Quelques éléments de statistiques

Avant-propos

Ces quelques éléments concernent essentiellement les statistiques au programme dans l'enseignement secondaire.

Ils prennent appui sur les documents utilisés par M. ARTIGUES, IA-IPR de Mathématiques, lors d'un stage – septembre 2001 à Buenos-Aires – à destination des professeurs de mathématiques des établissements français d'Amérique du Sud.

Deux objectifs essentiels présidaient à ce stage :

- faire appréhender la démarche propre aux statistiques à travers leur enseignement dans le secondaire,
- engager une réflexion sur la pratique actuelle (ou l'absence de pratique...) dans cet enseignement.

La plupart des documents qui suivent sont extraits ou largement inspirés des ouvrages cités dans la bibliographie qui est loin d'être exhaustive sur le sujet. Les lecteurs sont invités à les consulter sans modération afin de compléter le point de vue très parcellaire développé dans ces quelques pages.

Si quelques exemples de simulation avaient été donnés au cours du stage, ils n'ont pas été repris ici. Les travaux de nombreux collègues ou des IREM ont abondamment alimenté différents sites en la matière.

Sommaire

Préambule

I. Qu'est-ce que la statistique ?

Quelques citations amusantes glanées ici et là :

Edmond et Jules de GONCOURT :	« ***La statistique*** *est la première des sciences inexactes.* »
Alphonse ALLAIS :	« ***La statistique*** *a démontré que la mortalité dans l'armée augmente sensiblement en temps de guerre.* »
Mark TWAIN :	« *Il existe trois sortes de mensonges : les mensonges, les sacrés mensonges et* ***les statistiques*** »
Adolphe THIERS :	« *... L'art de préciser les choses que l'on ignore* »
LAVELEYE :	« *... L'art de mentir mathématiquement* »
MACAULEY :	« *Les chiffres disent toujours ce que souhaite l'homme habile qui sait en jouer* »
Louis ARMAND :	« ***Les statistiques****, c'est comme le bikini, ça montre tout, mais ça cache l'essentiel* »

Plus sérieusement :

ENCYCLOPEDIA UNIVERSALIS : « *Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation* »

PETIT LAROUSSE : (du latin status, état) « *Ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions* »

Daniel SCHWARTZ : « *La statistique est un mode de pensée permettant de recueillir, de traiter et d'interpréter les données qu'on rencontre dans divers domaines, et tout particulièrement dans les sciences de la vie, du fait que ces données présentent une caractéristique essentielle : la variabilité.* »

La variabilité est un concept fondamental en statistique : des individus apparemment semblables peuvent prendre des valeurs différentes.

Quelques exemples :

- Le nombre de quartiers dans le fruit du coquelicot : le plus souvent égal à 13, il peut varier de 6 à 20 !
- Le temps de latence d'une maladie virale du tabac (mosaïque) : il varie de 13 à 34 jours.
- En cours de fabrication, la longueur d'un profilé aluminium n'est jamais parfaitement constante.

...

Pour l'essentiel, l'analyse statistique est **une étude de la variabilité**.

2. Statistique et probabilités

La **théorie des probabilités** modélise des phénomènes où le « hasard » intervient. Comme toute théorie mathématique, elle est basée sur une axiomatique et se développe de façon autonome par rapport à la réalité physique.

La **statistique** repose, elle, sur l'observation de phénomènes.

Quels liens entre la statistique et les probabilités ?

G. Saporta en voit schématiquement trois.

- Les données observées sont souvent imprécises ou entachées d'erreur. La théorie des probabilités permet de représenter les déviations entre vraies valeurs et valeurs observées comme des variables aléatoires.
- On constate souvent que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques proposés par la théorie des probabilités (lois de probabilité).
- Les échantillons d'individus observés sont la plupart du temps tirés au hasard dans la population, ceci pour assurer mathématiquement leur représentativité : si le tirage est fait de manière équiprobable chaque individu de la population a une probabilité constante et bien définie d'appartenir à l'échantillon. Les caractéristiques observées sur l'échantillon deviennent, grâce à ce tirage au sort, des variables aléatoires et le calcul des probabilités permet d'étudier leurs répartitions.

Dans les deux premiers cas, la théorie des probabilités propose des modèles (simplificateurs mais peut-être contestables), du comportement d'un phénomène (par exemple la durée de vie X d'un composant électronique suit une loi exponentielle, c'est à dire $P(X > x) = \exp(-\lambda x)$).

Dans le dernier cas, la théorie des probabilités fournit des théorèmes si le processus d'échantillonnage est respecté : ainsi le **théorème central limite** permet d'établir que la moyenne $\overline{X}$ d'une variable numérique mesurée sur n individus s'écarte de la moyenne m de la population entière selon une loi approximativement gaussienne.

Le **calcul des probabilités** est donc un des outils essentiels de la statistique pour pouvoir extrapoler à la population les résultats constatés sur l'échantillon.

3. La démarche statistique

On distingue deux grands aspects :

- L'aspect exploratoire : **statistique descriptive**

La statistique descriptive a pour objectif de synthétiser, résumer, structurer l'information contenue dans les données concernant un phénomène étudié. On utilise des représentations graphiques ou des tableaux et on calcule quelques indicateurs. Le rôle des modèles probabilistes est quasiment inexistant.

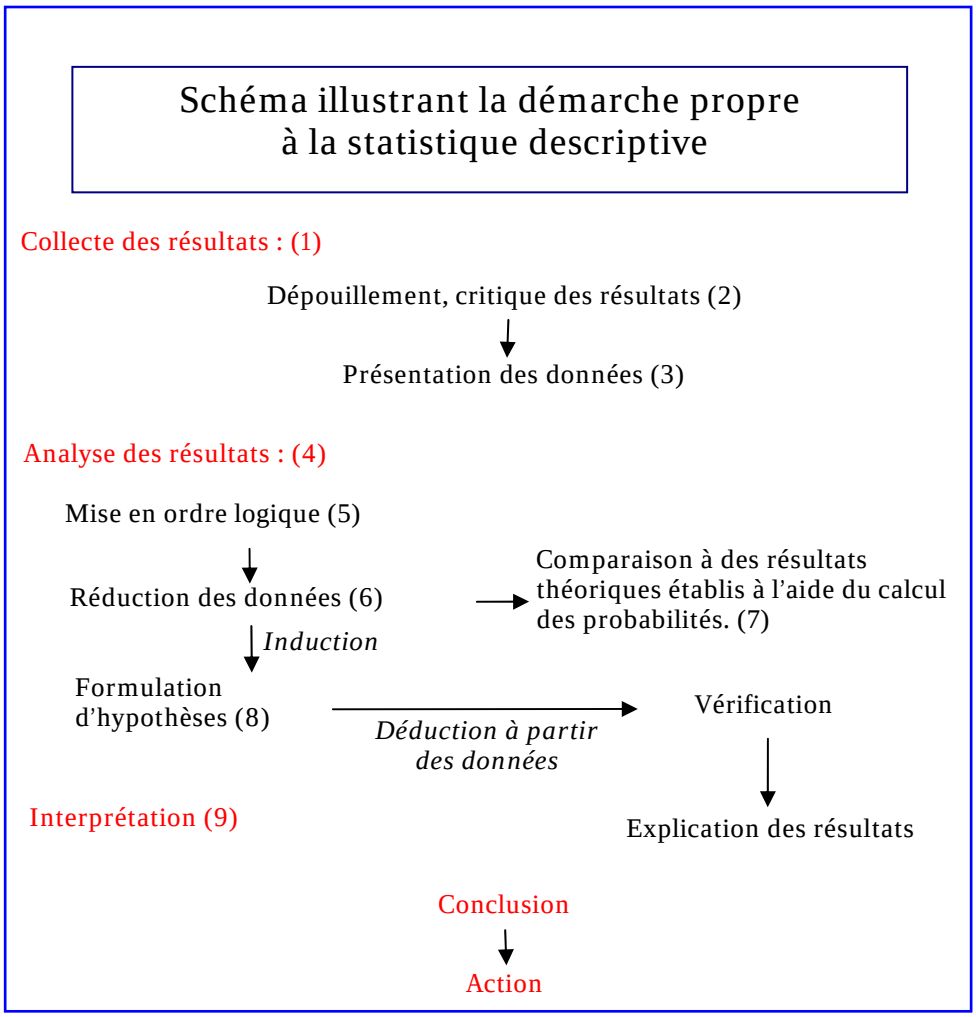
- L'aspect décisionnel : statistique inférentielle

