STATISTIQUES

INFIRMIER(E) GRADUE(E) SPECIALISE(E) EN SANTE COMMUNAUTAIRE

HAUTE ECOLE DE LA PROVINCE DE LIEGE

PROFESSEUR : RENARD X.

Année scolaire 2009-2010

TABLE DES MATIERES

CHAPITRE 1: ELÉMENTS DE STATISTIQUES DESCRIPTIVES

1. INTRODUCTION

Sur base des documents observés (extraits de journaux, revues, livres, …), on se rend compte sans difficulté que les études statistiques envahissent notre vie. Mais comment peut-on définir la statistique ?

La s***tatistique*** est une branche des mathématiques qui a pour but, dans un premier temps, de rassembler une série de données et de les présenter *(statistique descriptive)*. Dans un deuxième temps, ces données sont interprétées afin d'en tirer des conclusions et d'effectuer des prévisions éventuelles *(statistique inférentielle)*.

L'interprétation et l'utilisation de données statistiques se retrouve dans de très nombreux domaines dont notamment, les sciences humaines, les sciences économiques, les médias, la gestion des entreprises, la recherche médicale, ... Ce n'est pas pour rien qu'un cours de statistique est présent en première année de la plupart des graduats et des universités.

Exemple d'application *: l'accidentologie (études scientifiques des accidents)*

Grâce aux statistiques, nous avons aujourd'hui une meilleure connaissance de l'accidentologie. Les premières statistiques des accidents de la circulation remontent pratiquement à la naissance de l'industrie automobile. Mais, au fil des années, elles n'ont cessé de s'affiner, gagnant en fiabilité, en précision et en rapidité, afin de disposer d'une connaissance détaillée de notre accidentologie.

Il y a trois bonnes raisons à cela.

* *Il s'agit pour les pouvoirs publics d'avoir la **vision** la plus claire, la plus précise possible sur les causes et les conditions des accidents qui surviennent sur les différents réseaux. (les lieux, les conditions atmosphériques, l'éclairement, les individus et les véhicules impliqués, ...). Le choix des "armes" dépend étroitement de l'ennemi que l'on a à combattre… et seules les statistiques permettent de bien le définir.*

* *Une fois les **actions** décidées et mises en œuvre, il s'agit d'évaluer leur efficacité sur le terrain. Là encore, les statistiques permettent de mesurer objectivement les effets des actions entreprises, de façon à pouvoir les généraliser si celles-ci sont positives, ou bien à les amender si elles ne donnent pas entière satisfaction.*

* *La diffusion des statistiques permet de faire **partager** à l'ensemble des usagers notre **connaissance** sur l'accidentologie. C'est indispensable si l'on veut obtenir leur adhésion à une lutte qui passe forcément par un consensus social. Savoir, par exemple, qu'une baisse sensible des accidents qui coïncide avec la mise en place de nouvelles mesures peut convaincre des usagers, jusque-là incrédules, de l'intérêt de ces mesures ...*

Des mesures de vitesses sont réalisées sur les différents réseaux routiers. Les appareils utilisés permettent également de repérer les interdistances entre les véhicules. L'évolution des comportements sur certains points importants (l'alcool, le port de la ceinture et du casque, etc.) vient compléter les données strictement accidentologiques. Et d'autres enquêtes doivent s'y ajouter : sur le respect des feux rouges, l'utilisation des téléphones portables au volant, etc ...

Au départ, on se base généralement sur les résultats d'une enquête. La variable (ou le caractère) étudiée dans l'enquête peut être de différents types :

- ***données ou variables non chronologiques*** : Dans ce cas, la variable (ou le caractère étudié lors de l'enquête) peut être discrète ou continue ainsi que qualitative ou quantitative. Ce sont des données non chronologiques, c'est-à-dire des données dont on n'étudie pas l'évolution en fonction du temps. (Exemples : taille, poids, vote, … d'un ensemble de personnes à un moment donné)

- ***données ou variables chronologiques*** : On analyse l'évolution des valeurs de la variable en fonction du temps. (Exemples: chiffre d'affaire d'une société au cours des années, population de Schaerbeek de 1831 à 1970)

2. LES DIFFERENTS TYPES DE VARIABLES NON CHRONOLOGIQUES

Les différents types de variables non chronologiques sont les suivantes :

2.1. Variable qualitative

Une variable ***qualitative*** exprime une qualité et est une variable dont les valeurs ne sont ni mesurables ni repérables (ce ne sont pas des nombres).

Les valeurs prises par une variable qualitative sont appelées des « ***modalités*** » qui portent des noms. C'est la raison pour laquelle on parle aussi de « ***variable nominale*** ».

Exemples: couleur, profession, marque de voitures, ...

Variable ordinale

Les modalités d'une variable qualitative sont parfois ordonnées. On parle dans ce cas de variable ***ordinale***.

Exemples :

- Les grades aux examens : Aj, S, D, GD, PGD.
- Les degrés d'une brulure : 1^{er}, $2^{ème}$, $3^{ème}$.
- La pratique d'un sport : jamais, rarement, souvent, très souvent.

2.2. Variable quantitative

Une variable quantitative est une variable dont les valeurs sont mesurables ou repérables par des nombres réels. Les valeurs de la variable sont des nombres. *Exemples: salaire, température, taille, poids, ...*

2.2.1. Variable discrète quantitative

Une variable quantitative discrète est une variable dont les valeurs sont en nombre fini (un petit nombre de valeurs possibles).

Exemples: âges des élèves de $6^{ème}$ secondaire, nombre de filles dans une famille de cinq enfants, ...

2.2.2. Variable continue quantitative

Une variable quantitative continue est une variable qui peut prendre toutes les valeurs possibles dans un intervalle (un grand nombre de valeurs possibles).

Exemples: taille d'une population, poids, la pression artérielle, ...

Exercices

Parmi les exemples suivants, indique quelles sont les caractéristiques de la variable étudiée.

- ✓ Le nombre d'enfants de 0 à 24 ans par famille en France en 2003.
- ✓ Les marques de voitures neuves immatriculées en Belgique pendant le 1^{er} semestre de l'année 2002.
- ✓ La marque de GSM préférée des adolescents.
- ✓ Le temps quotidien passé devant la télévision
- ✓ Le nombre de lancés sur 5 réussis au basket-ball pour 25 élèves.
- ✓ La taille de cent personnes adultes de sexe masculin.
- ✓ Les marques de chocolat les plus appréciées par les élèves de l'école.

Voici un tableau brut donnant les puissances (en watts) des ampoules disponibles lors d'un inventaire :

60	100	40	100	150	60	100	40
100	100	60	60	60	40	75	60
75	150	40	40	100	75	75	150
60	100	150	75	60	100	100	100

Quelle variable est étudiée ? Est-elle quantitative ou qualitative ? Est-elle continue ou discrète ?

Réponse : .....

Voici un tableau brut reprenant la taille des élèves d'une classe. On obtient les résultats suivants :

165	172	181	158	172	156	190
192	168	175	180	184	159	178
162	161	185	195	178	189	175
159	160	182	186	192	187	152
168	165	178	175	175	182	180

Quelle variable est étudiée ? Est-elle quantitative ou qualitative ? Est-elle continue ou discrète ?

3. LA COLLECTE DES DONNÉES

3.1. L'échantillonnage

Si on veut résoudre le problème : "quel est l'âge moyen des belges ?", on peut envisager deux démarches :

- Relever l'âge et le nombre de belges ***(population)***
- Relever l'âge de quelques milliers de belges ***(échantillon)*** pris au hasard et considérer que les valeurs constatées dans cet échantillon sont identiques à celles que l'on cherche pour la population.

La première solution est la plus précise à condition, ce qui n'est pas certain, d'être réalisable dans un délai tel qu'aucune naissance ou qu'aucune mort ne viennent modifier le résultat.

La seconde solution présente l'avantage de la rapidité pour autant que les valeurs constatées dans l'échantillon puissent être reportées à la population entière, c'est-à-dire que l'échantillon soit ***représentatif*** de l'ensemble de la population.

Cette technique qui consiste à mesurer sur un échantillon des valeurs qu'il est impossible ou difficile de mesurer sur la population entière constitue une des bases de la statistique.

3.2. Représentativité d'un échantillon

Un échantillon représentatif d'une population pour une variable est un échantillon pour lequel on n'a pas de raison de penser que la valeur observée de la variable diffère dans l'échantillon et la population. En pratique, toute l'évaluation statistique est basée sur l'obtention d'échantillons représentatifs, la manière de les obtenir, le contrôle de leur représentativité, leur traitement mathématique.

> *Un bel exemple d'erreur d'échantillonnage conduisant à une évaluation erronée peut être cité : dans un laboratoire d'expérimentation en toxicologie, les animaux d'expérience sont prélevés dans des parcs sans méthode apparente de sélection. A première vue, les animaux sont prélevés au hasard. Il est fort possible que la capture des animaux les plus vifs soit moins probable que celle des animaux en mauvaise santé et que l'échantillonnage conduise à des conclusions fausses.*

3.3. Conditions d'obtention d'un échantillon représentatif

- Les individus de la population doivent tous avoir la même chance d'être sélectionnés,
- L'effectif de l'échantillon doit être grand (plusieurs milliers pour les sondages d'opinion),
- L'échantillon doit éviter les "mortalités", c'est-à-dire la perte d'un certain nombre d'individus choisis pour constituer l'échantillon (élèves malades lors d'un test par exemple)

3.4. Echantillon aléatoire

La première des conditions citées ci-dessus correspond à la notion d'obtention d'un échantillon aléatoire, c'est à dire que:

- l'extraction des individus de la population pour constituer l'échantillon s'est réellement faite au hasard
- la population n'a pas été significativement modifiée par l'extraction de l'échantillon. Si la population est très nombreuse, cette condition ne pose pas de problème; si elle est très réduite, il vaut mieux étudier la population dans son ensemble plutôt que d'en extraire un échantillon.
- la population est bien définie

Comme on l'a vu plus haut, constituer l'échantillon de manière aléatoire n'est pas aussi simple que l'on peut le croire à première vue; il faut éviter une série de pièges d'une sélection involontaire ou d'un rejet systématique de certains des individus. Une étude préalable sérieuse à la fois de la population et du paramètre considéré doit être conduite à terme; cette étude visera entre autres à préciser si le paramètre fluctue de manière continue dans la population.

3.5. Quelques types d'erreurs à éviter

A titre d'exemple, nous pouvons citer les erreurs les plus courantes lors de la constitution d'un échantillon :

- *des raisons de commodité inclinent parfois le chercheur à considérer comme valables des échantillons pris à proximité ou limités à un grand ensemble géographique, ville ou région. Or la localisation humaine joue souvent un grand rôle sur les paramètres des individus.*

- *l'élimination systématique d'une partie de la population lors de la constitution d'un échantillon est classique: on peut citer en exemple l'utilisation du bottin, qui élimine les non possesseurs de téléphone, l'appel aux téléspectateurs, la lettre à laquelle il faut répondre.*

- *la «mortalité» d'un échantillon résulte moins d'une erreur de méthode que de circonstances particulières ou temporaires: épidémie de grippe dans une école testée, enquêtes effectuées durant les vacances... . Lorsque l'on s'aperçoit qu'une partie non négligeable de l'échantillon envisagé est temporairement hors d'état de livrer ses paramètres, il vaut mieux remettre les tests à une date ultérieure.*

- *dans le monde de la médecine et de la pédagogie, les individus testés ont parfois un comportement particulier du au fait qu'ils se savent observés: l'inspecteur qui visite une classe sait que sa présence provoque, chez les élèves, des réactions diverses; le médecin qui examine un patient, surtout si celui-ci sait que l'examen participe à un plan d'ensemble, est susceptible de donner, sur ses paramètres personnels, des indications fausses.*

Cependant, dans une étude, même s'il faut éviter au maximum ***les biais*** *(écarts entre les "vraies" valeurs et les valeurs observées), ce n'est pas toujours possible.*

Par exemple, dans une étude récente sur les effets nocifs des ondes GSM sur la santé, on a interrogé des patients porteurs d'une tumeur dans le lobe frontal ou temporal (là où passent les circuits de la mémoire). La question était : "Combien de fois par jour téléphoniez-vous avec votre portable, il y a dix ou quinze ans et quelle était la durée de vos appels ?".

La maladie pourrait très bien dans ce cas altérer les souvenirs ! Il faut donc le vérifier et éventuellement effectuer des corrections.

4. PREMIER EXEMPLE : VARIABLE DISCRETE QUANTITATIVE

4.1. Vocabulaire statistique et mise en forme des données

En statistique, comme dans tous les domaines, il y a lieu d'utiliser un vocabulaire particulier. On partira d'un exemple pratique pour définir petit à petit le "langage statistique".

> ***On désire réaliser l'enquête suivante : Quel est l'âge moyen d'un étudiant de 1ère BSI en Belgique ?***

1ère partie : Pour étudier ce problème, il faut réaliser une enquête et poser la question à chaque étudiant de chaque classe de ***1ère kiné*** de chaque école de Belgique.

Population = ensemble de tous les étudiants de ***1ère BSI*** de chaque école de Belgique.

La population comporte un certain nombre d'***individus***. Il est évidemment difficile d'interroger tous les étudiants de 1ère ***BSI*** de Belgique. Pour l'exemple, on se limitera donc à un échantillon de 27 étudiants (l'échantillon n'étant ici pas représentatif de la population).

2ème partie : L'enquête ci-dessus est proposée à un ***échantillon*** c'est-à-dire une partie de la population.

Il compte 27 individus (27 étudiants).

3ème ***partie*** : Résultats de l'enquête.

18	*20*	*19*	*20*	*21*	*18*	*20*	*18*	*18*
21	*19*	*19*	*19*	*18*	*18*	*18*	*21*	*20*
19	*19*	*20*	*19*	*18*	*20*	*22*	*19*	*18*

Il s'agit d'un ***tableau brut*** c'est-à-dire un tableau où sont notés les résultats au fur et à mesure qu'ils se présentent.

Le nombre de données contenues dans ce tableau est appelé l'***effectif total n***. Ici, l'effectif total vaut : n = 27.

> *Remarque : dans cette enquête, l'effectif total correspond au nombre d'individus que contient l'échantillon puisque chaque individu a donné un résultat.*
>
> *Ce n'est pas toujours le cas. Exemple: enquête du type: Quel est l'âge des enfants de votre famille?*
>
> *Chaque individu peut donner plusieurs réponses. Par conséquent, le nombre d'individus n'est pas égal à l'effectif total.*

4ème partie : Classement des résultats de l'enquête.

Le tableau brut est difficilement utilisable tel quel surtout quand le nombre de données est très élevé.
On va donc classer les données pour rendre le tableau plus facile à exploiter.

i	X_i	n_i ***: effectifs***	f_i ***: fréquences***
1	18	*9*	*9/ 27 = 0,333 = 33,3 %*
2	19	*8*	*8 / 27 = 0,296 = 29,6 %*
3	20	*6*	*6 / 27 = 0,222 = 22,2 %*
5	21	*3*	*3 / 27 = 0,111 = 11,1 %*
6	22	*1*	*1 / 27 = 0,037 =3,7 %*
		$n = \sum_{i=1}^{6} n_i = 27$	$\sum_{i=1}^{6} f_i = 0{,}99$

1ère colonne: on y place la ***variable notée X*** qui est la caractéristique étudiée lors de l'enquête (âge d'un étudiant de 1ère BSI). Dans le cas présent, la variable X est discrète car elle peut prendre un nombre fini de valeurs (5 valeurs: 18, 19, 20, 21 ou 22).

2ème colonne: chaque valeur de la variable apparaît plusieurs fois. ***L'effectif*** (appelé aussi répétition pour une variable discrète) est le nombre de fois qu'apparaît chaque valeur de la variable dans le tableau brut. On le note n_i.
La somme des effectifs n_i est appelée ***l'effectif total n***.

3ème colonne: c'est la fréquence f_i c'est-à-dire le rapport entre chaque effectif et l'effectif total.

$$\text{fréquence } f_i = \frac{\text{effectif } n_i}{\text{effectif totaln}}$$

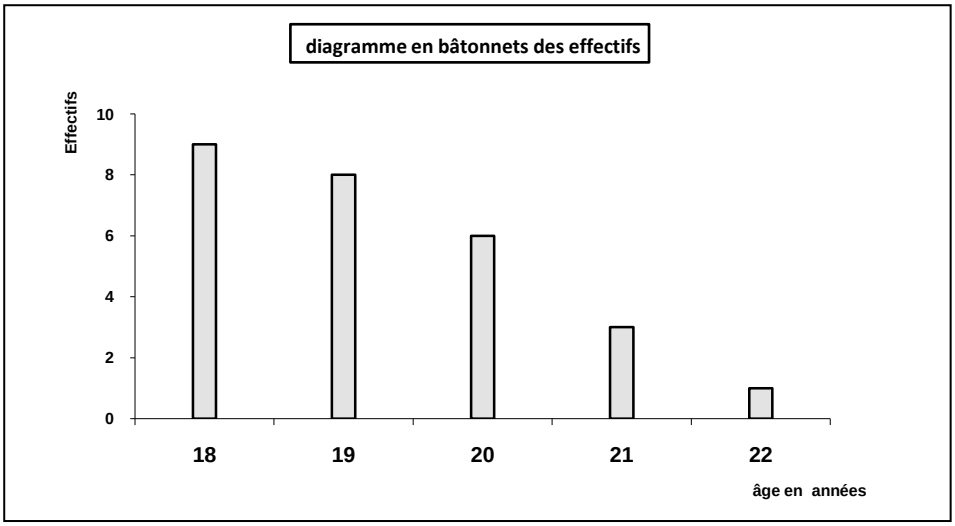
On constate que le total des fréquences n'est pas égal à 100 %. Ceci est du bien sûr aux erreurs d'arrondi. Pour éviter ces erreurs, on utilisera un tableur (comme EXCEL, par exemple).

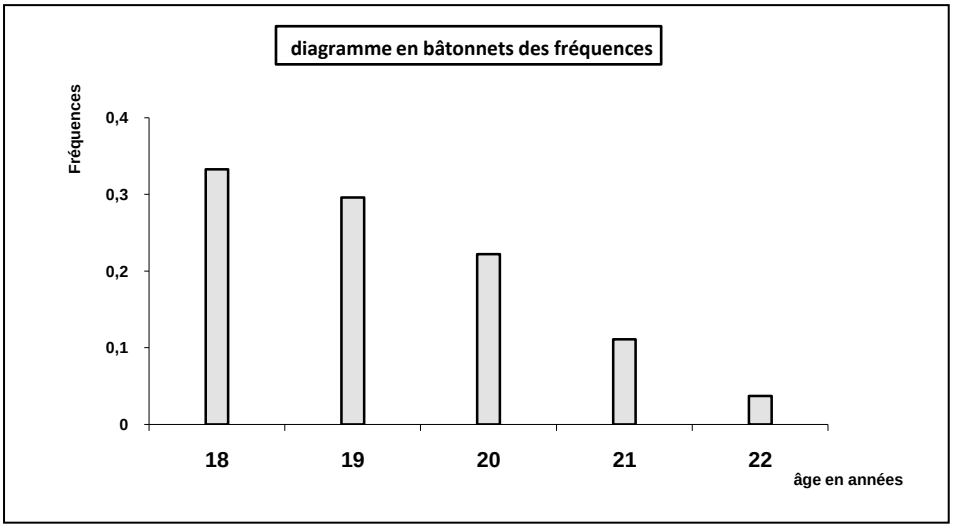
Le tableau résultant du classement des données est appelé ***tableau recensé ou ordonné***.

4.2. Représentations graphiques.

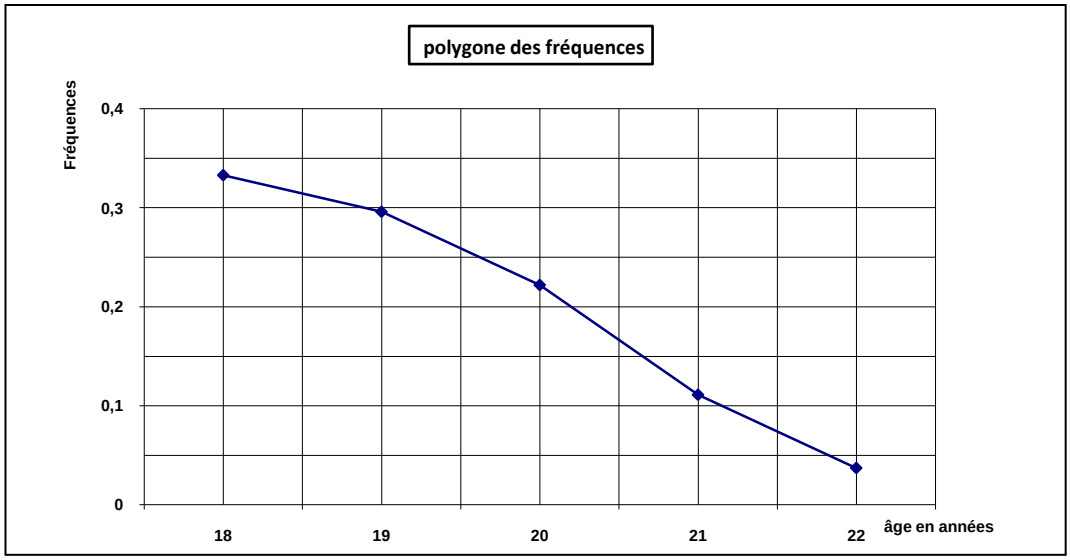
4.2.1. Le diagramme en bâtonnets

On porte en abscisse la variable (ici l'âge) et en ordonnée la fréquence ou l'effectif.



