment exploiter ces données ?

1ère exploitation :
Pour chaque décimal renvoyé, si il est strictement inférieur à 0,5 on associe PILE, si il est supérieur ou égal à 0,5 on associe FACE. Cette simulation appliquée à la capture ci-dessus donnerait

Face	PILE	FACE
Fréquence	$\frac{1}{5}=0,2$ donc 20%	$\frac{4}{5}=0,8$ donc 80%

jgcuaz@hotmail.com

2ème exploitation :
On exploite chacune des décimales du nombre renvoyé avec la convention :
Si la décimale est strictement inférieure à 5, on associe PILE
Si la décimale est supérieure ou égale à 5, on associe FACE
Cette simulation appliquée à la capture ci-dessus donnerait

Face	PILE	FACE
Fréquence	$\frac{23}{49} \approx 0,47$ donc environ 47 %	$\frac{26}{49} \approx 0,53$ donc environ 53 %

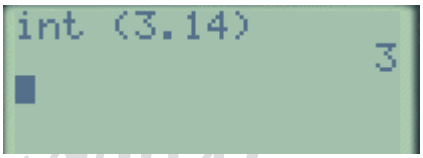
3ème exploitation :
On exploite chacune des décimales du nombre renvoyé avec la convention :
Si la décimale est paire, on associe PILE
Si la décimale est impaire, on associe FACE
Cette simulation appliquée à la capture ci-dessus donnerait

Face	PILE	FACE
Fréquence	$\frac{26}{49} \approx 0,53$ donc environ 53 %	$\frac{23}{49} \approx 0,47$ donc environ 47 %

L'instruction INT de la calculatrice

Les calculatrices possèdent une instruction permettant de calculer la partie entière d'un nombre décimal

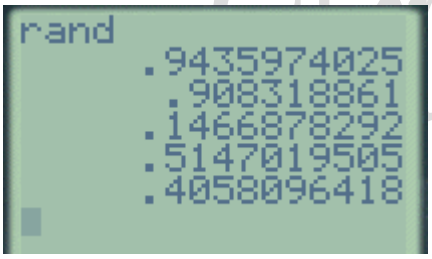
CASIO	TI
Menu MATH+NUM ou OPTN+NUM	Menu MATH+NUM
Instruction Int	Instruction Int



Exemples :
Int (3,14)=3 et Int (2,999999)=2

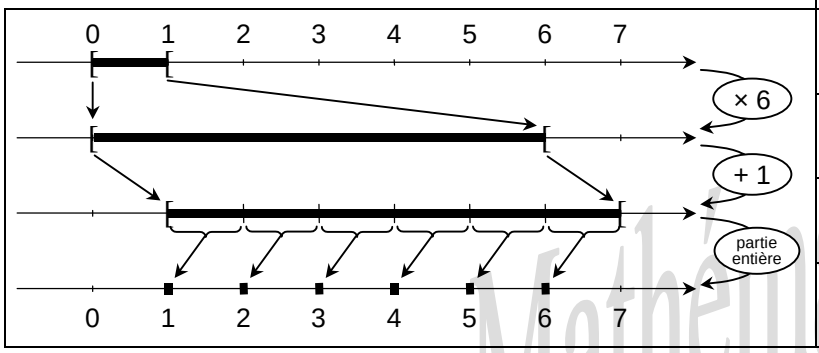
ATTENTION
Il s'agit bien de la troncature et non pas d'un arrondi

En reprenant la simulation précédente :



Nombre décimal	$10\times(\text{nombre décimal})$	$Int\left(10\times(\text{nombre décimal})\right)$
0,9435974025	9,435974025	9
0,908318861	9,08318861	9
0,1466878292	1,466878292	1
0,5147019505	5,147019505	5
0,4058096418	4,058096418	4

Propriété :
Il est possible d'obtenir une suite de nombres entiers compris entre a et $a+b-1$ en utilisant par exemple l'instruction INT(b×RAN#) + a



Instruction	Le résultat est alors :
RAN#	compris entre 0 et 1
INT(6×RAN#)	compris entre 0 et 6
INT(6×RAN#) + 1	compris entre 1 et 7

jgcuaz@hotmail.com