Dans son document, ***La statistique au collège***, Y. Olivier décrit la statistique inférentielle de la façon suivante :

En général, les ensembles d'observation correspondent à des échantillons liés au hasard (on dit aussi présentant un caractère aléatoire) et l'on essaie de modéliser le phénomène à l'aide de modèles probabilistes (on s'appuie sur certaines lois de probabilités classiques). Cela permettra sinon des prévisions tout du moins des présomptions qui sont précieuses dans l'étude de certains faits (sociaux, économiques ou industriels). L'**inférence** consiste donc à étudier les propriétés d'un échantillon représentatif d'un ensemble plus vaste et à généraliser ces propriétés en se souciant des questions suivantes : les faits étudiés sont-ils

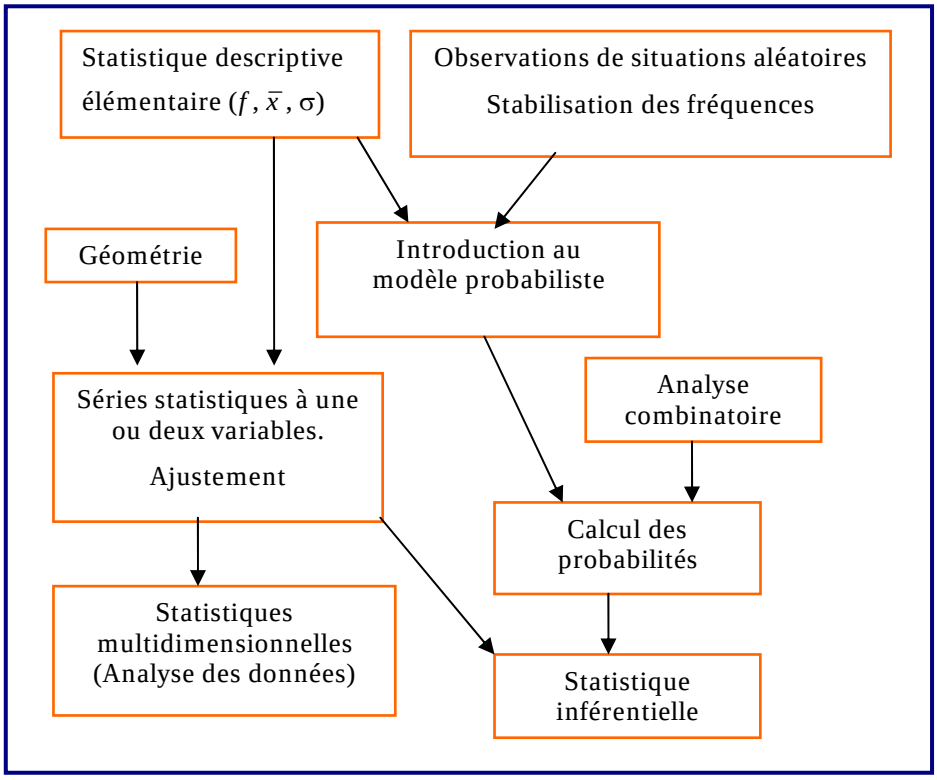
significatifs et sont-ils caractéristiques de propriétés plus générales ? Elles permettent de prendre de «bonnes» décisions malgré la présence d'incertitudes comme dans la recherche de qualité de produit fabriqué ou comme dans les analyses en laboratoires pharmaceutiques par exemple.

La démarche du statisticien peut être illustrée par le schéma suivant extrait de la brochure Inter-IREM *« Liaison Collège-Seconde »*.



Chapitre I
La statistique dans l'enseignement secondaire

1. Le schéma général



2. Les programmes

Il serait trop long ici de rappeler le contenu des programmes de statistiques de la sixième à la terminale. Ci-dessous un certain nombre de liens vers le site maths du rectorat de l'académie de Bordeaux.

Les statistiques au collège (PDF, 63 Ko)

Les statistiques au lycée

- Seconde (PDF, 71 Ko)
- Première L (PDF, 142 Ko)
- Première ES (PDF, 180 Ko)
- Première S (PDF, 200 Ko)
- Première STI (PDF, 44 Ko)
- Terminale ES (PDF, 434 Ko)
- Terminale S (PDF, 392 Ko)
- Terminale STI (PDF, 43 Ko)

Chapitre 2
Les graphiques

I. Généralités

C'est la partie des statistiques la moins souvent oubliée dans l'enseignement secondaire car elle mobilise la notion de proportionnalité sous ses différentes formes.

Les graphiques sont de natures très diverses : un tableur comme Excel offre de multiples possibilités.

Ils posent cependant de nombreuses interrogations qui ont été formulées par J.C. Girard dans la revue *Repères* n° 23 (avril 1996) :

- Sur le sens des graphiques
  - Quel est l'avantage d'un graphique sur un tableau de valeurs ?
  - Le graphique sert-il d'illustration ou permet-il de découvrir une structure des données que le tableau ne mettait pas en évidence ?
  - Peut-on repasser du graphique au tableau ?
  - Quelle perception de la réalité a-t-on en regardant un graphique ?
  - Pourquoi tel graphique plutôt que tel autre ? Dans quels cas, chacun est-il pertinent ?
- Sur des notions qui renvoient à différents domaines mathématiques
  - Les camemberts utilisent la notion d'angle et de mesure d'angle qui ne sont pas toujours acquises. Comment peut-on prendre en compte cet état de fait ? Que représente le disque complet ? Autrement dit, quel est l'ensemble sur lequel on calcule les *pourcentages ?*
  - Les histogrammes et les graphiques en barres ou en bâtons utilisent une échelle verticale sur laquelle on porte des effectifs ou des *fréquences*. Sur quel ensemble de référence ces fréquences ont-elles été calculées ?
  - Lorsqu'on représente des variations, sont-elles calculées de façon absolue ou relativement à une valeur de référence ?

Exemple 1 *(extrait d'un livre de CM1 :Objectif Calcul -Éditions Hatier)*

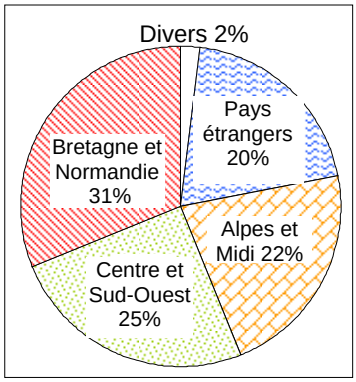
Le livre pose les questions suivantes :

1) Observe ce graphique.

2) Essaie de le lire.

3) Quels renseignements donne-t-il ?

4) Essaie de traduire ce graphique par un tableau de nombres.



On pourrait aussi demander (en CM2, en 6e ou plus tard !) :

– Sur quoi sont calculés les pourcentages ?

– Est-ce 20% des français qui partent en vacances à l'étranger ou 20% de ceux qui partent en vacances qui vont à l'étranger ?

– Peut-on calculer combien de français partent à l'étranger ? combien partent en vacances ? etc.