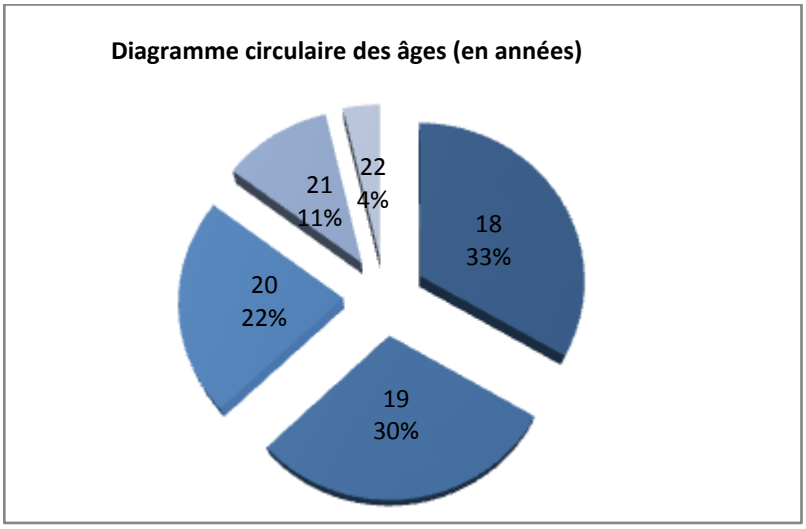


En joignant les extrémités des bâtonnets, on obtient un polygone appelé ***polygone des fréquences***.



4.2.2. *<u>Diagramme circulaire (fréquences)</u>*

i	***X_i***	***n_i***	***fréquences f_i en %***	***fréquences f_i en °***
1	18	3	33,3 %	33,3.3,6 = *120°*
2	19	7	29,6 %	106,7°
3	20	10	22,2 %	80°
4	21	4	11,1 %	40°
5	22	3	3,7 %	13,3°
		Effectif total: n = 27		***Somme = 360°***



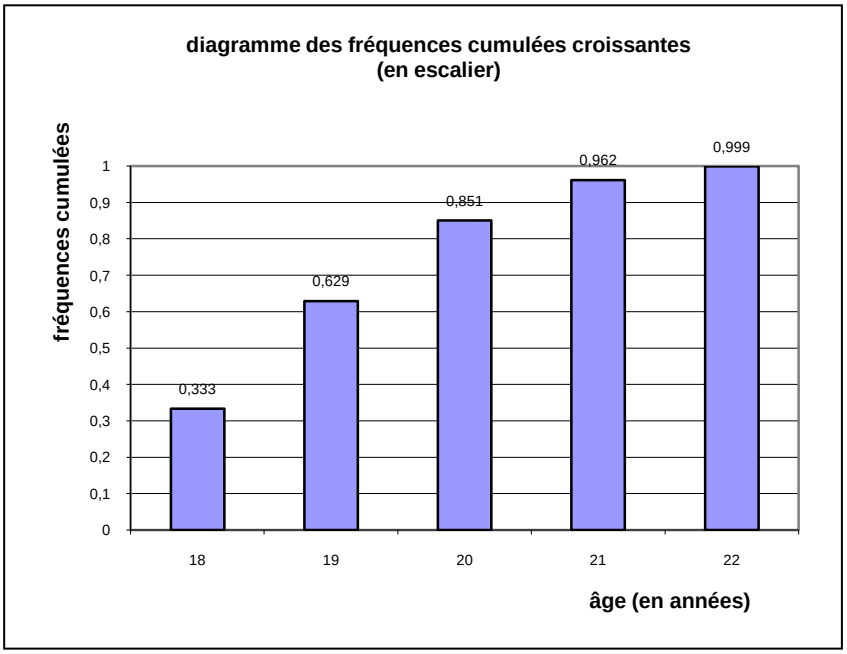
4.2.3. *<u>Le diagramme des fréquences cumulées ou diagramme en escalier</u>*

Reprenons le tableau de données.

i	X_i	n_i	f_i ***en %***	f_i ***cumulées croissantes***	f_i ***cumulées décroissantes***
1	18	*9*	33,3 %	33,3 %	99,9 %
2	19	*8*	29,6 %	62,9 %	66,6 %
3	20	*6*	22,2 %	85,1 %	37,0 %
5	21	*3*	11,1 %	96,2 %	14,8 %
6	22	*1*	3,7 %	99,9 %	3,7 %

On peut tracer un "diagramme des fréquences cumulées" avec en abscisse l'âge et en ordonnée les fréquences cumulées (croissantes ou décroissantes).

On voit directement sur ce graphique qu'il y a 62,9 % des élèves qui ont 19 ans ou moins.



5. DEUXIEME EXEMPLE : VARIABLE CONTINUE QUANTITATIVE

5.1. Mise en forme des données

Soit un tableau reprenant les salaires mensuels bruts en euros dans le secteur de la kinésithérapie (valeurs fictives, n'en déduisez rien !).

270 275 300 455 323 642 254 532 541 335 ..... (67 valeurs différentes)

C'est le tableau brut. L'échantillon est de 67 kinés. L'effectif total est aussi de 67 (un salaire par personne).

Si on classait ces données comme dans le premier exemple, on obtiendrait un très grand tableau avec une colonne effectif presque toujours égale à 1 (peu de valeurs sont identiques). On ne saurait que faire de ce tableau.

Pour classer ces données, on va procéder autrement, on va les regrouper en ***classes***.

i : numéro de la classe	***Classes (salaires en €)***	***Centres de classe***	***n_i: effectifs***	***f_i: fréquences***	***f_i cumulées croissantes***
1	< 300	275	4	$\frac{4}{67}$ = 0,060 (6,0 %)	6 %
2	[300; 350[	325	12	17,9 %	23,9 %
3	[350; 400[	375	20	29,9 %	53,8 %
4	[400; 450[	425	15	22,4 %	76,2 %
5	[450; 500[	475	10	14,9 %	91,1 %
6	≥ 500	525	6	9,0 %	100,1 %
			effectif total = 67	***somme = 100,1 % = 1***	

L'effectif total est égal à : n = $\sum_{i=1}^{6} n_i$ = 67.

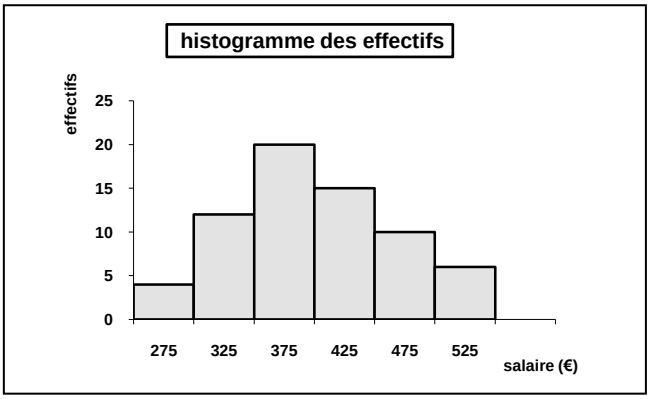
Dans la deuxième classe par exemple, les nombres 300 et 350 sont appelés respectivement ***borne inférieure et borne supérieure de la classe***.

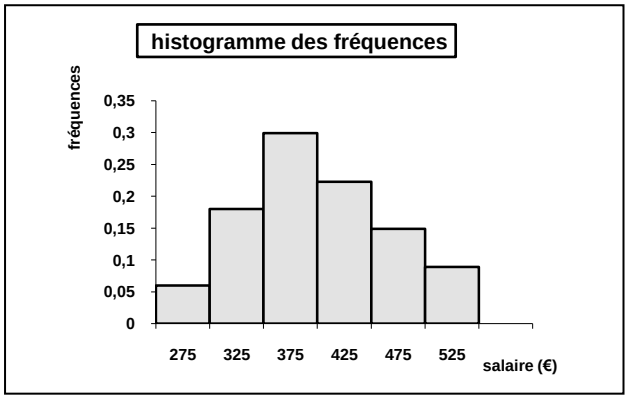
Le centre de la deuxième classe est 325.

5.2. Représentations graphiques

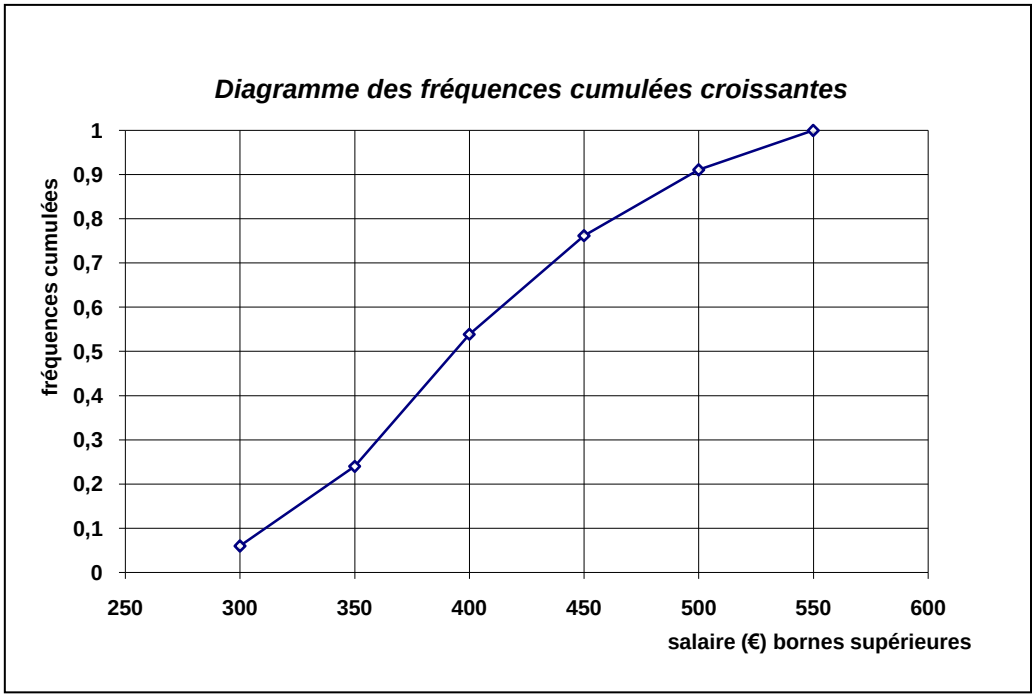
5.2.1. L'histogramme des effectifs ou des fréquences

On porte en abscisse les centres des classes (ou les intervalles de classe) et en ordonnée les effectifs ou les fréquences (uniquement dans le cas ou les classes ont même amplitude).





5.2.2. *Diagramme des fréquences cumulées*



Dans ce cas, on porte en abscisse les extrémités des classes (les bornes supérieures) et on suppose que les valeurs des individus se répartissent de manière linéaire entre ces points.

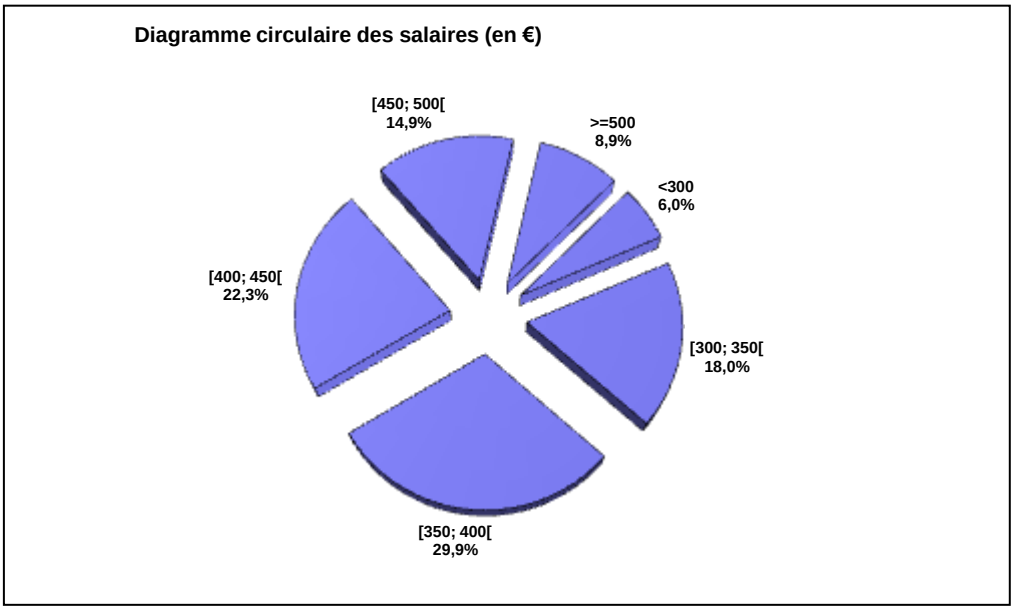
C'est sur ce type de graphique que l'on se basera pour la détermination de la médiane, des quartiles, déciles, …

Question: Quel est le pourcentage de kinés qui ont un salaire inférieur ou égal à 400 € ?

Remarque importante :

- dans le cas d'un diagramme des ***fréquences cumulées croissantes***, on porte en abscisse les extrémités des classes (**les bornes supérieures**).
- dans le cas d'un diagramme des ***fréquences cumulées décroissantes***, on porte en abscisse les origines des classes (**les bornes inférieures**).

5.2.3. *Diagramme circulaire*



5.2.4. Autres types de graphiques

Les graphiques tracés jusqu'ici sont les plus courants que l'on rencontre. Il est évident qu'il existe de nombreux autres types de graphiques qui permettent de caractériser des séries statistiques.

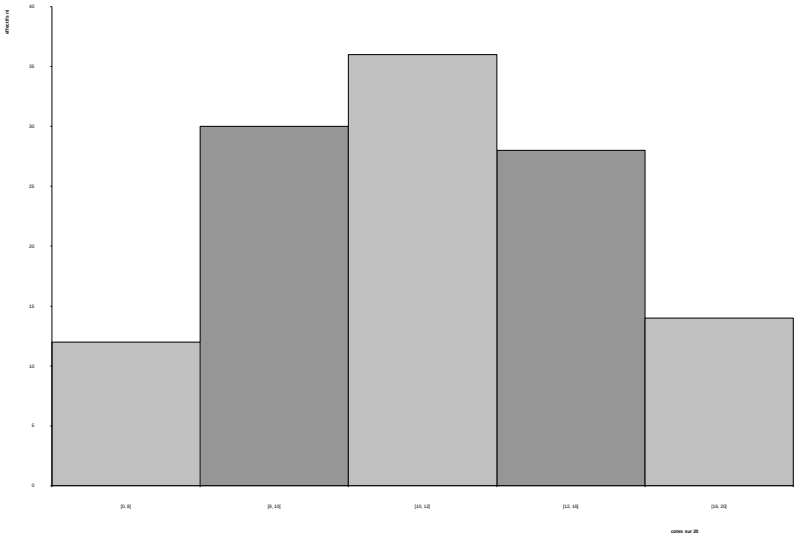
5.3. Remarque : classes d'amplitudes différentes

Dans l'exemple précédent, les classes avaient la même amplitude. Ce n'est cependant pas toujours le cas. Prenons l'exemple suivant.

Les résultats d'un groupe de 120 élèves dans une branche donnée sont représentés dans le tableau : (résultats exprimés sur 20)

Cotes sur 20	effectifs n_i
[0, 8[	12
[8, 10[	30
[10, 12[	36
[12, 16[	28
[16, 20]	14
	n = 120

Si on essaie de tracer un histogramme à l'aide d'EXCEL, on obtient le diagramme suivant :



Cette présentation n'est pas satisfaisante puisque les classes sont représentées comme si elles avaient la même amplitude.

En réalité le tableur Excel, lorsqu'on lui demande de tracer un histogramme, trace un diagramme en bâtonnets.

Dans un histogramme, chaque classe et son effectif sont représentés par un rectangle. La largeur du rectangle est l'amplitude de la classe ; la hauteur du rectangle est ajustée de manière que l'**aire** du rectangle soit proportionnelle à l'effectif de la classe.

Par conséquent, la première classe, qui a un effectif de 12 est donc représentée par un rectangle de largeur 8 (l'amplitude de la classe) et de hauteur 1,5, puisque 8 x 1,5 = 12.

Exercice : Suite à cette explication, tracer l'histogramme correct.

6. PETIT TEST RECAPITULATIF SUR LE VOCABULAIRE

Au cours d'une enquête dans une classe 20 élèves, on pose les questions suivantes :

> ***Combien avez-vous de frères et sœurs ?***
> ***Quelle est leur taille ?***
> ***Quel moyen de transport utilisez-vous pour venir à l'école ?***

Compléter les phrases suivantes :

La …….. sur laquelle porte l'enquête est constituée d'un …….. de 20 personnes ; chaque personne représente un …….. de la série statistique.

La …….. « nombre de frères et sœurs », s'exprime par des nombres : elle est ……, comme celle de la « taille ».

La …….. « moyen de transport » ne s'exprime pas avec un nombre ; elle s'exprime avec un mot : elle est ……..

La …….. « nombre de frères et sœurs » a également un autre caractère, elle est …….. ; de même la ……« taille » est ……..

Les données sont tout d'abord recueillies dans un ……..

Dans le cas d'une variable discrète, ce dernier est mis en forme pour donner alors un …….
Dans le cas d'une variable continue, ce dernier est mis en forme pour donner alors un ……

Ces tableaux retravaillés permettent alors de calculer la …….. de chaque valeur, en divisant l'effectif de cette valeur par l'effectif total. La …….. peut être calculée en ordre croissant et/ou en ordre décroissant.

Le traitement des données permet alors de faire des représentations graphiques statistiques. Dans le cas d'une variable discrète, les principales représentations sont :

- *……,*
- *……,*
- *……..*

Dans le cas d'une variable continue, l'une de ces représentations fait place à un ……

Réponses :

- continue - diagramme en bâtonnets - diagramme circulaire - discrète

- échantillon - fréquence - fréquence cumulée - histogramme - individu

- polygone des fréquences - population - qualitative - quantitative - tableau brut

- tableau recensé et ordonné - tableau recensé, ordonné et groupé - variable

- variable - variable - variable

7. LES PARAMETRES DE POSITION: LA MOYENNE, LE MODE , LA MEDIANE, LES QUANTILES

Quand les statisticiens se trouvent en face des résultats d'une enquête, ils trouvent intéressant d'en déterminer les "tendances moyennes". Pour cela, ils disposent de plusieurs outils: la moyenne arithmétique, le mode, la médiane et les quantiles. La moyenne arithmétique est le paramètre de position le plus utilisé. Il existe également d'autres types de moyenne, comme la moyenne harmonique ou la moyenne géométrique dont on ne parlera pas.

7.1. La moyenne arithmétique

Soit deux hommes affamés auxquels on donne un poulet rôti. Le premier s'en empare et le dévore entièrement. En moyenne, chacun a eu un demi-poulet.
(La Cité de chiffres, Jena-Louis Besson, éd. Autrement)

7.1.1. Cas d'une variable discrète

Reprenons l'exemple de l'âge des élèves de la classe de 6ème générale.

On avait obtenu:

18	20	17	17	17	16	20	18	18
18	19	19	19	18	18	18	19	18
18	18	17	16	16	17	20	17	17

On obtient la moyenne arithmétique en additionnant toutes ces valeurs et en divisant le nombre obtenu par le nombre de valeurs. Elle est désignée par le symbole $\overline{X}$ ("x barre").

Moyenne = $\overline{X}$ = (18+20+17+17+17+16+20+18+18+18+19+19+19+18+18+18+19+18+18+18+17+16+16+17 +20+17+17) / 27 = 483 / 27 = ***17,89***

Si on avait pris comme échantillon la totalité des classes de 6ème de l'Athénée, le calcul aurait été très long et on aurait fait des fautes.

C'est pour cette raison qu'on calcule généralement la moyenne à partir du tableau recensé ou ordonné.

i	X_i	n_i	f_i ***en %***
1	16	3	0,111 = 11,1 %
2	17	7	0,259 = 25,9 %
3	18	10	0,370 = 37,0 %
4	19	4	0,148 = 14,8 %
5	20	3	0,111 = 11,1 %
		n = 27	

On obtient la moyenne en multipliant chaque valeur de la variable par l'effectif correspondant. Le nombre obtenu est divisé par l'effectif total.

$$\overline{X} = \frac{16.3+17.7+18.10+19.4+20.3}{27} = \frac{483}{27} = 17{,}9$$

Dans ce cas, la formule générale de la moyenne s'écrit: $\overline{X} = \frac{\sum_{i=1}^{k} n_i . X_i}{n}$

où *k est le nombre de valeurs de la variable X*
X_i sont les valeurs de la variable
n_i est l'effectif correspondant à la variable i
n est l'effectif total

On peut aussi utiliser les fréquences pour calculer la moyenne. On sait que la fréquence f_i correspondant à la valeur X_i vaut : $f_i = \frac{n_i}{n} = \frac{\text{effectif i}}{\text{effectif total}}$.