Exemple 2

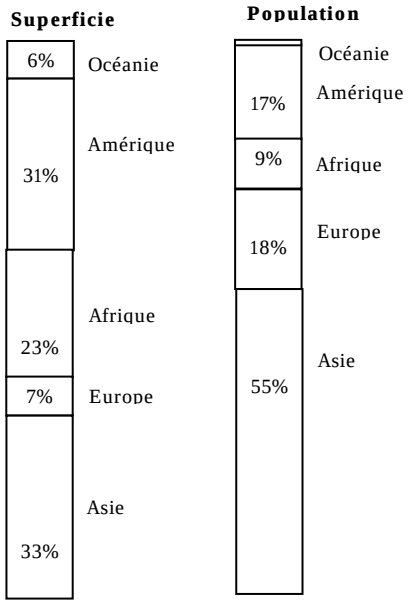
(extrait du même livre)

– Que représente la longueur de chaque barre ?

– Sur quoi ont été calculés les pourcentages ?

– Peut-on comparer ces différents pourcentages ?

– Quelle idée veut donner ce graphique ?



Superficie et population des continents

L'objectif essentiel des graphiques est de représenter la série statistique.

Comme toute représentation, ces graphiques doivent être :

- *lisibles* (les données représentées doivent pouvoir être lues),
- *fidèles* (la réalité des données ne doivent pas être déformées par la réalisation du graphique),
- *autosuffisants* (tous les renseignements doivent être mis dans la légende y compris l'ensemble de référence).

On l'a compris **chaque graphique** doit être pertinent par rapport aux données et à l'objectif poursuivi.

2. Cas d'un caractère qualitatif

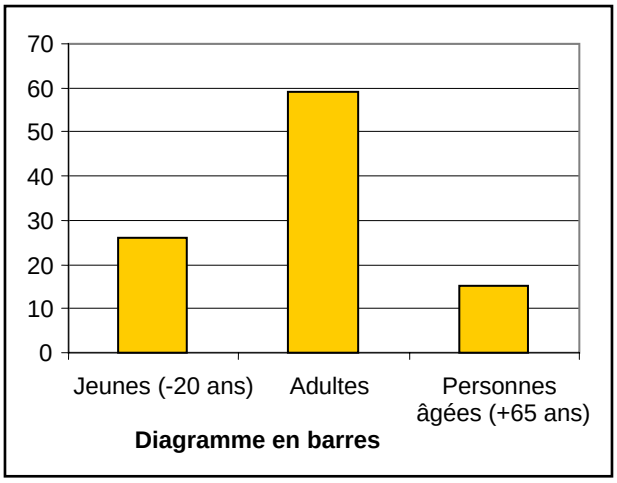


Diagramme en barres

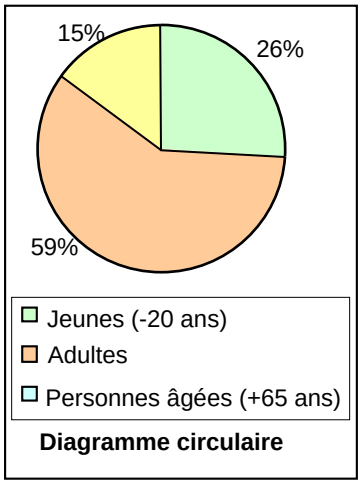


Diagramme circulaire

Le *diagramme en barres* permet de comparer les parties entre elles. Lorsque les modalités sont ordonnées par effectifs décroissants, on obtient un diagramme dit de Pareto. La longueur de la barre est proportionnelle aux effectifs ou à la fréquence.

Le *diagramme à secteurs* (circulaire ou semi-circulaire) : il permet de comparer la partie au tout. L'aire du secteur est proportionnelle à l'effectif ou à la fréquence.

Si l'on veut comparer plusieurs diagrammes à secteurs entre eux (sur plusieurs années par exemple), les rayons doivent être proportionnels à la racine carrée de l'effectif total.

Ces deux types de représentation nécessitent d'avoir une bonne perception mentale d'un pourcentage soit dans le domaine des longueurs soit dans celui des aires.

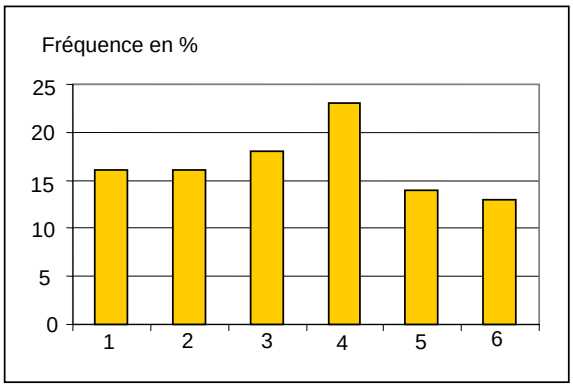
3. Cas d'un caractère quantitatif discret

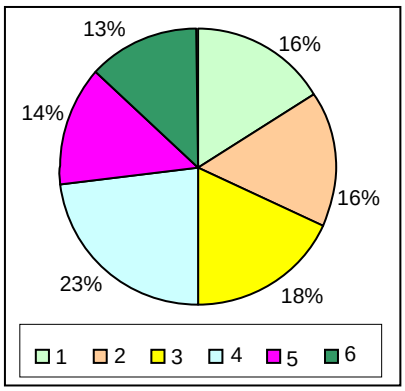
On utilisera :

Le *diagramme en bâtons (ou à bandes) :* il permet de facilement comparer les effectifs ou les fréquences entre eux.

> Relier les sommets entre eux pour constituer ce qui est nommé parfois le polygone des effectifs n'a pas de sens : en effet les points des segments autres que les extrémités n'ont aucune signification !
>
> On peut également utiliser un *diagramme à secteurs* si on souhaite comparer la partie au tout. Toutefois le diagramme semi-circulaire doit être privilégié pour respecter une structure d'ordre à mettre en évidence dans les modalités.

Exemple : Série de 100 lancers d'un dé équilibré





4. Cas d'un caractère quantitatif continu

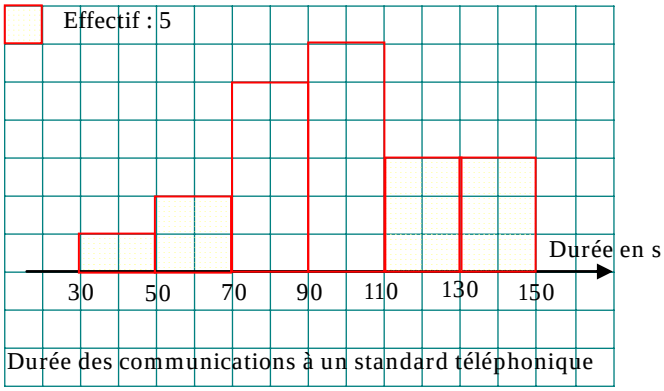
4.1. On utilise un histogramme

Il est constitué par des rectangles contigus ayant pour base chacune des classes et une aire proportionnelle à l'effectif ou à la fréquence de la classe correspondante.

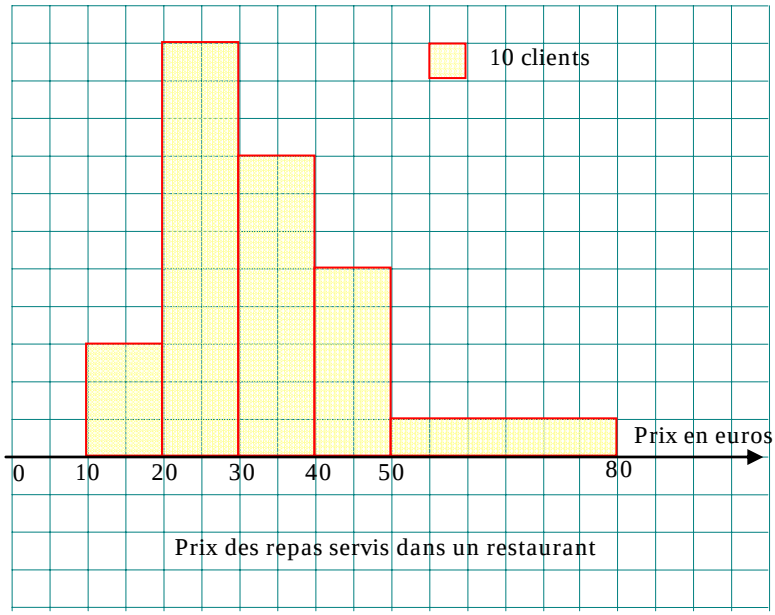
Si les amplitudes sont toutes égales, la hauteur du rectangle est proportionnelle à l'effectif (ou à la fréquence).

Si les amplitudes sont inégales, la hauteur est proportionnelle à la densité de la classe.

> En théorie, l'histogramme est la représentation graphique de la densité en tant que fonction des différents intervalles de la partition de l'ensemble des modalités.
>
> En pratique, on indique une unité d'aire correspondante à un certain effectif (ou à une certaine fréquence).



Durée des communications à un standard téléphonique



Prix des repas servis dans un restaurant

4.2. Sur la notion d'histogramme

On a trop tendance à considérer l'histogramme comme une juxtaposition de rectangles dont l'intérêt se limite :

– à l'obtention d'un dessin,

– à satisfaire un item du programme.

Prenons l'exemple de la modélisation du contrôle du réglage d'une machine fabriquant des profilés aluminium.

Il paraît pertinent de choisir une **variable quantitative continue** comme la longueur ***X*** en cm des profilés. Il est non moins pertinent de choisir cette variable ***X* absolument continue**.

RAPPEL

Une variable X est absolument continue s'il existe une fonction f définie sur $\mathbb{R}$ telle que :

• f est positive sur $\mathbb{R}$,

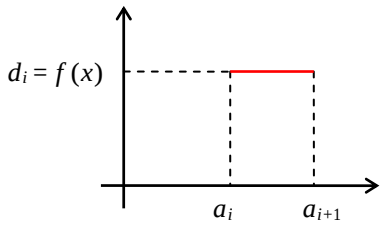
• f est continue sur $\mathbb{R}$ sauf peut-être en un nombre fini de points où elle admet une limite à droite et une limite à gauche,

• $\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t = 1$,

• La fonction de répartition F de X est liée à f par : $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$.

On dit que f est **une densité** de X. Abusivement, f est appelée **loi de *X*.**

Les données peuvent être placées dans des intervalles dont l'amplitude correspond à la précision de la machine (le mm par exemple). On choisit un nombre fini d'intervalles notés $[\ a_i\ ;\ a_{i+1}\ [$ pour $1 \leq i \leq k$.



A priori, tous les nombres réels appartenant à un tel intervalle ont la même chance d'être le résultat d'une mesure.

Cela se traduit pour le modèle adopté par une densité (de fréquence) constante sur chaque intervalle.

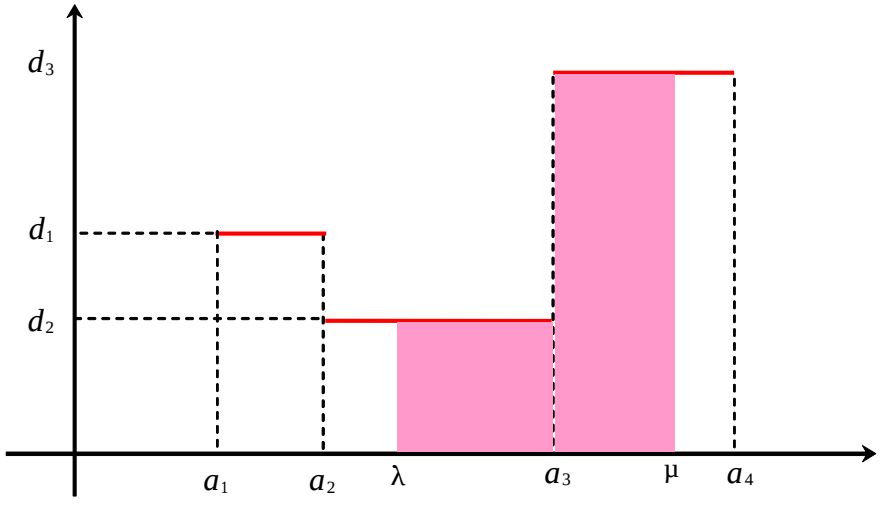
Pour estimer les valeurs d_i , on procède alors à un sondage en mesurant le plus grand nombre possible de profilés dans des conditions admises identiques. On obtient la série statistique des mesures qui sont réparties entre les différents intervalles :

Intervalles	$]-\infty\ ;\ a_1[$	$[a_1\ ;\ a_2[$	...	$[a_i\ ;\ a_{i+1}[$	...	$[a_k\ ;\ a_{k+1}\ [$	$[a_{k+1};\ +\infty[$
Effectifs	0	n_1	...	n_i	...	n_k	0
Fréquences	0	f_1	...	f_i	...	f_k	0

On a alors : $d_i = \dfrac{f_i}{a_{i+1} - a_i}$.

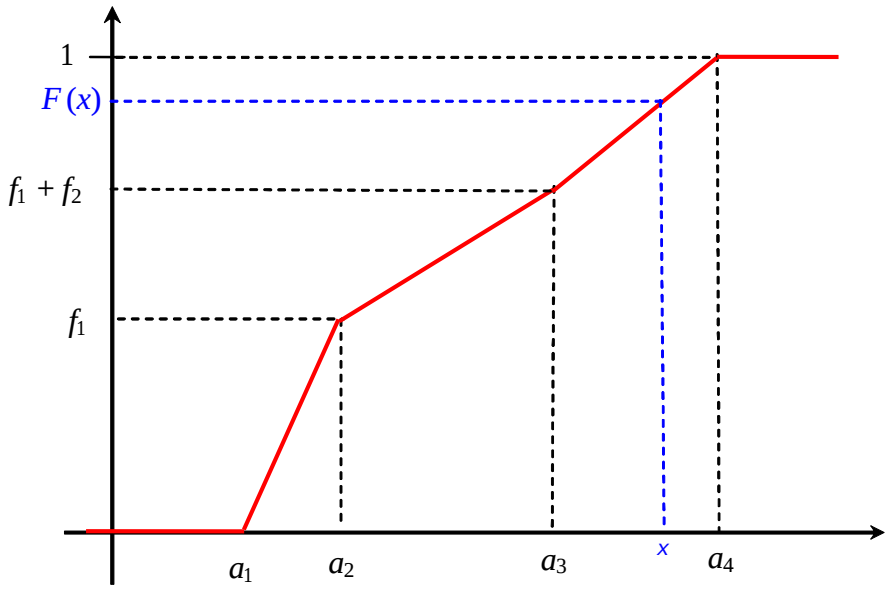
> REMARQUE
>
> L'unité sur l'axe des ordonnées est alors : **% /cm** . Il est donc erroné d'interpréter l'axe des ordonnées comme axe des effectifs ou des fréquences, ce qui se fait malheureusement souvent.

La fonction densité, obtenue par observation statistique et représentée par l'histogramme, est une fonction constante par intervalles :

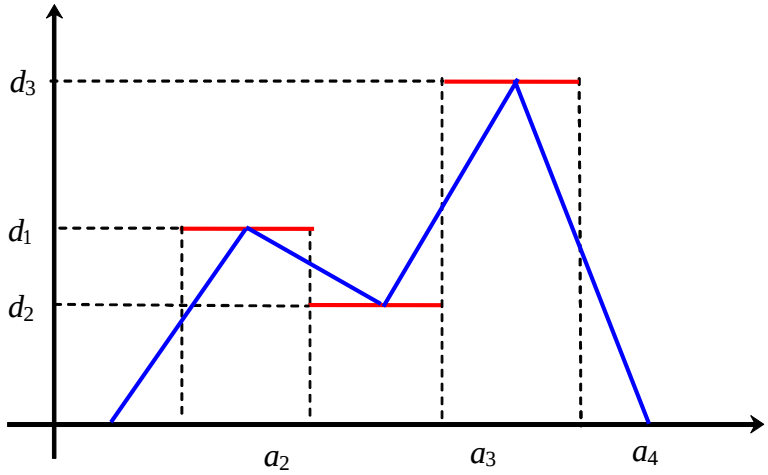


La fréquence possible des profilés dont la longueur est comprise entre λ et μ est estimée par l'aire du domaine rosé.