On obtient la formule:

$$\overline{X} = \sum_{i=1}^{k} f_i . X_i$$

$$\Rightarrow \overline{X} = 16.0{,}11 + 17.0{,}26 + 18.0{,}37 + 19.0{,}15 + 20.0{,}11 = 17{,}89$$

7.1.2. Cas d'une variable continue, répartition en classes

En 1954, une enquête sur la répartition selon l'âge de la population agricole masculine a donné les résultats suivants:

Age en années X_i	***Centres de classe***	***Effectifs n_i***	***Fréquences f_i***
[15; 25[	*20*	*197*	*0,197*
[25; 35[	*30*	*207*	*0,207*
[35; 45[	*40*	*151*	*0,151*
[45; 55[	*50*	*189*	*0,189*
[55; 65[	*60*	*127*	*0,127*
[65; 75[	*70*	*108*	*0,108*
75 et plus	*80*	*21*	*0,021*
		n = 1000	

Dans ce cas, les X_i qui représentaient les valeurs de la variable représentent les centres des classes. Les formules de calcul de la moyenne sont donc identiques.

$$\overline{X} = \frac{\sum_{i=1}^{k} n_i . X_i}{n}$$

$$\overline{X} = \sum_{i=1}^{k} f_i . X_i$$

où X_i représente le centre de la classe i et k le nombre de classes.

$$\overline{X} = \frac{20.197 + 30.207 + 40.151 + 50.189 + 60.127 + 70.108 + 80.21}{1000} = 42{,}5$$

7.2. Le mode

Dans une enquête relative au moyen de transport, on a obtenu le tableau suivant:

Moyens de transport	***Effectifs***
vélo	7
bus	10
tram	2
vélomoteur	5
à pied	6

Dans ce cas, il n'est pas possible de calculer une moyenne. On pourrait cependant se demander quel est le moyen de transport le plus utilisé ("à la mode"). C'est évidemment le bus qui correspond au plus grand nombre d'effectifs.

- *Dans le cas de variable à valeurs numériques, si on reprend l'exemple des âges de la classe de 6ème générale, on s'aperçoit que l'âge que l'on retrouve le plus souvent est 17 (10 effectifs).*
- *Dans l'exemple de l'âge de la population agricole masculine, le mode est la classe de 25 à 34 ans qui compte l'effectif le plus élevé. (on parle de « classe modale »)*

Le mode est la valeur de la variable (ou la classe) dont l'effectif est le plus important (la valeur de la variable la plus fréquente)

7.3. La médiane

La médiane d'une variable statistique est la valeur de la variable (ou la classe) qui partage l'effectif en deux parties égales.

Dans le cas d'une série ordonnée simple, on peut trouver aisément la médiane :

- ***si n est impair***, la médiane est la valeur de la variable de rang $\frac{n+1}{2}$.
  Par exemple, la série ordonnée {1, 3, 7, 8, 9, 15, 17} comporte 7 observations ; sa médiane est la quatrième observation ($\frac{7+1}{2}$) et est égale à 8.
- ***si n est pair***, la médiane est la moyenne arithmétique des valeurs de la variable de rang $\frac{n}{2}$ et $\frac{n+1}{2}$.

Par exemple, la série ordonnée {1, 3, 7, 8, 9, 15} comporte 6 observations et a pour médiane 7,5, obtenue en prenant la moyenne arithmétique entre la troisième observation (7) et la quatrième (8).

Dans le cas d'une distribution observée plus complexe, il existe diverses formules dans la littérature. ***On se limitera à déterminer la médiane sur le diagramme des fréquences cumulées (la médiane est l'abscisse correspondant à une fréquence de 50 %).***

7.4. Les quantiles

La médiane peut être considérée comme un cas particulier d'une valeur caractéristique plus générale, appelée ***quantile***.

Les quantiles que l'on rencontre le plus souvent sont (définis à partir d'un diagramme des fréquences cumulées) :

a) la ***médiane*** $X_{\frac{1}{2}}$: qui est la valeur de la variable correspondant à une fréquence (cumulée) de 50 %.

b) les ***quartiles*** $X_{\frac{1}{4}}, X_{\frac{1}{2}}, X_{\frac{3}{4}}$: qui correspondent aux fréquences cumulées 25, 50, et 75%. Ils partagent l'ensemble des observations en 3 parties de même effectif.

c) les ***déciles*** $X_{\frac{1}{10}}, X_{\frac{2}{10}}, ..., X_{\frac{9}{10}}$: qui correspondent aux fréquences cumulées 10, 20, …, 90%.

d) les ***percentiles*** $X_{\frac{1}{100}}, X_{\frac{2}{100}}, ..., X_{\frac{99}{100}}$ qui correspondent aux fréquences cumulées 1, 2, …, 99%.

Les quantiles $X_{\frac{1}{4}}, X_{\frac{3}{4}}, X_{\frac{1}{10}}, X_{\frac{9}{10}}$ sont notamment intéressants pour étudier la dispersion et l'asymétrie d'une distribution.

7.5. Exercices sur les paramètres de position

1) Un étudiant désirant aller faire une partie de ses études aux USA doit présenter un examen d'anglais pour pouvoir être inscrit. Son résultat lui est communiqué sous la forme suivante: *8540 personnes ont présenté l'examen; vous êtes dans le septième intervalle interdécile.*

Pouvez-vous traduire ce renseignement en termes d'évaluation scolaire qui nous est plus familier ?

2) Si on vous dit que sur un ensemble de pommes qu'on a pesées, le premier décile est à 140 g et le 9ème décile à 160 g, quel est le pourcentage de pommes pesant entre 140 g et 160 g ?

3) Le résultat d'une enquête sur la taille d'enfants de trois ans est reprise dans le tableau suivant:

Tailles (cm)	f_{icum} (%)
[88,5; 90[	3
[90; 91,5[	10
[91,5; 93[	22
[93; 94,5[	36
[94,5; 96[	60
[96; 97,5[	77
[97,5; 99[	90
[99; 100,5[	97
[100,5; 102[	100

Trace un diagramme des fréquences cumulées.

Détermine sur le graphique la médiane, les troisième et sixième déciles ainsi que les quartiles.

4) Le père de Sébastien n'est pas content du résultat du dernier contrôle de son rejeton. Mais, argumente ce dernier, j'ai tout de même une moyenne de 12. Sachant que Sébastien a passé 5 contrôles et que les résultats des 4 premiers ont été 13, 10, 16 et 14, quelle est la cote du dernier travail ?

5) Fabienne rentre chez elle avec une copie cotée 11/20. Comme elle a l'habitude de meilleurs résultats ses parents demandent des explications. Fabienne argue de la difficulté des questions et de la sévérité de la correction du professeur ; ses parents restent sceptiques. Pour les convaincre, elle ajoute : avec ce 11/20, je suis encore dans la première moitié de la classe. A quelle notion statistique fait appel cette argumentation ?

6) Dans une classe de 21 élèves, 20 élèves ont participé à un contrôle dont les résultats ont été :{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8}. La moyenne de ces cotes est de 12. Le lendemain, l'élève absent, Alain est interrogé à son tour et il obtient 20. La moyenne de ce deuxième contrôle est donc de 20. La cote de 16, moyenne des deux jours ((12 +20)/2) représente-t-elle la moyenne de la classe ?
Combien d'autres "Alain" faudra-t-il adjoindre à cette classe pour que la moyenne soit effectivement de 16 ?

7) Analyser la courbe de croissance de la page suivante. (d'autres courbes sont disponibles sur le site www.xrenard.sup.fr, page math, 4SCOM, "courbes de croissance")

7.6. Quelques remarques sur les paramètres de position

- ***Dans le cas d'une variable qualitative***, on se limite généralement à déterminer le mode (dans certains cas, on peut aussi déterminer la médiane). Il n'est pas possible de calculer la moyenne.

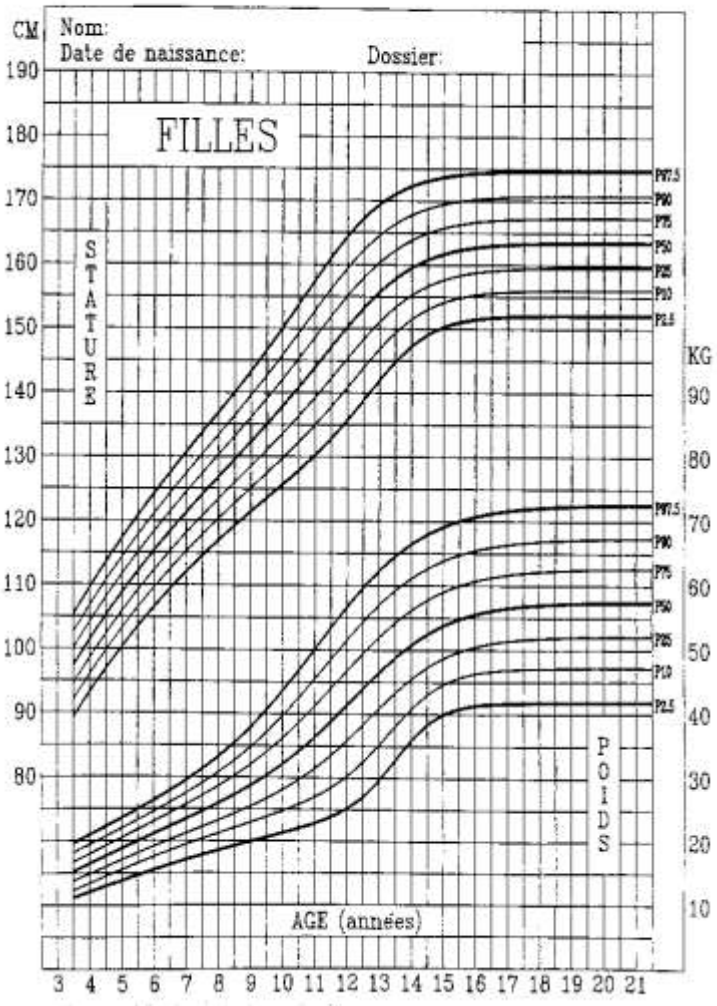
- ***Pour des variables quantitatives***, on peut déterminer soit la moyenne, soit le mode, soit la médiane.

Dans ce cas, le problème est de savoir quel paramètre il convient d'utiliser. Par exemple, considérons le cas d'une distribution de salaires moyens en Belgique. La moyenne est de 1398 € et la médiane de 1152 €, soit un écart de 246 € entre les deux. Si on examine la courbe des salaires, on s'aperçoit que si la moyenne est bien plus élevée que la médiane, c'est tout simplement parce que certaines personnes ont des revenus très élevés mais sont peu nombreuses. La moyenne est fortement influencée par ce petit nombre de personnes à revenu élevé alors que la médiane ne l'est pas. La moyenne vient donc fausser l'estimation de ce que gagne la plupart des belges. La médiane est, dans ce cas, un meilleur indicateur de tendance centrale et c'est ce paramètre que l'on choisira.

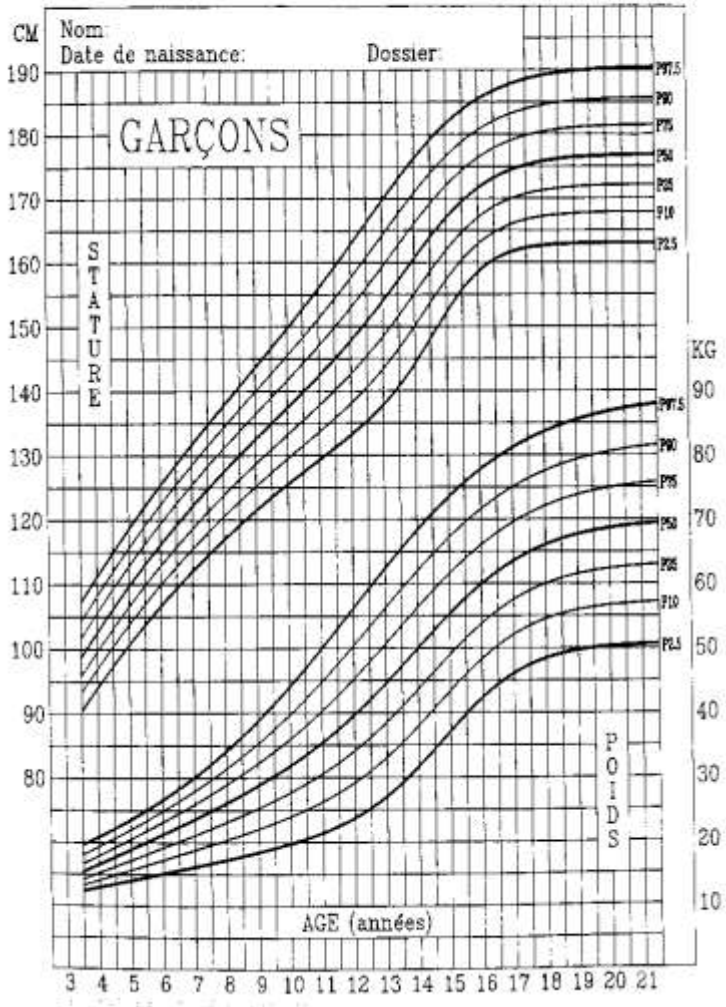
Un des avantages de la médiane est que sa valeur n'est pas influencée par des termes extrêmes, contrairement à la moyenne. Mais comme la médiane ne répond à aucune formule rigoureuse, elle ne peut être utilisée dans des calculs ultérieurs.

Remarquons que lorsque la moyenne et la médiane sont proches, cela témoigne d'une certaine symétrie de la distribution (comme dans le cas de données suivant une loi normale, voir plus loin)

POIDS ET TAILLE
FILLES - 3/21 ANS



POIDS ET TAILLE
GARÇONS - 3/21 ANS



8. LES PARAMETRES DE DISPERSION: L'ECART MOYEN, L'ECART-TYPE ET LA VARIANCE

8.1. Introduction

L'utilité des paramètres de position est d'indiquer d'une certaine manière autour de quelle valeur une série s'étend. Cependant, ce type de paramètre n'est pas suffisant pour caractériser une série.

En effet, prenons les trois exemples suivants :

a) Alice et Béatrice ont pris des vacances dans un endroit de rêve : des conditions de vie très proches de la nature (logement sous tente), mille et une possibilités d'activités sportives, de loisir, de détente, …, le tout à un prix intéressant. Le prospectus annonce pour la période envisagée, une température moyenne de 25°C et une brise légère.
Alice est revenue enchantée de ses vacances, Béatrice est très mécontente : un gros rhume, des mauvaises nuits fréquentes, des maux de tête, … lui ont gâché le séjour.

C'est que, alors que Béatrice, ayant lu très superficiellement le prospectus ne s'était équipée que de vêtements d'été, Alice au contraire avait lu que sur une journée, les écarts de température pouvaient être importants et que notamment, les nuits étaient plutôt fraîches. Elle avait donc encombré ses bagages de jeans et de petites laines qui lui ont permis de bien dormir et de ne pas frissonner en soirée et au petit matin.

b) Tout au long d'une année, on a fait des relevés statistiques d'un restaurant scolaire et on a trouvé qu'on servait en moyenne 100 repas complets par jour de fonctionnement du restaurant (4 jours par semaine), et que chaque repas comportait 100 g de pommes de terre. Ces données suffisent-elles à l'économe pour faire des achats hebdomadaires raisonnables de pomme de terre ?

c) Deux séries statistiques sont caractérisées par les données suivantes:

1.....100.....199

99 100 101

Si on calcule la moyenne de ces deux séries, on obtient:

$\overline{X} = (1 + 100 + 199) / 3 = 100$

$\overline{X} = (99 + 100 + 101) / 3 = 100$

Les moyennes sont les mêmes mais les séries sont très différentes. Dans la première série, la dispersion autour de la moyenne est très grande contrairement à la 2ème série !!!!

C'est pour cette raison que pour caractériser une série statistique, on doit définir des paramètres de dispersion en plus des paramètres de position. Ces paramètres de dispersion donnent des informations sur l' « étendue » de la série par rapport à la valeur centrale.

8.2. L'écart moyen, l'écart-type et la variance

Partons de l'exemple suivant : dans un carré de haricots, on a récolté 140 gousses (et non pas cosse qui est l'enveloppe du pois) et on a compté le nombre de grains dans chacune des gousses.

Voici le tableau de résultats.

Nombre de grains (variable discrète) X_i	1	2	3	4	5	6	7	8	9	10
Nombre de gousses effectifs n_i	2	6	9	18	32	38	20	7	6	2
Fréquences f_i	0,014	0,043	0,064	0,129	0,229	0,271	0,143	0,05	0,043	0,014

L'effectif total vaut: $n = \Sigma n_i = 140$

La moyenne vaut: $\overline{X} = \sum_{i=1}^{k} f_i.X_i = 5{,}51$ *(à vérifier comme exercice)*

Pour évaluer la dispersion autour de cette moyenne, l'idée qui vient spontanément à l'esprit est de déterminer les écarts entre cette moyenne et les diverses valeurs de la variable.

On appelle ***écart*** entre la valeur moyenne $\overline{X}$ et un nombre X, la valeur absolue de leur différence. On obtient pour la série étudiée:

Nombre de grains (variable discrète) X_i	1	2	3	4	5	6	7	8	9	10
Fréquences f_i	0,014	0,043	0,064	0,129	0,229	0,271	0,143	0,05	0,043	0,014
$\lvert X_i - \overline{X} \rvert$	4,51	3,51	2,51	1,51	0,51	0,49	1,49	2,49	3,49	4,49
$f_i . \lvert X_i - \overline{X} \rvert$										

En calculant la moyenne de la série des écarts, on obtient:

$\overline{\lvert X_i - \overline{X} \rvert} = \sum_{i=1}^{k} f_i.\lvert X_i - \overline{X} \rvert = 1{,}371$ Cette valeur s'appelle ***l'écart moyen*.**

En quelque sorte, on peut dire qu'en moyenne, le nombre de grains d'une gousse s'écarte, d'un côté comme de l'autre, de 1,371 grains de la valeur moyenne qui est de 5,51 grains.

Le défaut de l'écart-moyen est de donner la même importance à toutes les valeurs. Or, on a constaté qu'on obtient une meilleure mesure de la dispersion si on accorde plus de poids aux valeurs de la variable qui s'éloignent plus de la valeur moyenne. Au lieu de considérer les écarts, on considère les carrés des écarts. Plus l'écart est grand, plus son carré augmente. (puisqu'on élève les écarts au carré, la valeur absolue n'est plus nécessaire)

Nombre de grains (variable discrète) Xi	1	2	3	4	5	6	7	8	9	10
Fréquences fi	0,014	0,043	0,064	0,129	0,229	0,271	0,143	0,05	0,043	0,014
$\lvert Xi - \overline{X} \rvert$	4,51	3,51	2,51	1,51	0,51	0,49	1,49	2,49	3,49	4,49
$(X_i - \overline{X})^2$	$(-4{,}51)^2$	$(-3{,}51)^2$	$(-2{,}51)^2$	$(-1{,}51)^2$	$(-0{,}51)^2$	$(0{,}49)^2$	$(1{,}49)^2$	$(2{,}49)^2$	$(3{,}49)^2$	$(4{,}49)^2$
$f_i . (X_i - \overline{X})^2$										

La moyenne de la série des carrés des écarts vaut: $\overline{(X_i - \overline{X})^2} = \text{fi} .(X_i - \overline{X})^2 = 3{,}078.$

Elle est appelée la ***variance*** et elle a la même dimension que les carrés des valeurs de la variable.

On ne peut donc pas la comparer à l'écart-moyen. Pour cela il faut prendre la racine carrée de la variance que l'on appellera ***l'écart-type***.

$$\text{Ecart-type} = \sigma = \sqrt{\sum_{i=1}^{k} f_i.(X_i - \overline{X})^2} = \sqrt{3{,}078} = 1{,}75 \ldots$$

On peut dire que la plupart des gousses contient un nombre de grains compris entre 5,51 - 1,75 et 5,51 + 1,75. (environ 108 des 140 gousses sont comprises dans cet intervalle).

Remarque

La distribution de nombreux paramètres biologiques et autres (taille, poids, mensurations diverses, pouls, quotient intellectuel, mais aussi par exemple la dimension d'objets fabriqués en série ...) suit une loi dite normale (que nous étudierons dans la suite), représentée par une courbe en cloche symétrique autour de la moyenne.

Si la distribution de fréquence suit approximativement une loi normale :

- 68,27 % des éléments ont une valeur comprise entre $\overline{X} - \sigma$ et $\overline{X} + \sigma$,
- 95,45 % des éléments ont une valeur comprise entre $\overline{X} - 2\sigma$ et $\overline{X} + 2\sigma$,
- 99,73 % des éléments ont une valeur comprise entre $\overline{X} - 3\sigma$ et $\overline{X} + 3\sigma$.

L'intervalle entre $\overline{X} - 2\sigma$ et $\overline{X} + 2\sigma$ est appelé intervalle de référence (ou de tolérance) : seulement 5 % des sujets tombent en dehors des limites.

Par exemple, pour déterminer les valeurs de référence de l'urée (mmol/l), un échantillon de 284 sujets ont subi une prise de sang pour mesurer leur concentration d'urée. La moyenne obtenue est de 5,1 mmol/l et l'écart-type 1,1 mmol/l. Les limites de l'intervalle de référence sont donc $5,1 \pm 2.1,1$, soit 2,9 et 7,3 mmol/l.

Un résultat est considéré comme normal s'il tombe dans l'intervalle de référence. Il est très rare que l'on tombe en dehors de l'intervalle lorsqu'on est en bonne santé.

RESUME

Dans le cas d'une variable quantitative, on appelle:

- ***écart***: la valeur absolue de la différence entre la moyenne et une valeur de la variable.
- ***écart moyen***: la moyenne de la série des écarts (voir plus haut) de tous les individus de la population.
- ***variance***: la moyenne de la série des carrés des écarts entre la moyenne et les valeurs de la variable de tous les individus de la population.
- ***l'écart-type***: la racine carrée positive de la variance. C'est l'écart-type qui est généralement le plus utilisé.

$$\text{Ecart-type} = \sigma = \sqrt{\sum_{i=1}^{k} f_i.(X_i - \overline{X})^2}$$

On utilise souvent une formule plus pratique qui donne l'écart-type avec une bonne approximation:

$$\text{Ecart-type} = \sigma = \sqrt{\sum_{i=1}^{k} f_i.X_i^2 - \overline{X}^2}$$

9. POPULATION ET ECHANTILLON

Ce n'est généralement qu'en théorie qu'on travaille sur une bande de données complète (la population entière) et donc $\overline{X}$ et σ sont exacts.