On pourrait aussi estimer la fréquence possible $F(x)$ des profilés dont la longueur est inférieure ou égale à x. Cette fonction F est la **fonction cumulative croissante** ou **fonction de répartition** de X. C'est une primitive de la densité f. La courbe obtenue est une fonction continue et affine par intervalle.



Dans de nombreux ouvrages sur la statistique, on trouve la notion de *polygone des effectifs ou des fréquences*. Dans le cas où les classes ont même amplitude, on suggère de tracer sur l'histogramme une ligne brisée reliant les milieux des côtés supérieurs des rectangles de chaque classe. Lorsque les classes n'ont pas la même amplitude, on se ramène au cas précédent par un découpage *ad-hoc* des classes.



Il s'agit de remplacer l'histogramme précédent par un autre histogramme représenté par une courbe limitant un domaine d'aire égale à 1 (ou à l'effectif total) et **continue**. L'idée générale est de lisser la courbe de densité pour la comparer à celle de variables connues servant de modèles et le procédé suggéré précédemment fournit **une** solution à ce problème.

Pour conclure, on peut retenir cette définition proposée par J.C. Régnier :

Soit X une variable statistique (resp. aléatoire) absolument continue de densité f. L'histogramme est la surface limitée par la représentation graphique de f et l'axe des abscisses.

5. Graphiques en tiges et feuilles (stem and leaf)

La réalisation d'un histogramme n'est pas exempte de difficultés pour les élèves : notion d'intervalle semi-ouvert, utilisation d'une échelle, proportionnalité, concept de densité...

Le graphique tiges-feuilles (John W. Tukey) est assez proche de l'histogramme sans avoir les obstacles précédemment cités.

Prenons l'exemple cité par J.C. Girard : les données suivantes représentent la dureté, en indice Rockwell, de 60 pièces mécaniques après trempage.

Dans un premier temps les valeurs sont arrondies à l'unité pour pouvoir travailler sur des entiers.

51	54	55	56	57	58
51	54	55	56	57	58
52	54	55	56	57	58
53	55	55	56	57	58
53	55	55	56	57	59
53	55	55	56	57	59
54	55	55	56	57	59
54	55	55	56	57	60
54	55	56	56	57	60
54	55	56	57	58	61

	Effectif	
51	**2**	★★
52	**1**	★
53	**3**	★★★
54	**7**	★★★★★★★
55	**15**	★★★★★★★★★★★★★★★
56	**11**	★★★★★★★★★★★
57	**10**	★★★★★★★★★★
58	**5**	★★★★★
59	**3**	★★★
60	**2**	★★
61	**1**	★

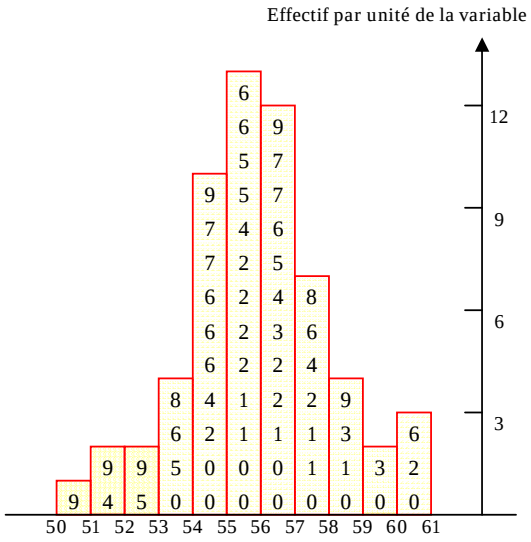
On travaille maintenant sur les valeurs au dixième : on obtient 44 valeurs différentes. Un schéma analogue au précédent ne présente guère d'intérêt. Le graphique « Stem and leaf » constitue un préalable intéressant avant d'aborder la notion d'histogramme.

50,9	54,0	55,0	55,6	56,6	57,8
51,4	54,2	55,1	55,6	56,7	58,0
51,9	54,4	55,1	56,0	56,7	58,1
52,5	54,6	55,2	56,0	56,9	58,3
52,9	54,6	55,2	56,1	57,0	58,9
53,0	54,6	55,2	56,2	57,1	59,0
53,5	54,7	55,2	56,2	57,1	59,3
53,6	54,7	55,4	56,3	57,2	60,0
53,8	54,9	55,5	56,4	57,4	60,2
54,0	55,0	55,5	56,5	57,6	60,6

50		9	1
51		49	3
52		59	5
53		0568	9
54		0024666779	19
55		0011222245566	(13)
56		001223456779	28
57		0112468	16
58		0139	9
59		03	5
60		026	3

Tige 52, feuilles 5 et 9

La simplicité de ce graphique est évidente, il revient à regrouper les valeurs dans des intervalles d'amplitude 1 ([50 ; 51 [, [51 ; 52 [, ...). Par rapport à un histogramme, il a l'avantage de ne perdre aucune information sur les données de départ. On n'est cependant pas loin de l'histogramme :



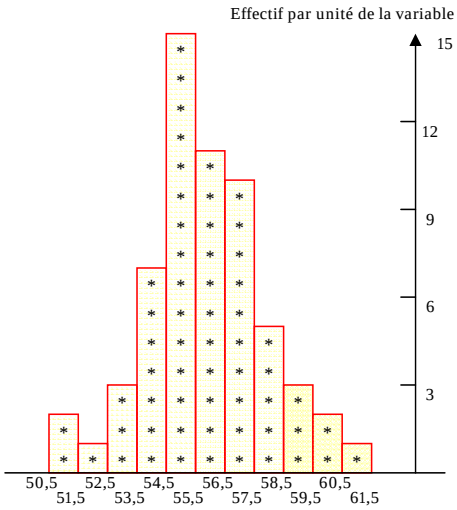
A partir des données précédentes, on peut aussi construire deux tiges par unité :

50	9	1
51	4	2
51	9	3
52		3
52	59	5
53	0	6
53	568	9
54	0024	13
54	666779	19
55	001122224	28
55	5566	(4)
56	0012234	28
56	56779	21
57	01124	16
57	68	11
58	013	9
58	9	6
59	03	5
59		3
60	02	3
60	6	1

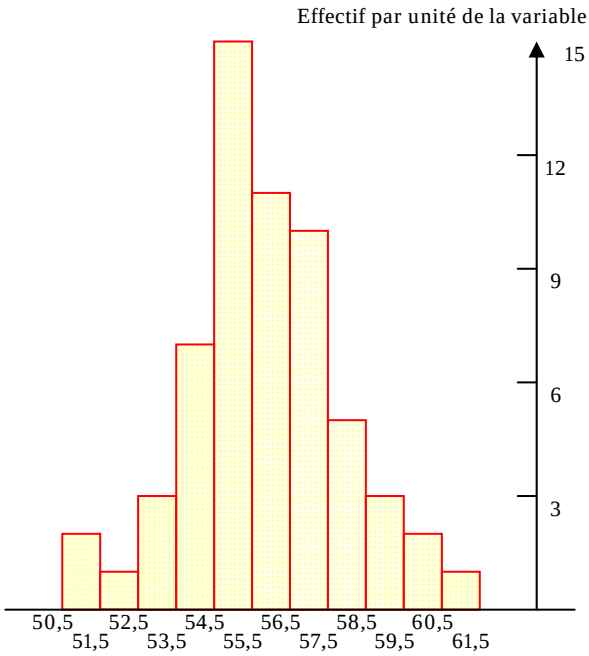
La première tige correspond à l'intervalle [50,5 ; 51 [,

la seconde à l'intervalle [51 ; 51,5 [.

L'ensemble de ces deux tiges représentent les valeurs de l'intervalle [50,5 ; 51,5 [, soit les valeurs qui ont été arrondies à 51.



On peut alors élaborer un histogramme dont les classes ont une amplitude d'une unité.


Effectif par unité de la variable
15
12
9
6
3
50,5
51,5
52,5
53,5
54,5
55,5
56,5
57,5
58,5
59,5
60,5
61,5

Chapitre 3
Les indicateurs

On se place uniquement dans le cas d'une variable **quantitative.**

L'objectif est de résumer l'ensemble des observations par des indicateurs. Il est toujours insuffisant de résumer une série par un seul indicateur.

D'après Guy Brousseau, un modèle doit :

- représenter correctement les observations *(pertinence)*,
- être un résumé plus simple que les observations *(communicabilité)*,
- permettre de reconstituer au mieux l'ensemble des observations *(fidélité)*,
- permettre de comprendre les données, c'est-à-dire de les placer par rapport à des modèles familiers, universels et donc de permettre la comparaison avec d'autres modèles *(intelligibilité)*,
- être accessible au contrôle mathématique *(consistance)*.

I. Les caractéristiques de position ou de tendance centrale

I.1. Le mode

Pour une variable statistique discrète, **le mode** est la valeur la plus fréquente.

Lorsque la variable est continue, on parle de **classe modale** : c'est la classe correspondant « au pic » de l'histogramme (G. Saporta), autrement dit c'est la classe pour laquelle d_i est maximale.

Plus généralement, si X est une variable statistique (resp. aléatoire) absolument continue de densité f, on appelle mode toute valeur de la variable pour laquelle f est maximum.

Bien entendu, il peut y avoir plusieurs valeurs (resp. classes) modales.

Exercice

Le but de cet exercice est de montrer que la classe modale n'est pas nécessairement celle dont l'effectif est le plus grand.

La répartition des salaires annuels, exprimés en milliers d'euros (k €) de 90 employés d'une entreprise est donnée dans le tableau suivant :

Salaires en k €	[13 ; 15[	[15 ;16[	[16 ;17[	[17 ;18[	[18 ;20[	[20 ;22[	[22 ;24[
Effectifs (n_i)	12	12	14	15	17	12	8

Tracer l'histogramme et déterminer la classe modale.

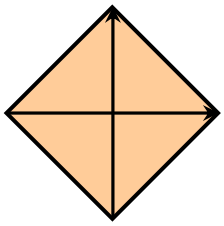
d'après Itinéraires en Statistiques et Probabilités (Ellipses)

1.2. Distances usuelles dans $\mathbb{R}^n$.

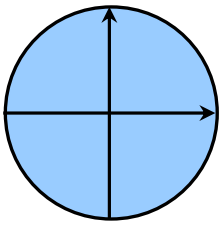
Pour les autres indicateurs de tendance centrale, il s'agit de résumer l'ensemble des observations par une valeur numérique relativement *proche*. La « proximité » se mesurant à l'aide de distances, il est utile de rappeler les distances usuelles dans $\mathbb{R}^n$.

Soit $X=(x_1,x_2,...,x_n)\in \mathrm{R}^n$.	Les distances associées sont :
On définit trois normes usuelles :	Pour tout $(X,Y)\in \mathrm{R}^n\times \mathrm{R}^n$:
$\|\| X \|\|_1=\sum_{i=1}^{n} \| x_i \|$	$d_1(X,Y)=\|\| X-Y \|\|_1$
$\|\| X \|\|_2=\left(\sum_{i=1}^{n} x_i^{\,2}\right)^{\frac{1}{2}}$	$d_2(X,Y)=\|\| X-Y \|\|_2$ (distance euclidienne)
$\|\| X \|\|_\infty=\underset{1\le i\le n}{\mathrm{Max}} \| x_i \|$	$d_\infty(X,Y)=\|\| X-Y \|\|_\infty$

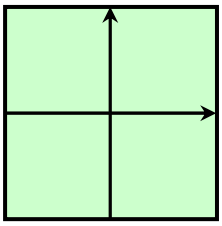
Pour chacune de ces distances, la boule unité dans le plan est donnée ci-après :



Distance d_1



Distance d_2



Distance d_∞

1.3. La moyenne

1.3.1. Cas d'une variable discrète

La moyenne $\bar{x}$ d'une série statistique est la somme de ses éléments divisée par leur nombre : $\bar{x}=\dfrac{x_1+x_2+...+x_n}{n}$.

Si $c_1,c_2,\dots,c_n$ sont les valeurs distinctes prises par les x_i et si n_i désigne l'effectif de la valeur c_i, on a :

$\bar{x}=\dfrac{\sum_{i=1}^{k} n_i c_i}{n}$ ou encore $\bar{x}=\sum_{i=1}^{k} f_i c_i$ avec $f_i=\dfrac{n_i}{n}$

La moyenne minimise la distance d_2. Cela signifie que, parmi les vecteurs "constants" $(a,a,\dots,a)$, le vecteur $\bar{X}=(\bar{x},\bar{x},\dots,\bar{x})$ est le plus proche du vecteur $X=(x_1,x_2,\dots,x_n)$ au sens de d_2.

Preuve :

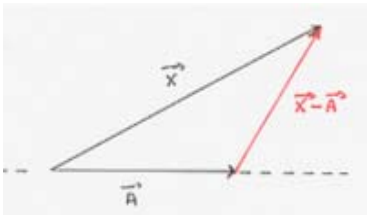
Soit $A=(a,a,...,a)$

$d_2^{\,2}(X,A)=\sum_{i=1}^{n}(x_i-a)^2=f(a)$

La fonction f est du second degré.

$$f(a) = na^2 - 2a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} {x_i}^2$$

f sera minimale pour $A = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$

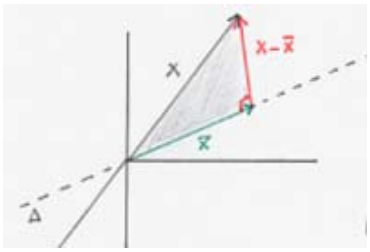


Alors $f(\overline{x}) = {d_2}^2(X, \overline{X}) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = nV(X)$

Le calcul de $f(\overline{x})$ donne alors :

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

Interprétation géométrique



${d_2}^2(X, A) = \| X - A \|_2$ sera minimum lorsque $\vec{A}$ est la projection orthogonale de $\vec{X}$ sur la droite vectorielle Δ engendrée par (1,1,...,1)

$\| X - \overline{X} \|_2^2 = nV(X) = ns^2$ où s est l'écrt-type.

Propriété : Linéarité de la moyenne

Soit deux séries statistiques $(x_1, x_2, \dots, x_n)$ et $(y_1, y_2, \dots, y_n)$ telles que, pour tout i de [1 ; n], on ait $y_i = ax_i + b$.
Alors $\overline{y} = a\overline{x} + b$.
En particulier, pour $a = 1$ et $b = -\overline{x}$, on a $\overline{y} = 0$. La nouvelle série a une moyenne nulle. On dit qu'on a centré les données x_i.

Propriété : Regroupement ou partition

Soit deux séries statistiques $(x_1, x_2, \dots, x_n)$ et $(y_1, y_2, \dots, y_m)$ de moyennes respectives $\overline{x}$ et $\overline{y}$.
Soit $(z_1, z_2, \dots, z_{n+m})$ la série obtenue par regroupement des deux séries précédentes et $\overline{z}$ sa moyenne.
Alors $\overline{z} = \frac{n\overline{x} + m\overline{y}}{n+m}$.