En pratique, on extrait et on analyse un échantillon de la population.

Pour l'écart-type de l'échantillon, on utilise la formule :

$$\sigma_{\text{échantillon}} = s = \sqrt{\left(\frac{\sum_{i=1}^{k} n_i.(X_i - \overline{X})^2}{n-1} \right)}$$

Il est évident que plus n est grand, plus l'écart-type de l'échantillon se rapproche de l'écart-type de la population $\sigma_{population}$.

Ces différences entre population et échantillon seront précisées dans le chapitre consacré à l'estimation.

Signalons cependant que :

- *Pour la population entière, la moyenne est notée μ, la variance σ² et l'écart-type σ.*
- *Pour un échantillon, la moyenne est notée* $\overline{X}$ *et la variance s² et l'écart-type s.*

Pour simplifier, en statistiques descriptives, on désignera la moyenne par $\overline{X}$ et l'écart-type par σ peu importe qu'il s'agisse d'un échantillon ou de la population et on utilisera la formule : $\sigma = \sqrt{\sum_{i=1}^{k} f_i.(X_i - \overline{X})^2}$.

Par contre, dans le chapitre consacré à l'estimation, on fera la différence entre les écart-types et moyennes de la population et de l'échantillon.

Remarque

Sur les claviers de beaucoup de calculatrices ayant les fonctions statistiques, on trouve les symboles σ_n et σ_{n-1} (parfois "s"). La fonction σ_n correspond à l'écart-type σ défini sur l'ensemble de la population. Dans les analyses d'échantillon, on utilise l'écart-type de l'échantillon σ_{n-1}.

Si la calculatrice n'a que le symbole σ, il faut vérifier, en consultant le mode d'emploi ou par un exemple simple, de quel écart-type il s'agit.

Exemple : soit la population {10, 15, 20} dont la moyenne est 15 ; σ_n = 4, 08, et σ_{n-1}=5.

Dans le logiciel Excel, σ_n est donné par la fonction ECARTYPEP (P pour population) et σ_{n-1} est donné par ECARTYPE. On a de même les fonctions VARP et VAR pour la variance.

10. LE COEFFICIENT DE VARIATION

Lorsque deux distributions sont analogues (par exemple, la taille en cm d'enfants de 8 ans et la taille en cm d'adultes de 25 ans), il est facile de comparer leurs dispersions et donc leurs écart-type.

Il pourrait cependant être utile de comparer des distributions issues d'échelles de grandeurs différentes ou d'unités de mesure différentes (par exemple un test dont les notes varient de 0 à 100 et un autre variant de 0 à 20). Dans ce cas, on peut utiliser un coefficient de variation qui permet de ramener la valeur de n'importe quelle dispersion sur une même échelle, en l'occurrence un pourcentage. La valeur de ce coefficient est donnée par la formule :

$$\text{Coefficient de variation (en \%)} = C_V = \frac{\text{écart-type}}{\text{moyenne}}.100 = \frac{\sigma}{\bar{X}}.100$$

Exemple :

La taille moyenne des jeunes de 18 ans est de 168 cm avec un écart-type de 14 cm; leur poids moyen est de 66 kg avec un écart-type de 8 kg. La taille varie-t-elle plus que le poids ?

$$C_V \text{ (taille)} = \frac{14}{168}.100 = 8{,}33..\%$$

$$C_V \text{ (poids)} = \frac{8}{66}.100 = 12{,}1..\%$$

Les poids varient plus que les tailles.

11. EXERCICES DIVERS

EXERCICE 1

On a relevé durant 30 jours la température extérieure, à midi (température exprimée en degrés Celsius):

12	10	11	13	15	16	16	17	19	18
19	17	16	15	14	17	19	18	19	21
22	21	21	23	22	24	25	27	26	24

Détermine :
a) la variable
b) l'effectif total
c) la nature de la variable: quantitatif ou qualitatif ? Discret ou continu ?
d) Dresse les diagrammes suivants:
- histogramme des fréquences
- diagramme des fréquences cumulées croissantes
- diagramme en camembert

e) Calcule la moyenne, le mode et la médiane. ***(Rép. :*** $\overline{X}$ ***=18,5)***
f) Calcule l'écart-type et la variance. ***(Rép. : 4,3 et 18,6)***

EXERCICE 2

Une série d'observations concernant les tailles d'un groupe d'adolescents de 11 à 14 ans a donné les résultats suivants:

Taille X_i	effectifs n_i	Taille X_i	effectifs n_i
140 < X ≤ 144 cm	3	160 < X ≤ 164 cm	31
144 < X ≤ 148 cm	17	164 < X ≤ 168 cm	20
148 < X ≤ 152 cm	63	168 < X ≤ 172 cm	4
152 < X ≤ 156 cm	82	172 < X ≤ 176 cm	1
156 < X ≤ 160 cm	69	176 < X ≤ 180 cm	1

a) Trace l'histogramme des effectifs
b) Trace le polygone des fréquences cumulées croissantes et en déduire la médiane.
c) Détermine le mode et la moyenne ***(Rép. :*** $\overline{X}$ ***= 155,5)***
d) Calcule l'écart-moyen, l'écart-type et la variance ***(Rép. :*** $\sigma = 5,8$)

EXERCICE 3

Voici, relevées au cours des jours ouvrables de l'année 1982, les recettes d'un magasin de détail:

Recettes en milliers de FB	Nombre de jours
0 ≤ X < 4	8
4 ≤ X < 8	24
8 ≤ X < 12	210
12 ≤ X < 16	42
16 ≤ X < 20	16

a) Calcule : - l'effectif total
- les fréquences en degrés

b) Trace : - un histogramme des fréquences
- un polygone des fréquences cumulées croissantes

c) Calcule la moyenne, le mode, l'écart-type et la variance. ***(Rép.:*** $\overline{X}$ ***= 10,4 et*** $\sigma = 2,9$ ***)***

EXERCICE 4

Voici les tailles de 34 garçons et 37 filles inscrits en 3ème année dans une école.

Les garçons mesurent (en cm) :

152	177	168	171	165	174	173	168	176	163	165	165	162
175	168	163	160	183	175	169	164	181	165	163	155	164
170	165	163	176	174	165	190	160					

Les filles mesurent (en cm) :

160	165	165	168	150	166	165	167	166	156	163	163	174	166
176	160	165	161	165	157	168	169	168	150	165	162	172	160
165	171	163	166	164	181	180	150	155					

Classe ces données dans deux tableaux.

1) Réalise un histogramme des effectifs pour les garçons.
2) Réalise un histogramme des effectifs pour les filles en conservant la même échelle que pour les garçons.
3) Compare les 2 histogrammes. Que constate-t-on ?
4) Calcule les moyennes des tailles des garçons et des filles en indiquant la formule utilisée.
5) Que constate-t-on ?
6) Réalise un polygone des fréquences cumulées (de haut en bas) pour les garçons.
7) Indique sur le graphique le pourcentage de garçons ayant une taille inférieure ou égale à 170 cm.
8) Réalise un diagramme en camembert pour les filles.
9) Calcule l'écart-type pour les garçons en indiquant la formule utilisée. Que signifie la valeur trouvée ?

EXERCICE 5 : Lors d'un rallye touristique, une des épreuves du parcours consiste à évaluer la distance qui sépare l'endroit où on se trouve d'un autre endroit identifiable (ou d'évaluer la hauteur d'un arbre ou d'un bâtiment, ou le poids d'un objet, …). Certaines équipes font la moyenne de l'évaluation de chaque membre de l'équipe. Cette approche statistique est-elle judicieuse ?

EXERCICE 6 : Un instituteur remet à ses élèves une feuille sur laquelle est dessiné un quadrilatère quelconque et demande à chacun de faire les mesures et les calculs nécessaires pour connaître l'aire de la figure. On devine qu'à la collecte des résultats, on observe des différences généralement plus grandes que ce qu'on pourrait croire. Quel nombre prendre pour la mesure de cette surface ?

EXERCICE 7 : Les statistiques montrent qu'on constate un plus grand nombre d'accidents mettant en cause des voitures roulant à vitesse modérée. Peut-on en conclure qu'il est recommandé de rouler comme un fou ?

EXERCICE 8 : Les statistiques montrent qu'il est faux de dire que les mathématiques constituent un très grand facteur de redoublements car c'est dans les classes où le nombre d'heures de math est le plus élevé que l'on constate le moins de redoublements à imputer à cette matière. Qu'en pensez-vous ?

EXERCICE 9 : Les propositions suivantes sont-elles vraies ou fausses ?

La moyenne d'une série de données statistiques divise les données en deux parts égales.

La moyenne arithmétique est influencée par des valeurs extrêmes d'une série de données.

Entre le premier quartile et le troisième quartile se trouvent toujours 50 % des effectifs.

L'effectif compris entre le premier et le deuxième quartile est toujours le même que celui compris entre le deuxième et le troisième.

Quelle que soit la série de données statistiques, la somme des écarts par rapport à la moyenne arithmétique est nulle.

EXERCICE 10

Voici un tableau reprenant les durées de vie de tubes électriques :

Durée (heures)	n_i
[201; 300[	6
[300; 399[	8
[399; 498[	46
[498; 597[	58
[597; 696[	76
[696; 795[	68
[795; 894[	62
[894; 993[	48
[993; 1092[	22
[1092 ; 1191[	4
[1191; 1290[	2

On donne :

$\overline{X} = 710{,}4$

$\sigma = 192{,}7$

Déterminer le pourcentage de tubes dont la durée de vie est :

a) "Normale"
b) Inférieure à la "normale"
c) Supérieure à la "normale"

Remarque :

Les éléments sont "normaux" s'ils sont compris entre $\overline{X} - \sigma$ et $\overline{X} + \sigma$, inférieurs à la normale s'ils sont compris entre $\overline{X} - \sigma$ et $\overline{X} - 2\sigma$ et supérieurs à la normale s'ils sont compris entre $\overline{X} + \sigma$ et $\overline{X} + 2\sigma$.

(Rép. : 64,22 %, 15,87 %, 16,44 %)

EXERCICE 11

Soit la taille de 210 étudiants :

Classes	***effectifs***
[155; 160[	5
[160; 165[	23
[165; 170[	42
[170; 175[	68
[175; 180[	47
[180; 185[	21
[185; 190[	4

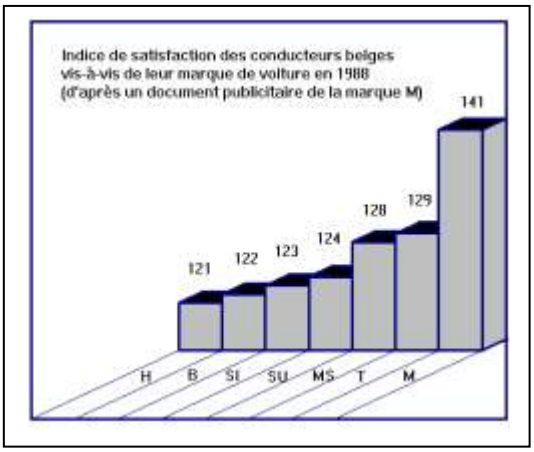
Déterminer le pourcentage d'étudiants dont la taille est :

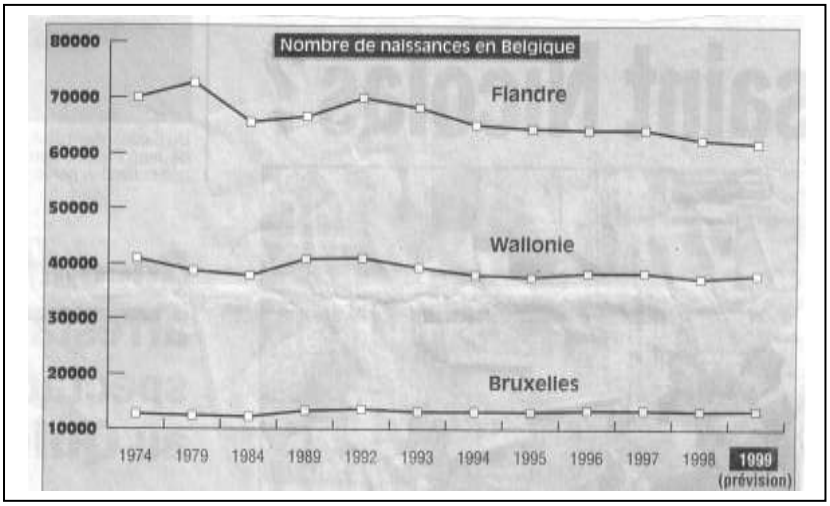
a) "Normale"
b) Inférieure à la "normale"
c) Supérieure à la "normale"

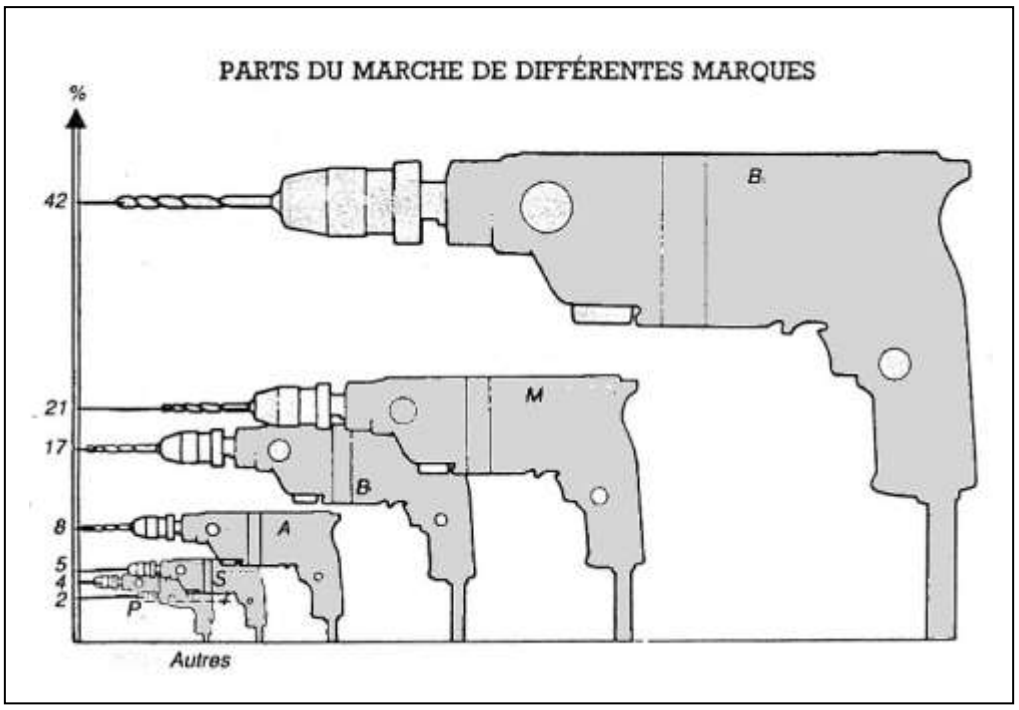
(Rép. : $\overline{X} = 172{,}5$, $\sigma = 6{,}4$; *65,6 %, 15,3 %, 15,2 %)*

12. ***PIEGES STATISTIQUES SOUS FORME DE GRAPHIQUES***

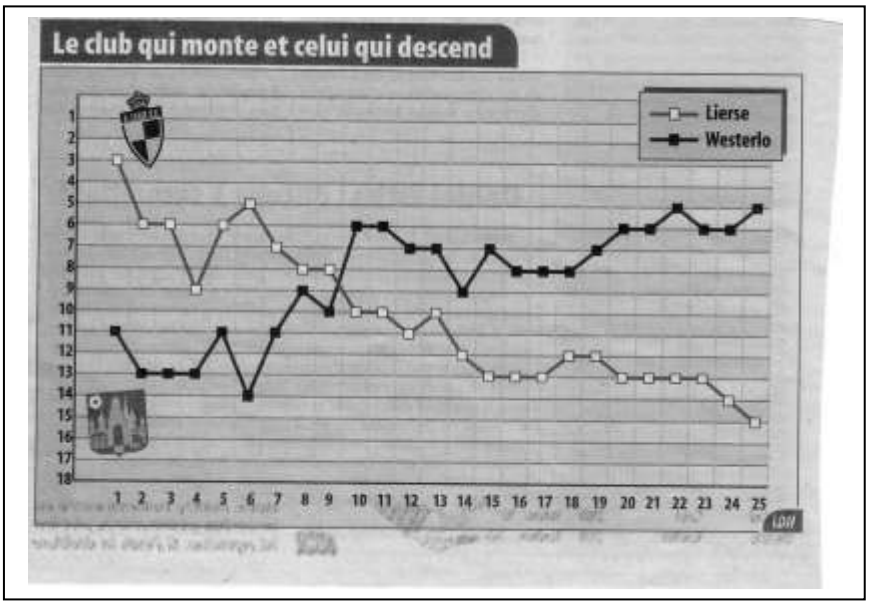
Graphique issu de la publicité de la marque M







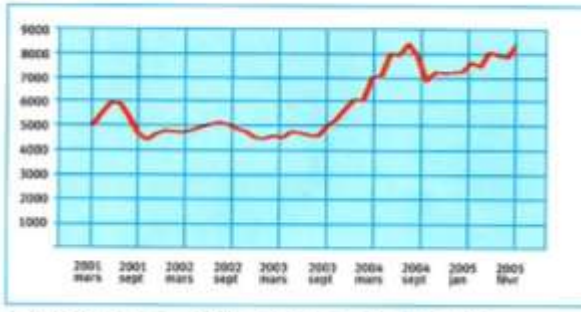
Voici un graphique comparant les parts de marché de différentes foreuses. Commente...



Le DJ Eurostoxx 50 est un indice de référence qui reprend les 50 principales valeurs de la zone euro. Les secteurs les plus représentés sont la finance, l'énergie et les télécommunications. Sa répartition géographique comprend notamment la France, l'Allemagne et les Pays-Bas. Les 5 principales actions de cet indice sont Total, Royal Dutch, Banco Santander, Telefonica et Sanofi.

Le FTSE Xinhua China 25 permet d'accéder au plus important nouveau marché du monde. Il est composé des 25 actions chinoises les plus significatives cotées à la bourse de Hong Kong. Parfaitement équilibré et diversifié, le fonds couvre tous les grands secteurs d'activités (télécommunications, énergie, transport, banque et assurance, industries, etc).

Le FTSE Xinhua China 25



Evolution de l' indice FTSE Xinhua China 25 ces 5 dernières années.

Le Hit-parade des jouets en 1998

		Part des ventes en 1998	Evolution par rapport à 1997
1.	Jeux de construction	22,8 %	-4 %
2.	Jeux de société	20,1 %	-11,8 %
3.	Jouets pour bébés	13,3 %	-28,3 %
4.	Jouets "à thèmes" (à l'effigie de personnages de télé, ciné,...)	11 %	+39,2 %
5.	Jouets informatiques, vidéos...	8,3 %	+3,4 %
6.	Trains électriques et circuits auto	7 %	-5 %
7.	Poupées et peluches	4,7 %	+182 %
8.	Puzzles	4,2 %	+48,6 %
9.	Bricolage	3,5 %	-0,4 %
10.	Maquettes, modèles réduits	2,2 %	+22,7 %

DOSSIER

dossier

Productivité ou compétitivité

Il faut distinguer deux concepts: productivité et compétitivité. Selon que l'on négocie les salaires en fonction de l'un ou de l'autre, il y aura une répercussion à la hausse, ou à la baisse sur notre pouvoir d'achat.

Tentative de décryptage :

1. Dans l'après-guerre, nous négociions les salaires en ayant à l'esprit les gains de productivité. Dans un premier temps, nous estimions les coûts des entreprises, dont les salaires ne sont qu'une partie. Il s'agissait alors de prendre en compte l'argent que toutes ces dépenses ont rapporté. Ensuite, nous nous partagions ces gains dits de productivité. Et ce qu'il faut rappeler, c'est que ces « gains de productivité » existent autant qu'hier et moins que demain ! En moyenne toutes les entreprises de Belgique sont productives. Elles créent chaque année plus de richesses. Il existe encore et toujours des bénéfices. Nous ne sommes donc pas en crise, loin de là.

2. Depuis 1996 nous sommes dans un cadre de compétitivité. Ce n'est plus l'ensemble des coûts de l'entreprise qui est pris en compte mais le seul coût salarial. Dire, nous ne négocions plus le salaire en fonction de l'activité économique de l'entreprise installée sur notre territoire, nous sommes dans un cadre de comparaison des coûts salariaux avec les autres pays. Il ne s'agit pas de produire plus et mieux, mais plus pour moins cher que les autres. Les autres, ce sont nos proches voisins mais aussi les voisins plus éloignés -Chine, Inde, Brésil- qui sont, de fait, plus « compétitifs » en terme de salaires. Pourtant, en terme de productivité, ce n'est pas toujours aussi évident ! N'oublions pas que s'il y a salaire plus bas, c'est à cause de la faiblesse des règles sociales et/ou des syndicats locaux. Tout est question de rapport de force social... local et international. C'est que là-bas plus qu'ailleurs, la pratique syndicale est dangereuse, y compris pour sa propre vie.

Comment expliquer l'obsession patronale de compétitivité ?