La preuve est immédiate.

Moyenne élaguée

La moyenne est sensible aux valeurs extrêmes. Pour pallier cet inconvénient, on peut décider de ne pas tenir compte des valeurs extrêmes dans le calcul de la moyenne.

Soit (x_1, x_2, ..., x_n) une série statistique et α un réel de [0 ; 1].
La moyenne élaguée de niveau $1-\alpha$ est la moyenne de la série privée d'un nombre de valeurs extrêmes égal à E($n\alpha$), soit à gauche, soit à droite, soit bilatéralement.
En principe α = 0,05 ou α =0,01.

1.3.2. Cas d'une variable continue

Le regroupement des valeurs en classes entraîne une perte d'information. Dans ce cas on ne peut calculer qu'**une valeur approchée de la moyenne.**

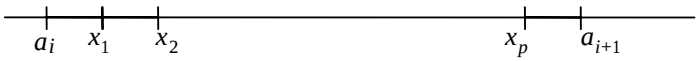
Pour trouver une telle valeur approchée, on considère que toutes les valeurs d'une classe sont rapportées au centre de cette classe. On remplace donc la série initiale par une série discrète.

Bien entendu, cette valeur approchée dépend de la nature du regroupement en classes effectué.

On pourrait prendre l'approximation de la distribution uniforme à l'intérieur d'une même classe : ce modèle conduit à la même valeur approchée que précédemment.

1.3.3. Comparaison des moyennes dans le cas d'une répartition uniforme

Les p valeurs $x_1, x_2, \ldots, x_p$ sont uniformément réparties sur l'intervalle $[\, a_i\, ,\, a_{i+1}\, [$ signifie :



$$\forall k \in [1\,;\,p],\ x_k = a_i + k\frac{(a_{i+1}-a_i)}{p+1}$$

Notons $\overline{x_i}$ la moyenne de $(x_1, x_2, \ldots, x_p)$.

$$p\,\overline{x_i} = \sum_{k=1}^{k=p} x_k = pa_i + \frac{(a_{i+1}-a_i)}{p+1}\sum_{k=1}^{p} k.$$

Or $\sum_{k=1}^{p} k = \frac{p(p+1)}{2}$. Donc $p\,\overline{x_i} = pa_i + p\,\frac{(a_{i+1}-a_i)}{2}$

soit $\overline{x_i} = \frac{a_i + a_{i+1}}{2}$. $\boxed{\overline{x_i} \text{ est donc le milieu du segment } [a_i\,;\,a_{i+1}[}$

Ainsi, quel que soit le modèle d'approximation choisi (regroupement au centre de classe ou répartition uniforme à l'intérieur de l'intervalle), la moyenne obtenue est la même.

Exercice

Lors du regroupement en classes de données abondantes, il y a évidemment perte d'information. Certes on peut espérer que les erreurs introduites par la concentration des données au centre de chaque classe se neutralisent dans le calcul de la moyenne, mais il n'en est pas toujours ainsi, comme le montre l'exemple suivant :

1. Dans une classe, la liste des notes obtenues à un devoir de mathématiques par les élèves classés par ordre alphabétique est la suivante :

8	16	9	18	9	11	9	13	7	3	14	7
10	10	10	17	13	14	10	13	5	15	13	19
10	6	12	5	12	1	9	9	8	8	4	

Déterminer une valeur approchée de la moyenne $\overline{x}$ de cette série statistique.

2. Le professeur décide de classer ses élèves en cinq groupes :

[0 ; 4 [	[4 ; 8 [	[8 ; 12 [	[12 ; 16 [	[16 ; 20 [
faible	médiocre	moyen	satisfaisant	très bon

Déterminer les effectifs de chaque classe.

En utilisant le centre des classes, calculer la moyenne $\bar{y}$ de cette série statistique.

3. Le professeur envisage une autre répartition et refait ses calculs avec le regroupement suivant :

[15 ; 20 [	[10 ; 15 [	[5 ; 10 [	[0 ; 5 [
très satisfaisant	convenable	insuffisant	très faible

Quelle est la moyenne $\bar{z}$ de cette dernière série statistique ?

Réponses : 1° : 10,2 ; 2° : 10,8 ; 3° : 10,5.

d'après Itinéraires en Statistiques et Probabilités (Ellipses)

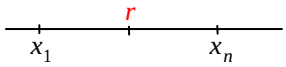
I.4. La moyenne des valeurs extrêmes

La moyenne des valeurs extrêmes d'une série $(x_1, x_2, \dots, x_n)$ est donnée par $r = \dfrac{\min(x_i) + \max(x_i)}{2}$.

Peu usité car très sensible aux valeurs extrêmes, elle minimise d_∞.

Cela signifie que, parmi les vecteurs "constants" $(a, a, \dots, a)$, le vecteur $(r, r, \dots, r)$ est le plus proche du vecteur $(x_1, x_2, \dots, x_n)$ au sens de d_∞.

r minimise la fonction f définie par $f(t) = \underset{1 \le i \le n}{\text{Max}} |t - x_i|$



I.5. La médiane

La médiane d'une série statistique ordonnée $(x_1, x_2, \dots, x_n)$ est x_p si $n = 2p+1$ et $\dfrac{x_p + x_{p+1}}{2}$ si $n = 2p$.

Dans le cas d'une variable continue, la pratique habituelle consiste à tracer la fonction de répartition en faisant l'hypothèse d'une répartition uniforme dans chaque intervalle puis d'exploiter cette représentation graphique pour déterminer l'antécédent de 0,5. D'après le document du GEPS sur les quantiles, cette pratique n'est pas usitée chez les statisticiens. Le GEPS préconise de parler de **classe médiane.**

« *La procédure qui consiste à tracer une courbe dite de fréquences cumulées croissante, continue, obtenue par interpolation linéaire à partir des valeurs F(ai) définies ci-dessus et à définir la médiane comme l'intersection de cette courbe avec la droite d'équation y=0,5, où avec une courbe analogue dite des fréquences cumulées décroissantes n'est pas une pratique usuelle en statistique et ne sera pas proposée au lycée.*

Si des données sont regroupées en classe, on parle de classe médiane. »

La médiane m_e minimise la distance d_1 . Cela signifie que, parmi les vecteurs « constants » $(a, a, \dots, a)$, le vecteur $M_e = (m_e, m_e, \dots, m_e)$ est le plus proche du vecteur $X = (x_1, x_2, \dots, x_n)$ au sens de d_1.

Preuve : $d_1(X,A)=\sum_{i=1}^{n}|a-x_i|=f(a)$

La fonction f est continue, dérivable sur chaque intervalle ne contenant pas x_i.

Pour tout $t \neq x_i$, $f'(t)$ sera négatif s'il y a plus de valeur x_i supérieures à t que de valeurs x_i inférieures... et $f'(t)$ sera nul s'il y a autant de valeurs x_i supérieures à t que de valeurs x_i inférieures. D'où le minimum est atteint pour $a = m_e$.

Le programme de 1ère S prévoit la notion de quartile. Le GEPS propose la définition suivante pour une notion plus générale de **quantile** :

> En statistique, pour toute série numérique de données à valeurs dans un intervalle I, on définit la fonction quantile Q, de [0,1] dans I, par : $Q(u) = \inf\{x, F(x) \geq u\}$, où $F(x)$ désigne la fréquence des éléments de la série inférieurs ou égaux à x.

Soit n la taille de la série ; si on ordonne la série par ordre croissant, $Q(u)$ est la valeur du terme de cette série dont l'indice est le plus petit entier supérieur ou égal à u.