Assez simplement. En se plaignant des coûts salariaux, vos patrons espèrent diminuer vos salaires et ainsi gonfler le leur... La preuve, dans notre gâteau commun de richesse nationale, depuis le début des années 1980 les salaires ont une part qui ne cesse de diminuer, au profit des revenus dit du capital (actionnaires, etc.). Mais comment ? En jouant sur la notion de compétitivité à la place de celle de productivité. Si vous êtes dans un cadre de comparaison constante des salaires avec les pays voisins, le seul résultat que vous aurez est celui d'une baisse des coûts salariaux chez vous, mais aussi chez vos voisins, qui évidemment font la même chose. Résultat, nous assistons à une drôle de bataille entre les Etats, dictée par les grandes firmes, pour savoir qui pourra produire des biens et services avec le moins de coûts salariaux, avec le moins d'imposition et donc avec le moins de revenus et de protection sociale pour les salariés... Chacun cherche à exporter son chômage chez les autres. C'est totalement inefficace et cette compétition à la baisse explique que, malgré la montée des besoins sociaux, nous dépensons en proportion plus ou moins autant maintenant qu'au début des années 1980 pour notre sécurité sociale. C'est ce mécanisme de comparaison à la baisse qui a pour résultat que la part des salaires dans le PIB, soit la répartition des revenus, qui était début 1980 à 67,2%, est passée en 1990 à 58,6% et en 2004 à +- 56,5%. (1) Voir tableau 1.

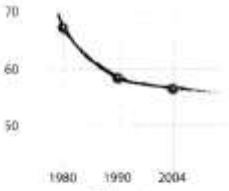


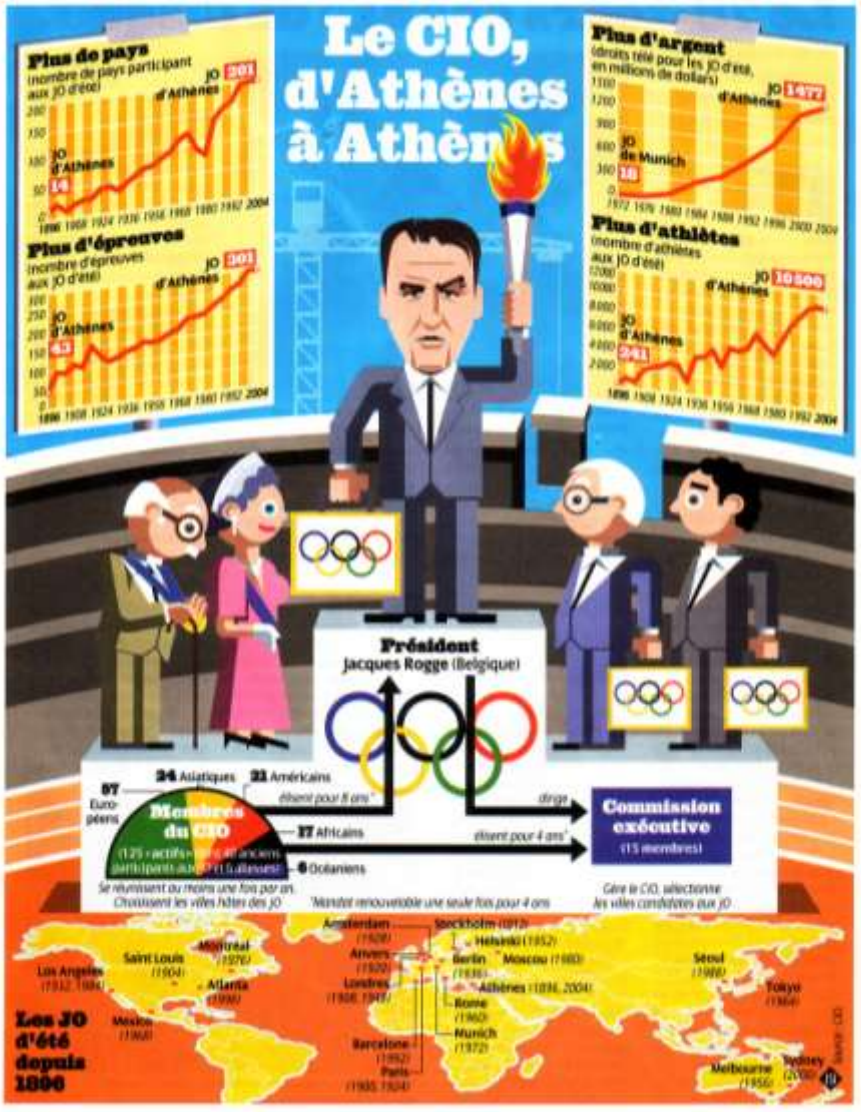
Tableau 1

Et cela se reflète aussi au niveau de l'Europe des 15 puisque d'après la commission européenne, dans l'ensemble des revenus, la part des salaires est tombée de +-70,5% en 1990 à +-67,7% en 2004. Donc, partout une tendance à la baisse des salaires... Voir tableau 2.

Tableau 2

Dans le cadre de pensée libéral de compétitivité, les salaires ne sont jamais assez bas puisque la référence est la profitabilité du capital investi ailleurs dans le monde. Il nous faut donc rappeler que la Belgique est un pays producteur de richesse. Dans son

LA BALISE



LE NOUVEL EXPRESS 15/8/2004 ● 65

CHAPITRE 2 : LES PROBABILITÉS,LES VARIABLES ALÉATOIRES ET LES LOIS DE PROBABILITÉ

1. PROBABILITES : DEFINITIONS

Un ***phénomène fortuit*** est une expérience qui donne lieu à plusieurs résultats dont on ne peut prédire à l'avance lequel se réalisera. Chacun de ces résultats porte le nom d'***épreuve*** du phénomène.

Exemples de phénomènes fortuits : tous les jeux de hasard, la transmission des caractères héréditaires dans les espèces (animales ou végétales) à reproduction sexuée, la physique des micro-particules, ...

L'ensemble de toutes les épreuves possibles d'un phénomène fortuit se nomme ***catégorie d'épreuves*** Ω du phénomène. Par exemple, si on tire cinq fois à pile ou face, on peut obtenir :

Épreuve 1 : pile
Épreuve 2 : face
Épreuve 3 : face
Épreuve 4 : pile
Épreuve 5 : pile

Dans ce cas, il n'y a que deux résultats possibles. La catégorie d'épreuve est : Ω = {pile, face}.

On pourrait se demander quelle est la probabilité d'avoir cinq fois "face" ou bien d'avoir deux "face" suivi de trois "pile", … On parle dans ce cas d'***événement***.

La **probabilité** P(E) qu'un événement E se produise est donnée par la formule :

$$P(E) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$$

qui est un nombre compris entre 0 et 1.

Deux ou plusieurs événements sont ***équiprobables*** s'ils ont la même probabilité d'apparition. (lancé d'un dé, d'une pièce de monnaie, ...)

Exemple : On lance deux fois de suite une pièce de monnaie. Quelle est la probabilité d'avoir l'événement "face" puis "face" ? (les événements "pile" et "face" sont équiprobables)

Nombre de cas favorables : 1 **Nombre de cas possibles** : 4

$\Rightarrow$ P (face puis face) = ¼

2. LES VARIABLES ALEATOIRES (V.A.)

2.1. Définition

Voici quelques exemples de définitions d'une variable aléatoire. Si vous ne comprenez pas la première, lisez la suivante et ainsi de suite …

Une variable aléatoire, notée X, est une variable susceptible de prendre des valeurs diverses (événements) en obéissant à une loi de probabilité déterminée (ou distribution de probabilité).

OU

Une variable aléatoire est une variable dont la valeur est un nombre déterminé par l'issue d'une épreuve.

OU (encore …)

Une variable aléatoire est une variable dont on ne peut prédire avec certitude, avant la mesure, le résultat qui surviendra.

OU (plus simple)

Une variable aléatoire est « quelque chose » dont il est impossible de connaître le résultat à l'avance.

En résumé, une variable aléatoire est une variable dont les valeurs dépendent du hasard. Le poids ou la taille ne sont pas des variables aléatoires mais si on les mesure sur un sujet tiré au hasard d'une population, ils le deviennent aussitôt.

Il y a deux grands types de variables aléatoires :

- on parle de « ***variable aléatoire discrète*** » si l'ensemble de ses valeurs est fini ou quand la catégorie d'épreuve comprend un nombre fini de valeurs.
- on parle de « ***variable aléatoire continue*** » si l'ensemble de ses valeurs est infini ou quand la catégorie d'épreuve est un ensemble continu. Dans ce cas, la variable peut prendre n'importe quelle valeur dans un intervalle.

Exemples :

- Le nombre de pannes journalières d'une machine est une ***variable aléatoire discrète***.
- Le nombre de pile obtenu lors de 5 lancers consécutifs est une ***variable aléatoire discrète***.
- Le poids réel de paquets de lessive pesant en principe 5 kg est une ***variable aléatoire continue.***
- La durée de vie d'un moteur soumis à des conditions difficiles est une ***variable aléatoire continue***.

En sciences, on manipule toute sorte de variables aléatoires, par exemple : le rendement d'une culture, la densité d'un matériau, le temps nécessaire pour accomplir une tâche, le nombre de pétales d'une fleur, la résistance à la flexion, le temps de réaction après un stimulus, ...

3. LES VARIABLES ALEATOIRES DISCRETES

Une variable aléatoire discrète X est une variable dont toutes les valeurs X_i sont connues et à chacune desquelles on peut attacher une probabilité de réalisation $P(X_i)$. ***La loi (ou distribution) de probabilité*** représente l'ensemble des probabilités $P(X_i)$ correspondant à chaque valeur de la variable aléatoire X.

Prenons l'exemple suivant.

Un vendeur de téléviseurs présente la synthèse du nombre d'articles vendus chaque jour au cours des 100 derniers jours de vente.

indice i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	
Nombre de TV vendues chaque jour	*0*	*1*	*2*	*3*	*4*	*5*	*6*	
Nombre de jours de vente	*2*	*8*	*20*	*25*	*30*	*12*	*3*	*Total = 100*

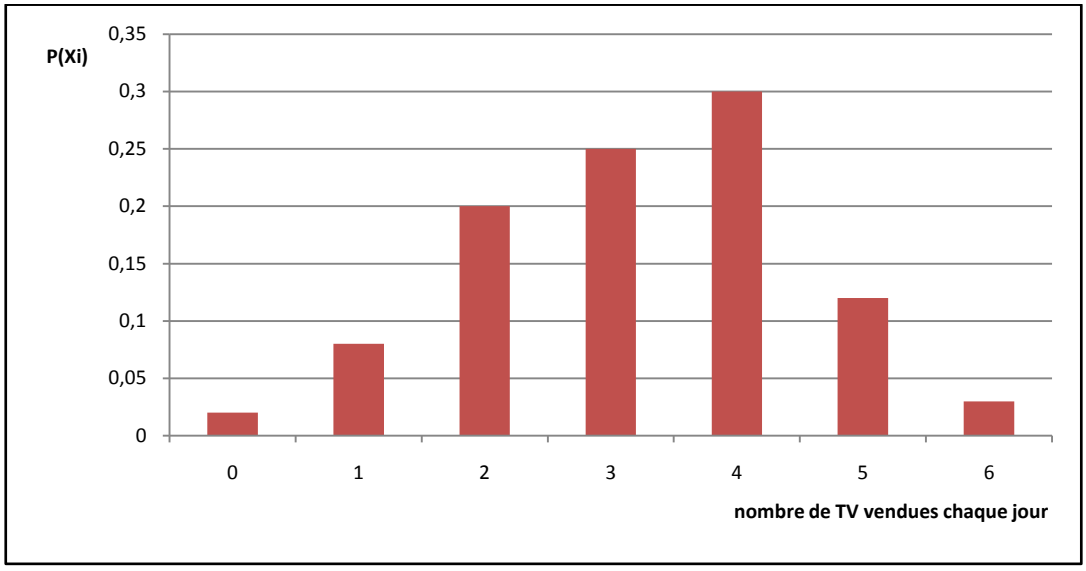
Soit X la variable aléatoire donnant « le nombre de TV vendues au cours d'une journée ».

A partir de ces données, on peut calculer les fréquences et considérer que chaque fréquence f_i correspond à une probabilité de réalisation $P(X_i)$. On obtient la distribution de probabilités suivante :

Nombre de TV vendues chaque jour X	*0*	*1*	*2*	*3*	*4*	*5*	*6*	
P(Xi)	*0,02*	*0,08*	*0,2*	*0,25*	*0,3*	*0,12*	*0,03*	$\sum_{i=1}^{7} P(X_i) = 1$

On peut évidemment tracer un diagramme en bâtonnets, un diagramme en escaliers, … (voir cours de statistique descriptive)

La loi (ou distribution) de probabilité est représentée sur le graphique suivant :



3.1. <u>Calcul de l'espérance mathématique.</u>

L'espérance mathématique est une caractéristique de position. Elle correspond à la moyenne arithmétique en statistique descriptive. Par analogie, on a :

$$E(X) = \sum_{i=1}^{k} X_i . P(X_i)$$

On obtient pour notre exemple : E(X) = 3,21.

Cela veut dire que sur une longue période, la vente quotidienne est de 3,21 TV.

3.2. ***Calcul de l'écart-type et de la variance.***

Par analogie avec la variance et l'écart-type d'une variable normale *(voir statistiques descriptives)*, on obtient :

L'écart-type : $\sigma(X) = \sqrt{\sum_{i=1}^{k} P(X_i).(X_i - E(X))^2} = 1,3$ La variance : $V(X) = \sigma^2(X) = 1,7$

3.3. ***Exercices supplémentaires***

1) 1000 billets sont vendus par une association qui organise une loterie. Parmi ceux-ci, 250 billets sont gagnants, 10 donnent droit à un lot de 75 €, 30 à un lot de 40 €, 80 à un lot de 18 € et 130 à un lot de 5 €.

   a) Détermine la variable aléatoire et la loi de probabilité qui en résulte.
   b) Un membre de l'association achète un billet. Quelle est la probabilité qu'il gagne un lot dont le montant peut aller jusque 18 €, un lot inférieur à 18 €, un lot d'un montant minimum de 5 €, un lot de 40 € ou de 75 €, un lot supérieur à 40 € et inférieur à 75 €, un lot d'un montant minimum de 40 € et inférieur à 75 € ?
   c) Calcule E(X), V(X) et σ(X). ***(E(X) = 4,04; σ(X) = 10,8)***
   d) En supposant que l'association veuille s'autofinancer totalement pour cette action, quel prix de vente minimum du billet faudra-t-il fixer ?

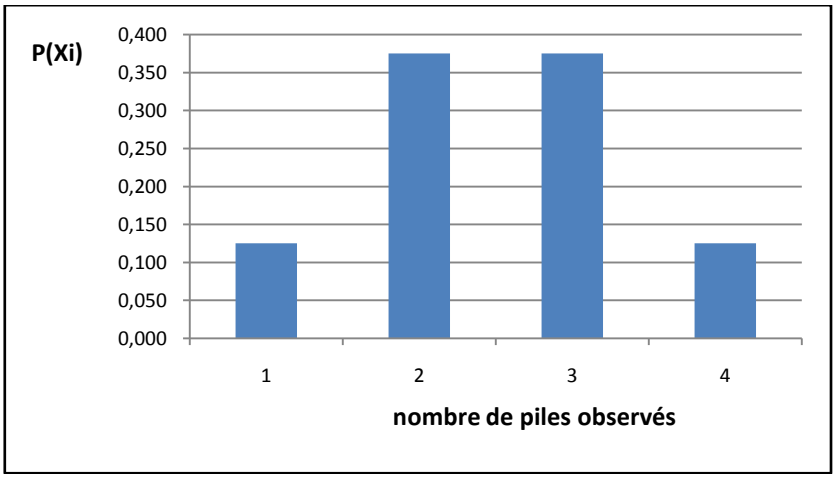
2) On lance 2 dés à jouer et on fait à chaque lancer la somme des points obtenus. Soit X la V.A. représentant cette somme. On suppose que les dés sont non pipés et ont 6 faces.

   a) Détermine la loi de probabilité de X.
   b) Calculer E(X), V(X) et σ(X).
   (E(X) = 7; σ(X) = 2,4)

3) On lance une pièce de monnaie bien équilibrée trois fois de suite. On note le "nombre de pile" observé.

   a) Quelle est la variable aléatoire ?
   b) Détermine la loi de probabilité (c'est-à-dire toutes les probabilités $P(X_i)$).
   c) Calcule l'espérance mathématique E(X), la variance V(X) et l'écart-type σ (X).
   (E(X) = 1,5; σ(X) = 0,87)
   d) Quelle est la signification de l'écart-type ?

Pour vous aider, on a représenté ci-après la loi de probabilité de cette variable aléatoire :



4) On lance 2 dés (1 rouge et un bleu) et on calcule le total des points.

Sur une mise de 50 €, ***si le total ≥ 9, on gagne 200 €.***
si le total < 9, on perd la mise.

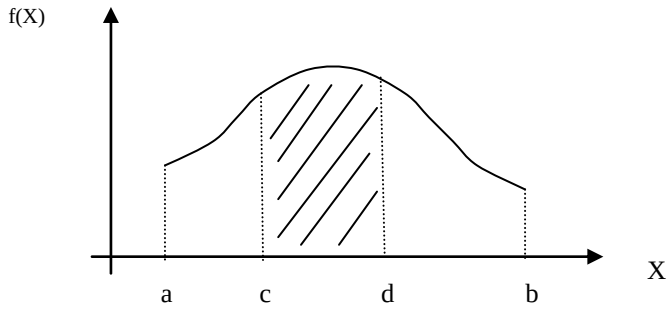
Calculer l'espérance mathématique des gains, la variance et l'écart-type des gains (et donner sa signification).

(5,6; 8024,7; 89,6)

4. *LES VARIABLES ALEATOIRES CONTINUES*

Une variable aléatoire X est dite continue si l'ensemble des valeurs prises par cette variable est infini non dénombrable, c'est-à-dire que X peut prendre n'importe quelles valeurs réelles comprises dans un intervalle [a, b].

Par conséquent, la probabilité que X prenne une valeur particulière quelconque est généralement nulle. On calculera plutôt la probabilité que X se trouve dans un intervalle donné.



L'ensemble des valeurs possibles de la variable aléatoire continue X et sa loi de probabilité (fonction de densité de probabilité de X) est représentée par une courbe qui peut avoir des allures différentes selon le type de distribution étudiée.

L'aire comprise entre la courbe, les 2 verticales d'abscisses a et b et l'axe des abscisses est égale à 1. Par conséquent dans l'intervalle [a, b], la somme des probabilités attachées aux valeurs de la variable est égale à 1.

D'une façon générale, la mesure d'une probabilité peut être traduite graphiquement par une surface. En considérant que la probabilité de se trouver dans l'intervalle [a, b] vaut 1, si on choisit un intervalle [c, d] inclus dans [a, b], la probabilité de se trouver dans l'intervalle [c, d] est inférieure à 1.

La différence fondamentale avec les variables aléatoires discrètes est l'utilisation de la notion d'intervalle de valeurs de X auxquelles sont liées les probabilités (plutôt qu'à des valeurs particulières et isolées).

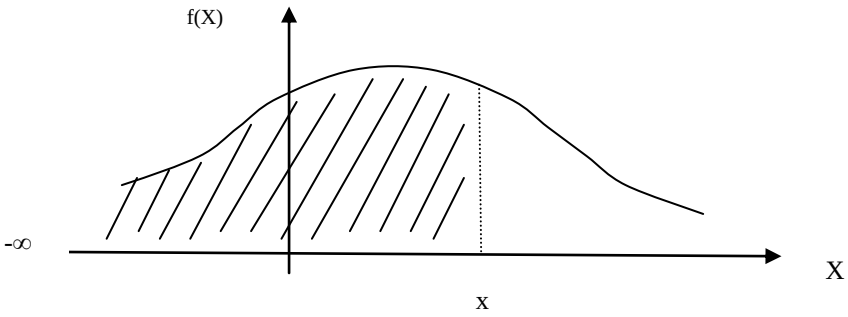
4.1. *Fonction de répartition*

On appelle ***fonction de répartition*** de la variable aléatoire continue X, la fonction F(x) qui à tout x de l'intervalle de définition fait correspondre la probabilité que X soit strictement inférieure à x :

$$F(x) = P(X<x)$$

Pour un intervalle de définition de]-∞, x[, on peut dire que $F(x) = \int_{-\infty}^{x} f(X).dX$ représente l'aire sous la courbe f(x) de -∞ à x.

Avec bien sûr : $\int_{-\infty}^{+\infty} f(X).dX = 1$.



On a de plus : $P(a \le X \le b) = F(b) - F(a) = \int_{a}^{b} f(X).dX$

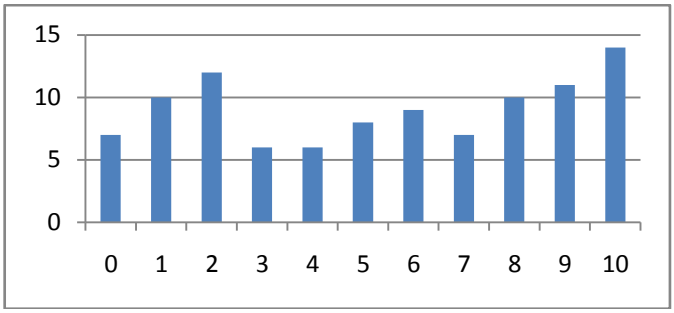
5. ***LOI DE PROBABILITE OBSERVEE ET LOI DE PROBABILITE THEORIQUE***

Pour bien comprendre la différence entre la loi de probabilité observée et la loi de probabilité théorique d'une variable aléatoire, on utilisera l'exemple suivant qu'on pourrait intituler : « simuler le hasard ».

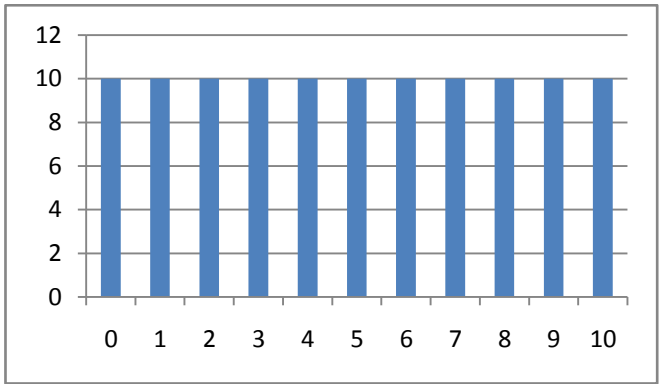
Ecrivez sur un papier un nombre choisi au hasard entre 0 et 10 (inclus).

On demande ensuite à un ordinateur de faire de même, c'est-à-dire de générer de manière aléatoire une suite de nombres compris entre 0 et 10 (on en générera 100).

Voici la loi (distribution) de probabilité observée :



La distribution théorique est évidemment celle-ci (c'est ce qu'on appelle une loi uniforme) :



On tracera à titre d'exercice la distribution de probabilité correspondant aux réponses données par les étudiants.

Que constate-t-on ?

La distribution observée sur ordinateur diffère de la distribution théorique. C'est un phénomène normal, nommé « fluctuations d'échantillonnage ». Si on recommençait un nouveau tirage de 100 nombres sur ordinateur, on obtiendrait un autre graphique.

Quant à la distribution obtenue par les étudiants, nous l'analyserons « en direct ».

Les distributions de probabilité des variables aléatoires ne prennent donc pas n'importe quelle forme et peuvent être proches de distributions théoriques. En particulier, les variables aléatoires utilisées en biologie présentent souvent (mais pas toujours !) une forme de courbe « en cloche », caractéristique de loi normale.

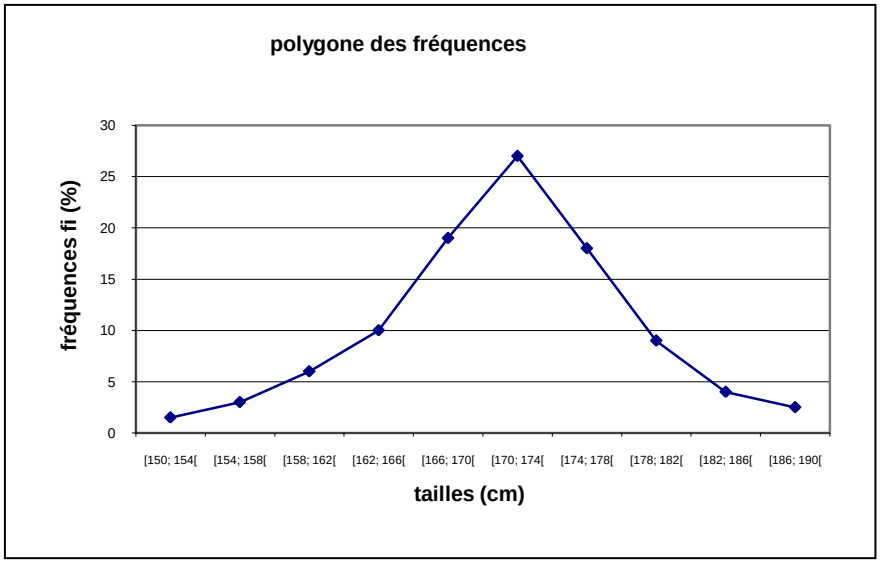
6. *LA LOI NORMALE (LOI DE LAPLACE-GAUSS)*

6.1. *Introduction*

Soit un ensemble de 200 étudiants dont on a noté la taille.

Classes	*effectifs*	*fréquences (%)*
[150; 154[	3	1,5
[154; 158[	6	3
[158; 162[	12	6
[162; 166[	20	10
[166; 170[	38	19
[170; 174[	54	27
[174; 178[	36	18
[178; 182[	18	9
[182; 186[	8	4
[186; 190[	5	2,5

Traçons le polygone des fréquences relatif à cette distribution.



Si, au lieu de se limiter à un échantillon de 200 tailles réparties en 10 classes, on considérait un échantillon de 10000 tailles réparties dans 50 classes s'échelonnant de 150 cm à 190 cm, on conçoit intuitivement que le polygone des fréquences s'apparenterait de plus en plus à ***une courbe en cloche*** symétrique par rapport à la verticale passant par la moyenne $\overline{X}$.

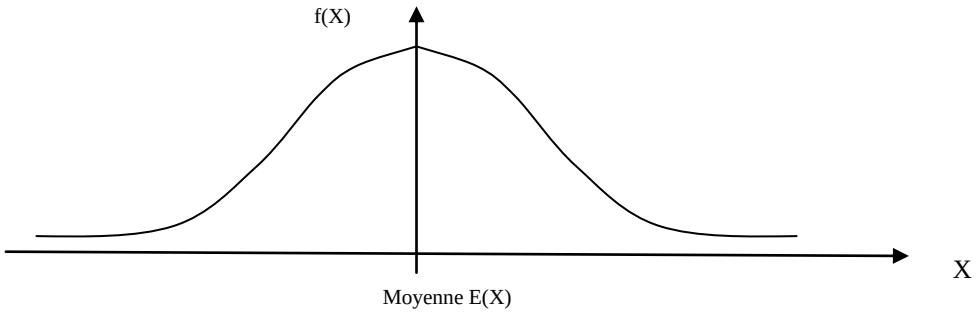
Cette courbe est appelée ***courbe normale*** ou ***courbe de Laplace-Gauss***.

6.2. *Formulation de la loi normale*

La loi normale ou loi de Laplace-Gauss pour les variables aléatoires continues est une des lois les plus importantes en statistiques. Elle constitue un modèle mathématique qui s'accommode de la plupart des distributions statistiques à variable continue.

La fonction qui la représente est la suivante : $$f(X) = \frac{e^{-\frac{1}{2}\cdot\frac{(X-E(X))^2}{\sigma^2}}}{\sigma\sqrt{2\pi}}$$ (fonction de densité de probabilité)

avec $-\infty < X < +\infty$.

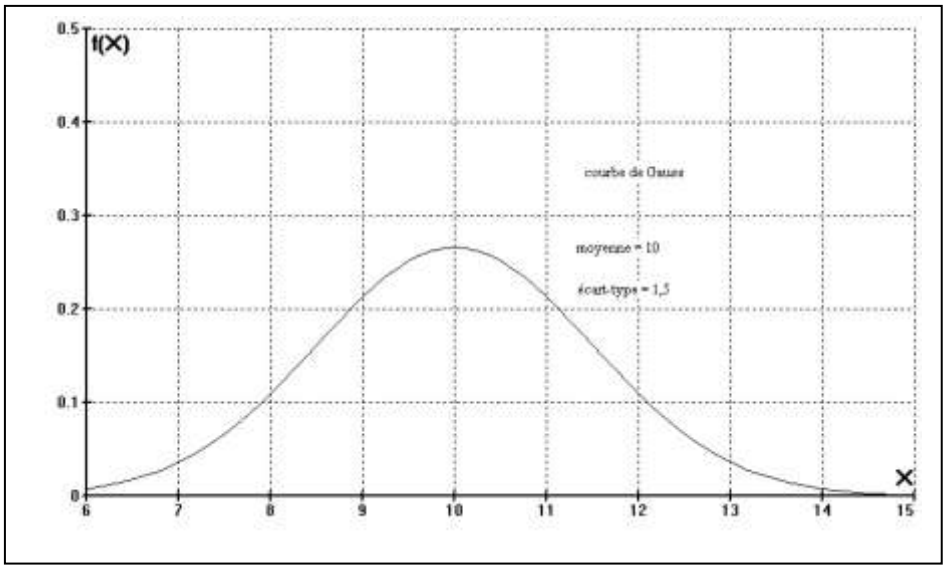


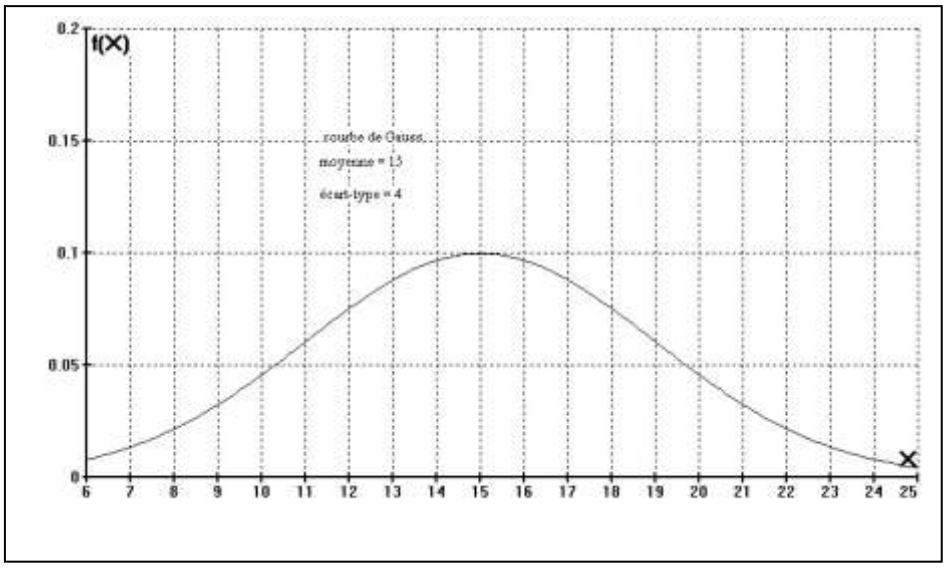
La courbe obtenue est une courbe en cloche symétrique (courbe de Gauss). On constate que les très grandes et les très petites valeurs sont peu probables et que les valeurs sont concentrées autour de la moyenne.

De nombreux caractères peuvent être représentés par une loi normale. C'est le cas notamment de la taille, le poids, les mensurations diverses, le pouls, le quotient intellectuel, mais aussi par exemple la dimension d'objets fabriqués en série.

Cette loi dépend donc des paramètres E(X) et $\sigma(X)$. Pour l'utiliser, il faudrait donc construire des tables pour chaque couple de valeurs (E(X, $\sigma(X)$), ce qui serait impossible car il y aurait une infinité de tables.

Par exemple, on a tracé ci-dessous deux courbes de Gauss ayant comme respectivement comme paramètres 10 et 15 pour l'espérance mathématique et 1,5 et 4 pour l'écart-type.

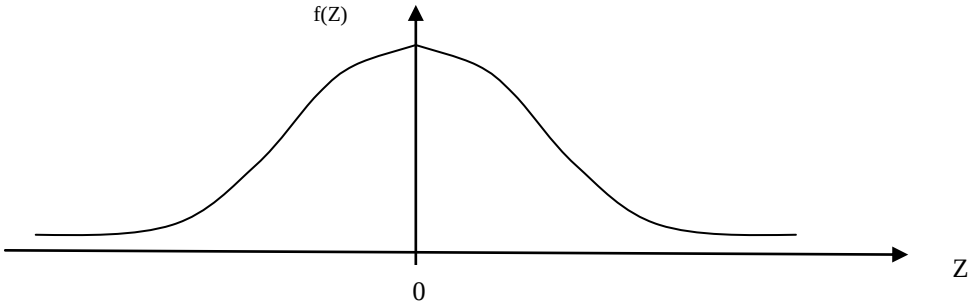




Pour cette raison, on effectue un changement de variables en posant :

$$Z = \frac{\mathbf{X - E(X)}}{\sigma}$$

Et on obtient après manipulation, on obtient : $\boxed{f(Z) = \frac{e^{-\frac{1}{2}.Z^2}}{\sigma\sqrt{2\pi}}}$ (loi normale centrée-réduite)



La courbe de la ***loi normale centrée-réduite*** est aussi une courbe en cloche avec E(Z) = 0 et σ(Z) = 1.

La loi réduite ne dépend donc plus que de la variable Z et il est donc facile de construire des tables. ***Tous les problèmes ayant trait à la loi normale demandent le passage à la loi réduite.***

6.3. Caractéristiques d'une distribution normale

On démontre mathématiquement, que dans une population distribuée normalement:

- *68,27 % des éléments ont une valeur comprise entre E(X) - σ et E(X) + σ,*
- *95,45 % des éléments ont une valeur comprise entre E(X) - 2σ et E(X) + 2σ,*
- *99,73 % des éléments ont une valeur comprise entre E(X) - 3σ et E(X) + 3σ,*
- *le mode est égal à la moyenne et à la médiane,*
- *la courbe est symétrique par rapport à la moyenne,*
- *la courbe a deux points d'inflexion à une distance +σ et -σ de la moyenne E(X).*

6.4. Exercices

1) Dans une population, la moyenne de la taille est de 169 cm avec un écart-type de 10 cm. Déterminer la proportion d'individus dont la taille est supérieure à 180 cm. ***(13,57 %)***

2) On sait que le poids des flacons de pénicilline livrés par une firme pharmaceutique, est distribué normalement avec une moyenne de 126 mg et un écart-type de 4 mg. Si on prélève un lot de 200 flacons, quelle devrait être la proportion d'entre eux dont le poids est:

    a) supérieur à 130 mg (0,1587)
    b) inférieur à 130 mg (0,8413)
    c) compris entre 120 et 130 mg (0,7745)
    d) compris entre 120 et 125 mg (0,3345)
    e) D'après les prescriptions légales, pas plus de 10% des flacons ne peuvent s'écarter de plus de 5% du poids moyen. Peut-on conclure que la firme respecte ces prescriptions ? (non)

3) Une machine fabrique des pièces cylindriques de diamètre nominal 50 mm. La tolérance admise est de – 0,2 à + 0,2 mm.

    La production montre « à la longue » que la dispersion correspond à un écart-type de 0,05 mm. Au cours d'une journée, la production s'est caractérisée par une moyenne des diamètres de 50,1 mm. En supposant que la distribution des diamètres est Gaussienne, quel est le pourcentage de rejet de pièces au cours de cette journée. (0,0228)

4) Une machine à cadence rapide produit des rondelles dont le diamètre suit une loi normale de moyenne 15 mm et d'écart-type 1,3 mm.

    On demande la proportion attendue de rondelles ayant :

    a) moins de 14 mm (0,2206)
    b) plus de 17 mm (0,0618)
    c) entre 14,6 et 15,2 mm (0,1813)
    d) quel est le diamètre de la rondelle la plus étroite dans le groupe des 15 % de rondelles les plus larges ? (16,352 mm)

5) Les pneus "Allseasons" peuvent parcourir en moyenne 56000 km, avec un écart-type de 8000 km et une distribution normale.

    a) Quelle est la probabilité qu'un pneu soit usé avant 50000 km ? *(0,2266)*

    b) Quelle est la probabilité que les quatre pneus Allseasons qui équipent ma voiture soient usés avant 50000 km ? *(0,0026)*

    c) Quelles sont les hypothèses posées en b) ? Comment des hypothèses plus réalistes modifieraient-elles la réponse ?

7. LE TEST DU KHI CARRE (χ2): VERIFICATION DE LA NORMALITE D'UNE DISTRIBUTION

7.1. Introduction

Voici par exemple un relevé du poids de 60 enfants dans une classe de maternelle :

Classes X : poids (kg)	***nombre d'enfants***
[9; 10[	8
[10; 10,5[	12
[10,5; 11[	18
[11; 11,5[	13
[11,5; 12,5[	9
	60

Il est très probable que la variable (le caractère) étudiée suit une loi normale. Si on calcule la moyenne et l'écart-type de manière classique, on obtient respectivement 10,8 et 0,7.

Il pourrait être utile de vérifier si cet échantillon peut être considéré comme extrait d'une population normale. Pour cela, on peut réaliser un test de Khi Carré (χ^2). (il existe d'autres tests comme le test de la droite de Henry ou le test de Kolmogorov-Smirnov).

7.2. Le test du χ2

Le test du χ^2 permet de vérifier s'il y a une différence significative entre une distribution statistique observée et une distribution théorique donnée (la loi normale en ce qui nous concerne)

L'utilisation d'un test statistique (et notamment le test du χ^2) est soumise à un préalable : la formulation d'une hypothèse de départ, appelée hypothèse nulle H_0, qui sera confirmée ou rejetée par le test.

Admettre l'hypothèse nulle H_0 en statistiques, c'est admettre que des différences observées entre la réalité des faits et les hypothèses émises ne sont pas dues à des causes systématiques mais au hasard de l'échantillonnage.

Deux erreurs d'interprétation sont évidemment possibles :

- *Rejeter l'hypothèse nulle alors qu'elle est vraie.*
- *Adopter l'hypothèse nulle alors qu'elle est fausse.*

Dans le cas du test du χ^2, l'hypothèse de départ (H_0 : hypothèse nulle), c'est de dire que ***la répartition expérimentale confirme la répartition théorique et donc que les différences observées sont uniquement dues au hasard***.

Toutefois, l'acceptation ou le rejet de l'hypothèse nulle est liée au risque d'erreur que le chercheur est prêt à prendre pour affirmer que l'hypothèse nulle est vraie ou fausse. Ce risque d'erreur est appelé ***"seuil de signification"*** α.

Quand l'hypothèse nulle est formulée, on calcule ensuite le $\chi^2_{calc.}$, qui est un indicateur d'écart entre la distribution observée et la distribution théorique.

Effectifs observés : $O_1, O_2, \ldots, O_k$

Effectifs théoriques : $A_1, A_2, \ldots, A_k$ (ce qu'on attend d'après la loi de probabilité choisie)

$$\chi^2_{calc} = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$

k : nombre de classes de la distribution

Comme les écarts $O_i - A_i$ peuvent être positifs ou négatifs, et comme on veut faire la sommation, on considère le carré des écarts.

Plus χ^2 calculé est grand, plus il y a discordances entre les effectifs observés et les effectifs théoriques.

Il est évident que si $\chi^2=0$, la théorie est égale à la réalité (« cas idéal »). Si ce n'est pas le cas, il faut rechercher dans la table le $\chi^2_{théorique}$.

Recherche du $\chi^2_{théorique}$.

On fixe la probabilité de voir $\chi^2_{calc.}$ dépasser une certaine limite ($\chi^2_{théorique}$). Cette probabilité est aussi appelée seuil de signification α.

On prend généralement un seuil de signification de 0,05, et donc la probabilité d'avoir $\chi^2_{calc} > \chi^2_{théorique}$ est supposée égale à 5 %. *(Dans 5 cas sur 100, le χ^2_{calc} dépassera le $\chi^2_{théorique}$ par le fait du hasard)*

On détermine le nombre de degrés de libertés (d.d.l.)

d.d.l. = k – 1 – nombre de paramètres à calculer pour trouver les effectifs théoriques

Quand on connaît le seuil de signification et le nombre de d.d.l., on va rechercher la valeur de $\chi^2_{théorique}$ dans la table ci-dessous. Deux cas peuvent alors se présenter :

1) Si $\chi^2_{calc} > \chi^2_{théorique}$, alors la différence est trop forte pour être due aux seuls faits du hasard et on est conduit à rejeter l'hypothèse de départ. (les discordances sont significatives, elles sont trop grandes)
2) Si $\chi^2_{calc} < \chi^2_{théorique}$, on accepte l'hypothèse de départ. (les discordances observées ne sont pas significatives, elles ne sont pas trop fortes, elles sont dues au hasard)

Remarques :

- le « moins 1 » dans le nombre de degrés de libertés vient du fait qu'une fois les effectifs répartis dans k – 1 classes, les observations restantes doivent normalement revenir à la dernière classe. Cette dernière classe n'a par conséquent pas la « liberté » (l'indépendance) de prendre la valeur qu'elle veut.

- plus le nombre de degrés de liberté est élevé, plus le $\chi^2_{théorique}$ est élevé. On acceptera donc plus de discordances entre les effectifs observés et théoriques.

- plus on a de paramètres à calculer, plus le nombre de degrés de liberté sera faible, et plus le $\chi^2_{théorique}$ sera faible. On acceptera donc moins de discordances entre les effectifs observés et théoriques.

- plus le seuil de signification est faible, plus le $\chi^2_{théorique}$ est élevé et on est donc moins restrictif (on acceptera plus de discordances).

  Par conséquent, un seuil de signification de 0,05 est moins « sévère » qu'un seuil de signification de 0,25.

  Avec 0,05, on accorde moins de signification aux discordances qu'avec 0,25.

Conditions de validité du test

- la taille de l'échantillon doit être suffisamment grande : $n \geq 50$.
- Les effectifs observés de chaque classe ne doivent pas être trop faibles. En pratique, un effectif observé de 5 est considéré comme un minimum.

  Si certaines classes ne satisfont pas à cette condition, on doit les regrouper avec les classes voisines jusqu'à ce que la règle soit respectée.
  Dans ce cas, c'est ce nouveau nombre de classes qui sera pris en compte pour le calcul du nombre de degrés de liberté.

Prenons l'exemple suivant :

Supposons qu'une pièce de monnaie ait été lancée en l'air 50 fois et soit retombée en ne montrant que le côté face que 15 fois. Peut-on affirmer que la pièce était faussée ou que celui qui la lançait était moins honnête qu'il le paraissait ?

Hypothèse nulle *: la pièce n'est pas faussée et les divergences observées sont dues au hasard.*

Calcul du χ^2_{calc} : $\chi^2_{calc} = \frac{(35-25)^2}{25} + \frac{(15-25)^2}{25} = 8$

Nombres de ddl : 2 – 1 = 1

- Avec $\alpha = 0{,}05$, (c'est-à-dire que si on réalise 100 fois l'expérience correctement, le χ^2_{calc} ne dépassera $\chi^2_{théorique}$ par le fait du hasard que 5 fois), on obtient $\chi^2_{théorique} = 3{,}84$.

  Puisque $\chi^2_{calc} > \chi^2_{théorique}$ (les différences sont trop importantes), on rejette donc l'hypothèse nulle, le risque de se tromper étant de 5 %.

- Avec $\alpha = 0{,}01$, (c'est-à-dire que si on réalise 100 fois l'expérience correctement, le χ^2_{calc} ne dépassera $\chi^2_{théorique}$ par le fait du hasard que 1 fois), on obtient $\chi^2_{théorique} = 6{,}64$.

  Puisque $\chi^2_{calc} > \chi^2_{théorique}$ (les différences sont trop importantes), on rejette donc l'hypothèse nulle, le risque de se tromper étant de 1 %.

- Si on veut un degré de certitude très élevé, on peut prendre $\alpha = 0{,}001$, (c'est-à-dire que si on réalise 1000 fois l'expérience correctement, le χ^2_{calc} ne dépassera $\chi^2_{théorique}$ par le fait du hasard que 1 fois), on obtient $\chi^2_{théorique} = 10{,}83$.

  Dans ce cas, $\chi^2_{calc} < \chi^2_{théorique}$, on ne peut pas se prononcer sur le fait que la pièce est faussée.

7.3. Applications

1) Reprendre l'exemple du relevé du poids de 60 enfants dans une classe de maternelle et vérifier l'hypothèse de normalité de la distribution.

Classes X : poids (kg)	*nombre d'enfants*
[9; 10[	8
[10; 10,5[	12
[10,5; 11[	18
[11; 11,5[	13
[11,5; 12,5[	9
	60

2) Le tableau suivant donne le résultat d'une enquête sur la taille des étudiants en informatique.

X_i : tailles en cm	*effectifs observés*
[150,160[	28
[160,170[	110
[170,180[	69
[180,190[	61
[190,200[	32
	300

a) Vérifier que les tailles de ces étudiants ne suivent pas du tout une loi normale.

b) Critiquez le résultat obtenu. Cela vous semble-t-il possible ? Observez attentivement le tableau de données.

3) Vous réalisez une enquête donnant sur le nombre de pulsations par minute (pouls) d'un échantillon de 250 adultes de 20 à 35 ans.

a) Compléter le tableau ci-dessous en indiquant des effectifs observés tels que la distribution se rapproche d'une loi normale et le vérifier par le test du khi carré.

b) Prendre des seuils α de 0,05 et de 0,01 et expliquer leur signification.

c) Tracer un graphique permettant de confirmer "grossièrement" que la variable suit une loi normale".

Nombre de pulsations par minute	*Effectifs observés O_i*
[40; 50[	
[50; 60[	25
[60; 70[	
[70; 80[	
[80; 90[	49
[90; 100[	
[100; 110[	12

4) Un cadre d'une entreprise métallurgique effectue une étude statistique sur l'épaisseur de pièces en acier.

L'enquête dure un mois et les résultats sont repris dans le tableau suivant :

X_i : épaisseur (en 1/10 mm)	n_i
[100; 105[	13
[105; 110[	101
[110; 115[	282
[115; 120[	430
[120; 125[	253
[125; 130[	92
[130; 135[	8

a) Vérifier par le test du Khi carré que ces données suivent une loi normale.

b) En particulier, montrer ***sur papier*** la méthode utilisée pour calculer les effectifs théoriques de la classe [100; 105[.

c) En supposant que le test est concluant, calculer ***sur papier*** avec la loi normale le pourcentage de rebuts si l'on considère que l'on rejette les pièces d'épaisseur inférieure à 108 et supérieure à 128 (1/10 de mm).

d) Expliquer la signification de l'écart-type.

CHAPITRE 3 : INFÉRENCE STATISTIQUE

1. PRINCIPES DE L'INFERENCE STATISTIQUE

En statistique descriptive, nous nous sommes limités à décrire un ensemble de données, à en extraire l'information au moyen de graphiques, tableaux de fréquences et calculs de paramètres, et ensuite à décrire quelques distributions théoriques de variables, à savoir binomiale, de Poisson et normale.

L'inférence statistique a pour but d'émettre un jugement sur une ou plusieurs caractéristiques de la population à partir d'une ou plusieurs caractéristiques observées sur un échantillon extrait de cette population.

Par population, on entend la totalité des observations individuelles existant dans une aire spécifiée, limitée dans l'espace et dans le temps, et au sujet desquelles on désire tirer des conclusions. Par exemple, la population des tailles de tous les Belges, la population des concentrations hépatiques d'une enzyme chez tous les rats mâles d'une même espèce, la population des mesures d'erreurs si on pouvait prélever un millilitre de façon infinie, ...

L'échantillon constitue un sous-ensemble limité de la population.

Exemple : les tailles de quarante Belges, la concentration hépatique d'une enzyme mesurée chez dix rats, ...

Pour que l'inférence statistique ait un sens, ***l'échantillon doit être représentatif de la population***. On admettra en général son caractère représentatif si l'échantillon est prélevé de façon aléatoire, c'est-à-dire lorsque chaque individu a une même chance d'être prélevé et que les individus sont prélevés indépendamment les uns des autres. (voir précédemment)

Toute caractéristique numérique calculée à partir des observations de l'échantillon est appelée une ***statistique*** et toute caractéristique numérique associée à la population sera dorénavant appelée un ***paramètre***.

Les principes de l'inférence statistique concourent à deux objectifs :

- *Les problèmes d'estimation : à partir d'une statistique d'échantillon, obtenir une estimation précise du paramètre de la population.*
- *Les tests d'hypothèse : vérifier la vraisemblance d'une hypothèse concernant le ou les paramètres d'une ou de plusieurs populations à partir des statistiques calculées dans le ou les échantillons.*

2. L'ESTIMATION

2.1 Introduction

L'estimation, c'est l'ensemble des méthodes utilisées pour évaluer *un paramètre* θ d'une population à l'aide *d'un estimateurs* $\hat{\theta}$ pris dans un échantillon extrait de cette population.

Par exemple, dans le cas de l'étude de la taille d'une population, les paramètres que sont la moyenne μ et la variance σ^2 sont des constantes (qui sont généralement inconnues).

Si on choisit plusieurs échantillons différents à partir de cette population, les moyennes $\overline{X}$ et les variances s^2 de ces échantillons seront différentes pour chaque échantillon choisi. Ce sont par conséquent des variables aléatoires.

Un grand nombre de problèmes statistiques consistent en la détermination de la moyenne "vraie" μ sur base d'observations réalisées sur un échantillon, mais on peut aussi chercher à connaître d'autres caractéristiques, comme la variance par exemple.

Exemples :

- *Quelle est la fréquence d'apparition de tel type de cancer chez les souris ?*
- *Quelle est la "vraie" valeur de la glycémie chez un patient ?*
- *Quelle est la variance de la glycémie mesurée chez ce même patient ?*

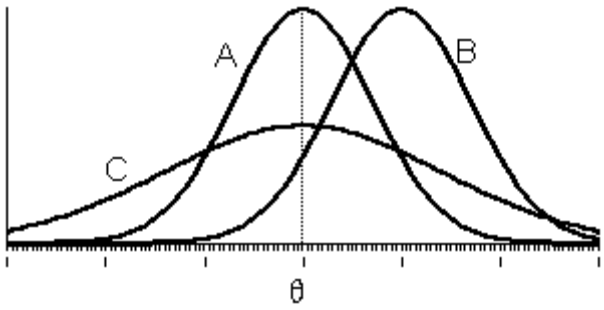
2.2 Estimateurs biaisés et non biaisés

Si la moyenne (l'espérance mathématique) d'un estimateur d'échantillonnage est égale au paramètre correspondant de la population, on dit que l'estimateur est un estimateur ***non biaisé*** de ce paramètre. Dans le cas contraire, on dit que l'on a un estimateur ***biaisé***.

- *Estimateur non biaisé (ou sans biais)* : $E(\hat{\theta}) = \theta$
- *Sinon, le biais est défini par* : $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Les valeurs correspondantes de ces estimateurs sont respectivement des estimations non biaisées ou biaisées.

Dans la figure ci-dessous, les estimateurs dont les distributions de probabilité sont A et C sont non biaisés (sa moyenne est égale à θ) alors que celui dont la distribution est B est biaisé.



Si $\overline{X}$ et s² sont respectivement la moyenne et la variance d'un échantillon, on peut démontrer que $\overline{X}$ et s² sont des estimateurs sans biais de la moyenne μ et de la variance σ^2 de la population correspondante.

$$E(\overline{X}) = \mu \text{ et } E(s^2) = \sigma^2.$$

(s² est la variance de l'échantillon définie par : $s^2 = \frac{\sum_{i=1}^{k} n_i . \left| X_i - \overline{X} \right|^2}{n-1}$)

Remarque : on peut montrer que si on avait pris comme estimateur de la variance σ^2 de la population $s_n^2 = \frac{\sum_{i=1}^{k} n_i . \left| X_i - \overline{X} \right|^2}{n}$, cet estimateur serait biaisé. On préférera donc utiliser s².

2.3 Estimateurs efficaces

Si on compare les distributions de probabilité A et C des deux estimateurs du graphique ci-dessus, celui qui a la variance la plus faible est considéré comme le plus efficace. Cela correspond évidemment à une distribution la plus concentrée possible.

3. ESTIMATION PONCTUELLE

Cette méthode consiste à attribuer au paramètre inconnu θ une valeur approchée de l'estimateur mesurée ***dans un échantillon*** pris au hasard.

Paramètres de la population	***Estimateurs***
Moyenne de μ	$\hat{\mu} = \overline{X}$
Proportion d'individus π possédant la caractéristique A	$\hat{\pi} = \frac{n_A}{n}$
Ecart-type σ	$\hat{\sigma} = s$

Cette méthode a comme inconvénient de ne pas donner d'indication sur l'erreur possible entre l'estimé et le paramètre.

Exemples

1) Si 150 étudiants de 6ème ont un QI de 135 avec un écart type de 15 et si 38 de ces étudiants fréquentent l'enseignement technique, estimer le QI moyen, l'écart type et la proportion d'étudiants fréquentant le technique de tous les étudiants de 6ème de la ville.

2) Pour étudier la consommation d'essence des voitures d'une certaine marque, on prélève un échantillon de 12 voitures pour lesquelles on note les consommations en litres par 100 km.

9,7	10,3	9,9	10,4	10,5	10,8	11,2	11	8,9	10	10,7	10,8

Estimer de façon ponctuelle la moyenne et l'écart type de la consommation d'essence des voitures de cette marque.

4. LE THEOREME CENTRAL LIMITE

Les informations relatives à la distribution d'échantillonnage de la moyenne se résument en un théorème : ***le théorème central limite***.

Etant donné une population ayant une moyenne μ et d'écart-type σ :

Considérons l'ensemble des échantillons aléatoires possibles de taille n. On a vu précédemment que : E ($\overline{X}$) = μ.

Il est évident que les moyennes d'échantillon $\overline{X_1}$, $\overline{X_2}$, … varient entre elles, puisque les échantillons varient de l'un à l'autre. On peut mesurer cette variabilité des moyennes d'échantillon en calculant l'écart-type $\sigma(\overline{X})$ de celles-ci.

Le théorème central limite nous dit que : $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

Et donc, la moyenne de l'échantillon $\overline{X}$ varie autour de la moyenne de la population μ avec un écart-type égal à $\frac{\sigma}{\sqrt{n}}$.

Dans la réalité, $\sigma(\overline{X})$ n'est pas connu puisque σ est inconnu. En pratique, on a vu que l'écart-type de l'échantillon s est une estimation de σ. On peut dès lors remplacer σ par s et obtenir une estimation de $\sigma(\overline{X})$, notée $s(\overline{X})$:

$$s(\overline{X}) = \frac{s}{\sqrt{n}}$$

Remarquons que cette formule est très intéressante car elle permet d'estimer la variabilité des moyennes d'échantillon $\overline{X}$ ***alors qu'on ne dispose en pratique que d'un seul échantillon*** !

Autre résultat important

Lorsque la population mère est normale, ou lorsque la taille de l'échantillon est suffisamment grande (n > 30), la distribution d'échantillonnage de $\overline{X}$ a une forme approximativement normale.

5. ESTIMATION PAR INTERVALLE DE CONFIANCE

Estimer un paramètre θ à partir d'un échantillon aléatoire simple fournit, comme on vient de le voir, une estimation ponctuelle.

On peut par exemple estimer la taille moyenne μ des élèves de toutes les élèves de toutes les classes de $6^{ème}$ secondaire de Belgique à l'aide de la moyenne $\overline{X}$ d'un échantillon restreint d'élèves.

Une telle opération ne nous permet pas de savoir si la valeur obtenue est proche ou non de la valeur inconnue. Pour remédier à ce problème, on utilise la notion ***d'intervalle de confiance***.

Définir un intervalle de confiance, c'est rechercher un encadrement d'une valeur inconnue (moyenne, pourcentage,...) qui soit à la fois très probable et le plus serré possible. La probabilité que l'intervalle contienne la valeur inconnue est égale à 1 - α quelle que soit cette valeur.

Cette probabilité 1 - α est appelée niveau de confiance, α étant un seuil ou niveau de probabilité habituellement choisi en-dessous de 0,1.

6. INTERVALLE DE CONFIANCE DE LA MOYENNE μ

6.1 Introduction

Si $\overline{X}$ (moyenne d'un échantillon) est un bon estimateur de μ (moyenne de la population), la valeur observée de $\overline{X}$ sur un échantillon est toujours un peu plus grande ou plus petite que μ. On ne peut plus affirmer que $\mu = \overline{X}$, on construit alors un intervalle de confiance de la forme :

$\mu = \overline{X} \pm$ marge (ou erreur) d'échantillonnage

Quelle sera l'importance de cette marge d'erreur ?

On doit décider du degré de confiance souhaité $1 - \alpha$. Si on choisit un niveau de confiance de 0,95 (et donc α = 0,05), cela signifie que sur 100 échantillons prélevés, 95 donneront un intervalle qui contient la vraie valeur à estimer.

Prenons l'exemple suivant. Supposons que l'on veuille construire un intervalle de confiance pour la moyenne μ des tailles de la population des étudiants de l'Université de Liège en se basant sur un échantillons de 10 étudiants. On suppose également qu'on connaît (ce qui est étonnant !) les valeurs de μ et σ de la population, soit respectivement 169 cm et 3,22 cm.

Si on prend 50 échantillons, on va obtenir des moyennes $\overline{X_1}$, $\overline{X_2}$, $\overline{X_3}$, …, $\overline{X_{50}}$ différentes (dont la moyenne est égale à μ et l'écart-type égal à $\frac{\sigma}{\sqrt{n}}$, voir le théorème central limite).

Pour chacun des 50 échantillons, on peut calculer un intervalle de confiance (de niveau de confiance 95 %). On a donc 50 intervalles de confiance différents et on regardera s'ils contiennent ou non la moyenne "vraie" μ (ici 169 cm).

Dans le cas d'un degré de confiance de 95 %, on devrait constater qu'environ 47 ou 48 intervalles de confiance sur 50 (soit environ 95 %) contiennent la moyenne "vraie" μ.

Remarquons que si on avait choisit un degré de confiance de 99 %, on devrait constater qu'environ 49 ou 50 intervalles de confiance sur 50 contiennent la moyenne "vraie" μ. Il est évident par conséquent que les intervalles de confiance "à 99 %" seront plus étendus que ceux "à 95 %".

En pratique, on ne prélève qu'un seul échantillon et on calcule un seul intervalle de confiance. La probabilité que cet intervalle contienne la vraie valeur sera égale au niveau de confiance choisi (90 %, 95 %, 99 %, …)

6.2 Echantillonnage avec ou sans remise

Lors du prélèvement d'un échantillon, deux méthodes sont possibles. Prenons le cas d'une population de 1000 boules numérotées dont on veut extraire un échantillon de 10 boules. On peut choisir une boule puis décider de la mettre sur le côté ou de la remettre dans l'urne (tirages sans ou avec remise). Dans le premier cas, cette boule ne pourra pas être tirée une deuxième fois, il y aura une boule de moins dans l'urne et donc les tirages successifs seront dépendants. Dans le deuxième cas, l'urne est à chaque fois reconstituée et les tirages sont indépendants.

Si la population est très grande, la différence entre ces deux cas sera faible et on la négligera. Par contre, dans le cas de petite population, il faut utiliser des facteurs de correction (facteur d'exhaustivité). Pour simplifier, nous considérerons dans la suite que les différentes observations (ou tirages) sont indépendantes l'une de l'autre et il n'y aura donc pas de facteur de correction à appliquer.

6.3 Cas de grands échantillons ($n \geq 30$)

On pourrait démontrer que la marge d'erreur est de la forme :

$$Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ (si } \sigma \text{ est connu)} \qquad \text{ou} \qquad Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \text{ (si } \sigma \text{ est inconnu)}$$

où $Z_{\alpha/2}$ dépend du seuil de confiance considéré et tel que $P(Z < Z_{\alpha/2}) = 1 - \alpha/2$. On recherche la valeur de $Z_{\alpha/2}$ dans les tables de la loi normale.

Par exemple :

- *pour α = 10 %, on a : $Z_{\alpha/2}$ = 1,645*
- *pour α = 5 %, on a : $Z_{\alpha/2}$ = 1,96*
- *pour α = 1 %, on a : $Z_{\alpha/2}$ = 2,575*

Donc en risquant de se tromper avec une probabilité égale à α, on peut affirmer que :

$$\overline{X} - Z_{\alpha/2}.\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}.\frac{s}{\sqrt{n}}$$

L'intervalle $[\overline{X} - Z_{\alpha/2}.\frac{s}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}.\frac{s}{\sqrt{n}}]$ est appelé intervalle de confiance de μ au niveau de confiance $1 - \alpha$.

On peut aussi écrire que : $P(\overline{X} - Z_{\alpha/2}.\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}.\frac{s}{\sqrt{n}}) = 1 - \alpha$

Exemple

Dans un ensemble de 150 étudiants, on a constaté dans un exercice d'expression orale que le temps de parole était distribué normalement avec un écart type de 2 min. On prélève sans remise un échantillon de 30 étudiants pour lesquels le temps moyen de parole utilisé pour cet exercice est de 14 min.

Construire un intervalle de confiance au niveau de confiance de 99% pour estimer le temps moyen de parole utilisé pour la population des 150 étudiants.

Déterminer la taille d'échantillon qu'il faudrait considérer pour réduire la marge d'erreur de moitié.

6.4 Cas des petits échantillons extraits d'une population normale (n < 30)

Au paragraphe 4, on a vu que si la variable aléatoire X suit une loi normale, la distribution des moyennes $\overline{X}$ suit une loi normale $N(\mu, \frac{\sigma}{n})$.
Si σ est connu, l'intervalle de confiance de la moyenne se construit comme dans le cas des grands échantillons.

Mais en général, σ est inconnu et estimé par s. Dans le cas des petits échantillons, en remplaçant σ par s et on modifie la nature de la loi suivie par m.

La variable aléatoire $t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ suit la loi de Student à n-1 degrés de liberté.
Le coefficient de risque α étant choisi et le nombre de degré de liberté étant connu, les tables permettent de lire t_α tel que $P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$.

On en déduit la marge d'erreur sur la moyenne de la population :

$$t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

L'intervalle [$\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$, $\overline{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$] est appelé intervalle de confiance de μ au niveau de confiance $1 - \alpha$.

Lorsque le nombre de degrés de liberté tend vers l'infini, la fonction de répartition de la loi de Student tend vers celle de la loi normale centrée réduite.

7. INTERVALLE DE CONFIANCE D'UNE FREQUENCE (N ≥30)

Dans un échantillon de taille n tiré d'une population binomiale pour laquelle p est la fréquence de succès, les limites de confiance de p sont données par $p \pm Z_{\alpha/2}\,\sigma_p$ où p est la proportion de succès dans l'échantillon de taille n.

Les conditions suivantes étant réalisées : $np \geq 5$ et $n.(1-p) \geq 5$, les limites de confiance sont données par :

$$p \pm Z_{\alpha/2}\sqrt{\frac{p.(1-p)}{n}}$$

8. EXERCICES

1) La durée de fonctionnement d'une pile suit une loi normale d'écart type 0,73 ans. On prélève un échantillon de 36 piles et on a observé une moyenne de 4,4 ans. Construire un intervalle de confiance au niveau de confiance 95% pour estimer la durée de fonctionnement moyenne des piles de ce type. On suppose que l'échantillonnage a été fait avec remise.

   Quelle devrait être la taille d'échantillon de façon à n'avoir une marge d'erreur que de 0,1 au même niveau de confiance ?
   (4,16 - 4,64 ; 205)

2) Le personnel d'un bureau d'experts fiscaux sait que le temps nécessaire pour remplir une déclaration fiscale est une variable aléatoire d'écart type 15,3 min. Pour estimer le temps moyen, on prélève un échantillon de 60 déclarations et on note le temps moyen 77,8 min. On demande de construire un intervalle de confiance au niveau de confiance de 90% pour estimer le temps moyen nécessaire pour remplir une déclaration fiscale. (échantillon avec remise)
   (74,55 - 81,05)

3) Sur un échantillon de 83 coquillages, on a relevé une cote déterminée x. On a trouvé une moyenne de 52,17 mm et un écart type de 1,8 mm.

   a) estimer la moyenne de la population par un intervalle de confiance à 95%

   b) quel devrait être l'effectif de l'échantillon pour situer la moyenne dans un intervalle de 0,4 mm avec une sécurité de 95 % ?
      (51,78 - 52,56 ; 311)

4) Un fabricant de stylos met un nouveau stylo sur le marché. Pour estimer la durée moyenne d'écriture de ces stylos, il prélève un échantillon de 16 stylos et il note leur durée d'écriture (en heures).

   87 92 97 93 91 96 95 92 68 85 90 86

   102 101 91 100

   En supposant que la durée de vie suit une loi normale, on demande de construire un intervalle de confiance au niveau de confiance de 95% pour estimer la durée moyenne d'écriture.
   (87,3 - 95,9)

5) Pour connaître les intentions de vote des belges, on prélève un échantillon de 900 personnes et on trouve que 426 d'entre elles ont l'intention de voter pour le parti "XYZ". On demande de construire un intervalle de confiance au niveau de confiance de 95 % pour estimer la proportion des électeurs qui ont l'intention de voter pour le parti "XYZ" aux prochaines élections. Déterminer le nombre de personnes qu'il faudrait consulter de façon à ce que la marge d'erreur ne dépasse pas 0,02.
   (0,44 - 0,51 ; 2395)

6) Marc Dupont, un étudiant, veut savoir s'il vaut la peine de se porter candidat à la présidence de l'Association des Etudiants. Un échantillon de 50 étudiants a montré que 22% des étudiants voteraient pour lui. Estimer le véritable pourcentage à un niveau de confiance de 99%.
   (0,07 - 0,37)

7) Dans un échantillon aléatoire de 400 personnes, 320 approuvent la nouvelle politique gouvernementale en matière d'immigration. Au niveau de confiance de 95%, estimer le pourcentage de la population en accord avec la nouvelle politique. (0,76 - 0,84)

8) Madame X, député, est inquiète. La victoire aux prochaines élections est loin d'être assurée. Prise de panique, elle s'empresse de commander un sondage pour savoir comment elle est perçue par l'électorat. Des 1200 votants interrogés, 532 ont affirmé qu'ils voteraient pour madame X, tous les autres semblent préférer son adversaire ou sont indécis. Au niveau de confiance de 95%, estimer le pourcentage de l'électorat qui appuie la députée.
   (41,53 - 47,13)

9) Un échantillon de 100 votants choisis au hasard parmi tous les votants d'une circonscription donnée a montré que 55 % d'entre eux étaient favorables à un certain candidat.

Déterminer les limites de confiance (a) à 95% (b) à 99% (c) à 99,73 % de la proportion de tous les votants favorables à ce candidat. (0,55 ± 0,10 ; 0,55 ± 0,13, 0,55 ± 0,15)

10) Une entreprise de production de graines veut vérifier la faculté germinative d'une espèce, c'est-à-dire la probabilité p pour qu'une graine, prise au hasard dans la production, germe.
Sur un échantillon de 400 graines, on observe que 330 graines germent. Quel est l'intervalle de confiance de p au risque 5 % ? au risque 1 % ?
([0,788 - 0,862] ; [0,776 - 0,874])

11) On a mesuré la densité minérale osseuse (DMO) chez 12 sujets hémiplégiques.

Côté paralysé	0,772	0,630	0,757	0,541	0,607	0,474	0,489	0,570	0,721	0,824	0,525	0,824
Côté sain	0,750	0,628	0,824	0,599	0,667	0,649	0,499	0,570	0,691	0,883	0,599	0,827

Déterminer l'intervalle de confiance sur la DMO au niveau de confiance de 95 %.

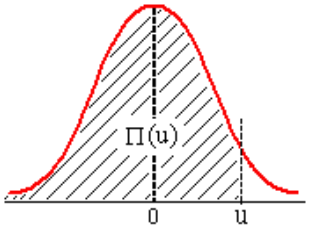
([0,608 – 0,7564] ; [0,5622 – 0,7268])

ANNEXE 1 : TABLES

Table de Loi Normale

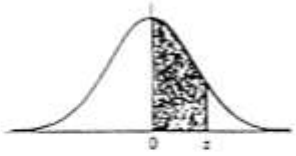
Fonction de répartition de la loi normale centrée réduite.

Probabilité de trouver une valeur inférieure à u.



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

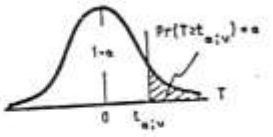
AREAS under the STANDARD NORMAL CURVE from 0 to z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Distribution de Student

Distribution de Student



ν α	0,25	0,1	0,05	0,025	0,01	0,005
1	1,0000	3,0777	6,3137	12,7062	31,8210	63,6559
2	0,8165	1,8856	2,9200	4,3027	6,9645	9,9250
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8408
4	0,7407	1,5332	2,1318	2,7765	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9979	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467
16	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,6864	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,6848	1,3178	1,7109	2,0639	2,4922	2,7970
25	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,6840	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500
31	0,6825	1,3095	1,6955	2,0395	2,4528	2,7440
32	0,6822	1,3086	1,6939	2,0369	2,4487	2,7385
33	0,6820	1,3077	1,6924	2,0345	2,4448	2,7333
34	0,6818	1,3070	1,6909	2,0322	2,4411	2,7284
35	0,6816	1,3062	1,6896	2,0301	2,4377	2,7238
36	0,6814	1,3055	1,6883	2,0281	2,4345	2,7195
37	0,6812	1,3049	1,6871	2,0262	2,4314	2,7154
38	0,6810	1,3042	1,6860	2,0244	2,4286	2,7116
39	0,6808	1,3036	1,6849	2,0227	2,4258	2,7079
40	0,6807	1,3031	1,6839	2,0211	2,4233	2,7045
41	0,6805	1,3025	1,6829	2,0195	2,4208	2,7012
42	0,6804	1,3020	1,6820	2,0181	2,4185	2,6981
43	0,6802	1,3016	1,6811	2,0167	2,4163	2,6951
44	0,6801	1,3011	1,6802	2,0154	2,4141	2,6923
45	0,6800	1,3007	1,6794	2,0141	2,4121	2,6896
46	0,6799	1,3002	1,6787	2,0129	2,4102	2,6870
47	0,6797	1,2998	1,6779	2,0117	2,4083	2,6846
48	0,6796	1,2994	1,6772	2,0106	2,4066	2,6822
49	0,6795	1,2991	1,6766	2,0096	2,4049	2,6800
50	0,6794	1,2987	1,6759	2,0086	2,4033	2,6778
51	0,6793	1,2984	1,6753	2,0076	2,4017	2,6757
52	0,6792	1,2980	1,6747	2,0066	2,4002	2,6737
53	0,6791	1,2977	1,6741	2,0057	2,3988	2,6718
54	0,6791	1,2974	1,6736	2,0049	2,3974	2,6700
55	0,6790	1,2971	1,6730	2,0040	2,3961	2,6682

Table du χ^2

		$\alpha \leftrightarrow$			
ddl ↓	0,99	0,95	0,05	0,025	0,01
1			3,84	5,02	6,63
2	0,02	0,10	5,99	7,38	9,21
3	0,11	0,35	7,81	9,35	11,34
4	0,30	0,71	9,49	11,1	13,28
5	0,55	1,15	11,07	12,8	15,09
6	0,87	1,64	12,59	14,4	16,81
7	1,24	2,17	14,07	16,0	18,48
8	1,65	2,73	15,51	17,5	20,09
9	2,09	3,33	16,92	19,0	21,67
10	2,56	3,94	18,31	20,5	23,21
11	3,05	4,57	19,68	21,9	24,72
12	3,57	5,23	21,03	23,3	26,22
13	4,11	5,89	22,36	24,7	27,69
14	4,66	6,57	23,68	26,1	29,14
15	5,23	7,26	25,00	27,5	30,58
16	5,81	7,96	26,30	28,8	32,00
17	6,41	8,67	27,59	30,2	33,41
18	7,01	9,39	28,87	31,5	34,81
19	7,63	10,12	30,14	32,9	36,19
20	8,26	10,85	31,41	34,2	37,57
21	8,90	11,59	32,67	35,5	38,93
22	9,54	12,34	33,92	36,8	40,29
23	10,20	13,09	35,17	38,1	41,64
24	10,86	13,85	36,42	39,4	42,98
25	11,52	14,61	37,65	40,6	44,31
26	12,20	15,38	38,89	41,9	45,64
27	12,88	16,15	40,11	43,2	46,96
28	13,56	16,93	41,34	44,5	48,28
29	14,26	17,71	42,56	45,7	49,59
30	14,95	18,49	43,77	47,0	50,89
40			55,8	59,3	63,7

ANNEXE 2 : EXEMPLE INTRODUCTIF

EXEMPLE INTRODUCTIF : VARIABLE QUANTITATIVE CONTINUE

On a mesuré en millisecondes, à quelle vitesse 50 enfants de quatre ans identifiaient des images simples (ours, lapin, chat, ...).

Les résultats sont présentés dans le tableau suivant :

Enfants	1	2	3	4	5	6	7	8	9	10
Temps (en ms)	24	27	38	21	27	19	23	23	24	19

Enfants	11	12	13	14	15	16	17	18	19	20
Temps (en ms)	27	30	15	27	24	35	18	20	21	15

Enfants	21	22	23	24	25	26	27	28	29	30
Temps (en ms)	33	27	20	32	28	27	22	17	30	18

Enfants	31	32	33	34	35	36	37	38	39	40
Temps (en ms)	21	25	25	29	25	24	32	31	28	20

Enfants	41	42	43	44	45	46	47	48	49	50
Temps (en ms)	29	24	23	34	17	15	21	28	24	23

Vu le nombre de valeurs différentes relevées, il y a peu d'intérêt à les représenter sur un graphique. Il est préférable de constituer des classes.

Par exemple, on peut obtenir le tableau suivant :

i : numéro de la classe	1	2	3	4	5	6
Intervalles de temps (en ms) X_i	[15, 18]	[19, 22]	[23, 26]	[27, 30]	[31, 34]	[35, 38]
Effectifs n_i	5	11	14	11	6	3

La variable, notée X_i, est le temps en millisecondes.

ANNEXE 3 : TOUTES SORTES DE "DISTRACTIONS STATISTIQUES"

Lisez ce texte et découvrez la tromperie

Bigot quitte la RTBF pour Endemol France (Journal Le Soir)

JEAN-FRANCOIS LAUWENS

jeudi 10 avril 2008, 21:01

Le Français Yves Bigot, directeur des antennes de la RTBF depuis avril 2006, quittera la télévision publique le 1er septembre prochain. Il a en effet trouvé un accord avec Endemol France dont il prendra la direction des programmes.

Deux ans après avoir pris ses fonctions de directeur des antennes de la RTBF (le 1er avril 2006), Yves Bigot a décidé de faire le chemin inverse et de reprendre la direction de Paris. Bigot va en effet rallier Endemol France, société de production qui a la mainmise sur la télé française puisqu'elle arrose ses chaînes avec des programmes comme la *Star Academy, A prendre ou à laisser, Attention à la marche, Miss France, Les enfants de la télé, T'empêches tout le monde de dormir*, *Secret Story* et même + *Clair*.

A 52 ans, le Tropézien n'abandonne pas le paquebot RTBF en pleine mer même si, évidemment, il n'ira pas au bout de son mandat de six ans. D'ailleurs, il ne quittera la RTBF que le 31 août prochain. *« Ce n'est évidemment pas une nouvelle que l'on accueille avec plaisir, même si cela démontre que nous ne nous étions pas trompés lorsque nous l'avons recruté voici deux ans*, dit l'administrateur général de la RTBF, Jean-Paul Philippot. *Ma première préoccupation a été de savoir ce que nous allions faire pour les grilles de la rentrée de septembre. Et Yves m'a répondu spontanément qu'il allait les faire avec son équipe et les présenter à la presse fin août. Ce n'est qu'après cette échéance qu'il rejoindra Endemol France. »*

Désigner son successeur

Par ailleurs, Bigot conservera, deux ans durant, une mission de conseil auprès de la RTBF. D'ici à septembre, il participera à la désignation de son successeur, qui sera choisi via un appel à candidatures lancé d'ici une semaine. Bigot appartiendra effectivement au collège d'experts mis en place à cet effet en compagnie du Suisse Gilles Marchand (TSR) et du Flamand Aimé Van Hecke (ex-VRT, Sanoma).

Pour Yves Bigot, il s'agit d'une nouvelle ligne sur un CV impressionnant qui l'a amené à travailler dans la radio et la presse écrite, à diriger des maisons de disques et à faire de la télé (il a été directeur des variétés et des programmes de France 2, et directeur général adjoint chargé des programmes de France 4).

« C'est l'amour de l'aventure et le goût du risque qui m'avaient amené en Belgique et j'y ai passé deux années merveilleuses, explique-t-il. *Mais le challenge que me propose Virginie Calmels, la présidente d'Endemol France, est ambitieux et enthousiasmant. Je ne remplace pas Alexia Laroche-Joubert* (ndlr : la directrice de la *Star Ac* vient de quitter son poste de directrice des programmes d'Endemol France pour rejoindre son ancien patron, Stéphane Courbit) *puisque je prends un poste nouvellement créé, celui de directeur général adjoint chargé des programmes. Tous les programmes, donc y compris la téléréalité mais pas seulement la téléréalité, ce qui ne m'aurait pas intéressé. Endemol ne serait pas venu me chercher moi pour ça. »*

Quant au bilan de Bigot à Reyers, il reste globalement très positif malgré quelques échecs. C'est en tout cas l'avis de Jean-Paul Philippot : *« La Une a stabilisé ses audiences, la Deux les a augmentées de 43 %. On a investi dans la production belge, comme* Melting-pot café, *lancé Arte Belgique, des talk-shows, le* 12 minutes *et mis* Matin première *en télé. Surtout, la RTBF a renoué avec des événements* (Stars of Europe, Tenue de soirée) *et avec une politique remarquable en termes d'acquisition de fictions et de droits sportifs. Nous aurions pu travailler ensemble longtemps encore…* »

COMMENTAIRE :

La Deux augmente son audience de 43 % !! En fait, elle passe de 3 à 4,5 % !!!!

Halte aux idées reçues

Les jeux vidéo améliorent le niveau scolaire

Les parents qui se lamentent en voyant leur progéniture affalée sur le canapé du salon à jouer à des jeux vidéo et qui n'hésitent pas à dire *" Et tes devoirs ? Travaille plutôt à l'école "* seront sans doute forcés de réviser leurs positions, car non, le jeu vidéo ne nuit pas aux performances scolaires, bien au contraire, c'est du moins ce que tend à prouver une étude consacrée à ce sujet et dont nous nous faisons l'écho ici.
Le professeur Astreau, chercheur indépendant attaché à l'Institut d'études pédiatriques de Niort a effectué des mesures très précises à ce sujet, mesure dont il a tiré des conclusions qui en étonneront plus d'un.
Tout d'abord, le rapport entre la fréquence d'utilisation des consoles de jeux vidéo et les performances scolaires chez les enfants comme chez les adolescents, loin d'être défavorable, s'avère au contraire étonnamment bon. Qu'on en juge : un enfant qui ne dispose pas d'une console de jeux vidéos chez lui a quatre fois plus de chances de redoubler au cours de son parcours scolaire qu'un enfant qui dispose d'un tel équipement. Sans entrer plus avant dans les détails, on peut retenir les grandes lignes de cette étude, selon laquelle la précocité de l'usage des consoles, le temps quotidien qui y est consacré et la variété des jeux ont tous une influence positive directe indubitable sur les performances scolaires des enfants et des adolescents, à l'exception des jeux en réseau qui, au delà d'un certain nombre d'heures de pratique hebdomadaires s'avèrent néfastes au dossier scolaire.
Après avoir analysé ses observations, Gérard Astreau s'est demandé ce qui pouvait expliquer le bon niveau scolaire des jeunes joueurs.

" Il semble qu'il y ait un faisceau de raisons à cela... " explique-t-il, *" et non une raison unique "*. La première explication qui vient à l'esprit, c'est la stimulation intellectuelle. Forcé de réfléchir rapidement, de résoudre des énigmes, de se montrer observateur, le joueur développe des capacités intellectuelles intéressantes. Mais, tempère le chercheur, *" les aptitudes développées par la pratique du jeu vidéo ont un usage principal : jouer aux jeux vidéo. C'est le serpent qui se mord la queue ! "*. L'explication se trouve donc ailleurs.
Les joueurs apprennent très tôt l'immobilité spatiale. Ils peuvent rester assis au même endroit sans éprouver la moindre lassitude, plongés dans un état de concentration qu'ils parviennent à tenir pendant des heures. Une telle aptitude s'avère capitale en situation scolaire où, de la même manière, on exige des écoliers d'être physiquement apathique et intellectuellement vifs et concentrés.
Des électrodes ont été disposées sur le cuir chevelu des sujets étudiés afin d'enregistrer dans leur détail leur activité cérébrale lorsqu'ils jouaient aux jeux vidéo.

Il s'avère à l'analyse de ces tests que le cerveau est fortement stimulé par les jeux vidéo dans l'hippocampe et dans la région du para-hippocampe, des zones partiellement dédiées à l'appréhension cognitive de l'environnement spatial et à l'orientation. Or, en agissant sur l'étendue et la qualité de la "carte mentale" des joueurs, les jeux vidéo ne stimulent pas uniquement leur sens de l'orientation, ils agissent aussi fortement sur la mémoire, au point qu'un chercheur de l'Université de Californie à Los Angeles teste actuellement un traitement de la maladie d'Alzheimer à base de jeu vidéo.

Tout ceci explique donc les résultats obtenus. Cependant, G. Astreau se refuse à dresser un tableau idyllique du sujet : *" 12% des joueurs se plongent involontairement en apnée lorsqu'ils jouent. Cette privation temporaire d'oxygène peut remettre en cause le bénéfice scolaire obtenu par la pratique des jeux vidéo. De plus les enfants épileptiques risquent des crises lorsqu'ils sont trop longtemps confrontés à des images frénétiques et lumineuses, et ces crises peuvent imposer une médication importante qui à son tour peut grever les performances scolaires de l'enfant "*. Enfin, rappelle-t-il, *" les notes des passionnés de jeux vidéos sont généralement médiocres en cours d'éducation physique et sportive "*.

Jenny L. Voight
Spécialiste des souris

Commentaire : *Qu'en-pensez-vous ?*

BON COURAGE !!

Ça tousse un peu partout…

Infections respiratoires en forte hausse… comme chaque année à la rentrée

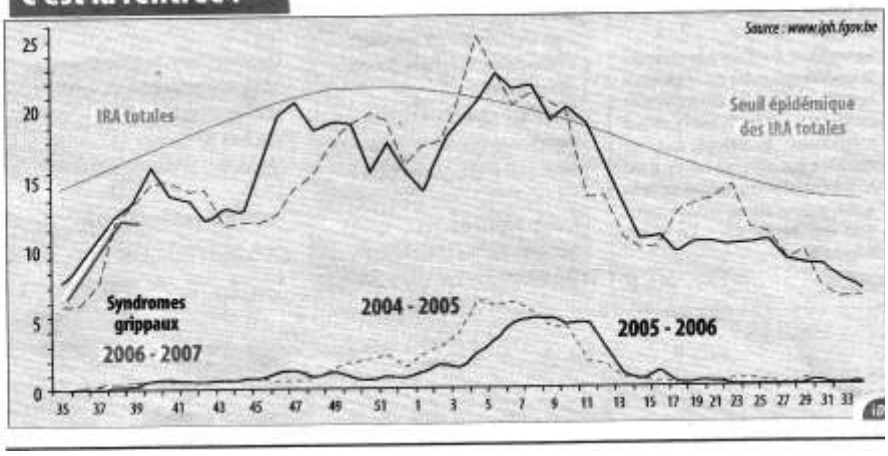
BRUXELLES ▽ Rhumes, bronchites, pharyngites… : depuis une ou deux semaines, les infections respiratoires se multiplient en Belgique. La faute aux changements de conditions climatiques, qui ont basculé de manière spectaculaire depuis la fin septembre ? Pas du tout, du moins si on s'en réfère aux observations des années antérieures.

Selon les données diffusées par l'Institut scientifique de santé publique (ISP), les infections respiratoires aiguës motivent, en ce moment, 12 % des consultations et visites des médecins généralistes. Nous frôlons, en fait, le seuil épidémique, fixé à 15 % environ.

L'explication ? La rentrée scolaire ! Le regroupement des enfants déclenche – comme le montrent clairement les statistiques – une flambée de transmission de germes infectieux, qu'ils se transmettent l'un à l'autre, souvent sans qu'ils développent d'ailleurs de troubles bien spécifiques. Et de retour à la maison, ce sont les parents qui sont à leur tour contaminés, avec des conséquences bien plus ennuyeuses. Et puis on va au boulot, et le tour est joué.

J. M.

C'est la rentrée !



BIBLIOGRAPHIE

- *Adam A., Lousberg F., Espace Math 54, De Boeck, 1999.*
- *Albert A., Biostatistique, Université de Liège, Faculté de Médecine, 2005.*
- *d'Odemont P., Statistique biomédicale, Haute Ecole de la Province de Liège.*
- *Droesbeke J.-J., Eléments de Statistiques, Editions Ellipse, 1997.*
- *Jaffard P., Initiation aux méthodes de la statistique et du calcul des probabilités, Masson, 1986.*
- *Lambermont M-F., cours de statistique, graduat en informatique, IPEPS Seraing, 2004.*
- *Masiéri, W., Notions essentielles de statistique et calcul des probabilités, travaux pratiques, Sirey, 1976.*
- *Monfort F., Eléments du calcul des probabilités et des méthodes statistiques. Ulg. 1970*
- *Moroney M.J., Comprendre la statistique, Marabout Université, 1970.*
- *Poinsot D., Statistiques pour statophobes, cours en ligne, 2004.*
- *Rateau P., Méthode et statistique expérimentales en sciences humaines, Edition Ellipses, 2001.*
- *SBPMef, Explorations didactiques, Statistiques 1ère approche, dossier n°6, 1999.*
- *Van Vyve-Genette A., Gohy J-M., Feytmans E., Statistique élémentaire pour les sciences bio-médicales, De Boeck Université, 1988.*
- *Tassi P., Méthodes Statistiques, Ed. Economica, 1989.*
- *Wonnacott H. et J., Statistique : Economie, Gestion, Sciences, Médecine, Ed. Economica, 4ème édition, 1995.*