Dans le cadre de cette définition, les trois quartiles sont $Q_1 = Q(0,25)$, $Q_2 = Q(0,50)$ et $Q_3 = Q(0,75)$. Les 9 déciles sont les valeurs de $Q(i/10)$, $i = 1...9$, les 99 centiles sont les valeurs de $Q(i/100)$, $i = 1...99$. On définit assez souvent la médiane m_e par $m_e = Q(0,5)$: la médiane est alors le second quartile, le cinquième décile, le cinquantième centile, etc....

Voir le document du GEPS... (PDF, 58 Ko)

2. Les caractéristiques de dispersion

2.1. L'étendue

> L'étendue de la série $(x_1, x_2, ..., x_n)$ est égale à : $\max(x_i) - \min(x_i)$.

Comme la moyenne des valeurs extrêmes, elle est très sensible à ces valeurs extrêmes.

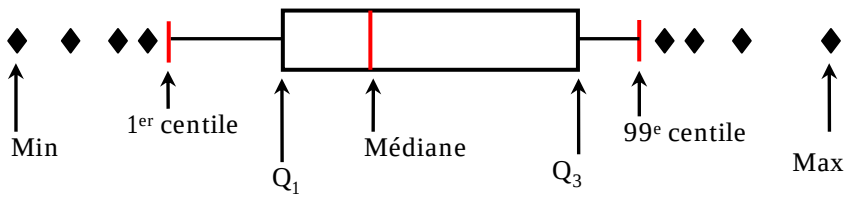
2.2. L'écart interquartile

> **L'écart interquartile** est la quantité $Q_3 - Q_1$.

2.3. Une autre représentation : « la boîte à moustaches ».

Elle est due à JW. Tukey et est appelée « box plot » en anglais.

Le dessin suffit à l'explication :



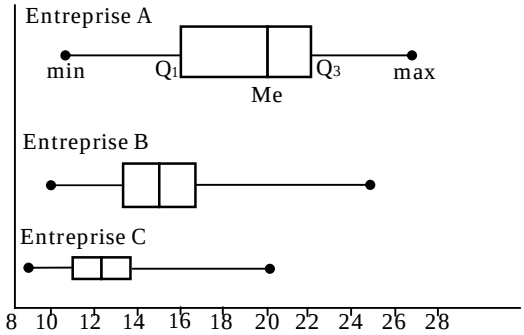
Pour comparer des populations qui n'ont pas le même effectif, on trace la largeur du rectangle **proportionnelle à la racine carrée de la population.**

Exercice

Comparer les salaires dans les trois entreprises suivantes d'un même secteur industriel.

Entreprise	Taille	min	Q1	Me	Q3	max
A	125	10 500	16 000	20 000	22 000	27 000
B	75	10 000	13 500	15 000	17 000	25 000
C	25	8 500	11 000	12 500	14 000	20 500

À partir des données, on obtient la représentation suivante :



d'après Itinéraires en Statistiques et Probabilités (Ellipses)

Dans les premiers diagrammes de Tukey, la longueur des « moustaches » est 1,5 fois l'écart interquartile. Les diagrammes de Tukey étaient utilisés dans des secteurs où les données peuvent le plus souvent être modélisées en utilisant une loi de Gauss ; dans ce cas, au niveau théorique, les extrémités des « moustaches » sont voisines du premier et 99e centile : ces diagrammes étaient surtout utilisés pour détecter la présence de données exceptionnelles. On utilise aujourd'hui les diagrammes en boites pour représenter des distributions empiriques de données quelconques, non nécessairement symétriques autour de la moyenne, et le choix de moustaches de longueurs 1,5 fois l'écart interquartile ne se justifie plus. (*Document d'accompagnement des programmes de 1re S*)

2.4. Variance et écart type

On a déjà rencontré la variance dans l'interprétation géométrique de la moyenne.

Pour une série $(x_1, x_2, \dots, x_n)$ de moyenne m, on définit la variance $V(X)$ et l'écart-type s par : $V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ et $s = \sqrt{V(X)}$

Propriétés

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - m^2$$

Pour tous a et b réels :

$$V(aX+b) = a^2\, V(X)$$

$$s(aX+b) = |a|\, s(X)$$

Exercice

On considère deux séries statistiques portant sur le même caractère :

- $(x_1, n_1), \ldots, (x_p, n_p)$, effectif total n, moyenne $\bar{x}$, écart-type σ_x ;
- $(y_1, m_1), \ldots, (y_q, m_q)$, effectif total m, moyenne $\bar{y}$, écart-type σ_y.

On note (z_k, r_k) la série statistique obtenue en regroupant les deux séries, $\bar{z}$ sa moyenne et σ_z son écart-type.

1. Montrer que $\bar{z} = \dfrac{n\bar{x} + m\bar{y}}{n+m}$.
2. Démontrer que : $(m+n)\,\sigma_z{}^2 = n\left(\sigma_x{}^2 + (\bar{x}-\bar{z})^2\right) + m\left(\sigma_y{}^2 + (\bar{y}-\bar{z})^2\right)$

   En déduire : $\sigma_z{}^2 = \dfrac{n\,\sigma_x{}^2 + m\,\sigma_y{}^2}{n+m} + \dfrac{n\,m}{(n+m)^2}(\bar{x}-\bar{y})^2$
3. Un professeur a corrigé n copies d'examen. La moyenne des notes est $\bar{x}$ et l'écart-type de la série de notes est σ.
   Une copie supplémentaire (à corriger) lui est attribuée. On désigne par y la note obtenue pour cette copie.
   Exprimer en fonction des données la moyenne et l'écart-type de la série de $(n+1)$ notes ainsi obtenue. Existe-t-il une valeur de y qui ne modifie pas la moyenne ? l'écart-type ? les deux ?

Dans le cas d'une variable continue et d'un regroupement par classes, on obtient une **valeur approchée** de la variance à l'aide de la modélisation utilisée pour obtenir une valeur approchée de la moyenne, c'est à dire en ramenant toutes les valeurs d'une classe au centre de cette classe. Rappelons que pour avoir une valeur approchée de la médiane, on avait utilisé la modélisation de la répartition uniforme par classe.

2.5. À propos du regroupement en classes

2.5.1. Comparaison des variances

Soit $(x_1, x_2, \ldots, x_n)$ une série répartie en m classes $[a_1;a_2[\,;[a_2;a_3[\ldots[a_m;a_{m+1}[$ d'effectifs respectifs n_i pour $1 \le i \le m$ (et $\sum_{i=1}^{m} n_i = n$).

On note $\bar{x}$ la moyenne obtenue par l'un ou l'autre des modèles choisis.

On note σ l'écart-type réel de la série, σ_1 l'écart-type obtenu en ramenant les valeurs au centre de chaque classe, et σ_2 l'écart-type obtenu en supposant une répartition uniforme à l'intérieur de chaque classe.

La variance σ^2 est systématiquement sous estimée par $\sigma_1{}^2$, car on néglige la variation à l'intérieur de chaque classe ($\sigma_1 < \sigma$).

En revanche, $\sigma_2{}^2$ est en général assez proche de σ^2.

Comparons σ_1 et σ_2 :

$$n\,\sigma_2{}^2 = \sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2$$

$$n\,\sigma_1{}^2 = \sum_{i=1}^{m} n_i\left(c_i - \bar{x}\right)^2 \quad \text{où} \quad c_i = \frac{a_i + a_{i+1}}{2}$$

$n\sigma_2{}^2 = \sum_{i=1}^{m}\left(\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2\right)$ en regroupant les valeurs par classe.

D'après ce qui précède, pour la ième classe, on a :

$x_k = a_i + \frac{k\,d_i}{n_i+1}$ où $d_i = a_{i+1} - a_i$.

$$\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2 = \sum_{k=1}^{n_i}\left(x_k - c_i + c_i - \bar{x}\right)^2 =$$

$$= n_i(c_i - \bar{x})^2 + \sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 + 2(c_i - \bar{x})\sum_{k=1}^{n_i}\left(x_k - c_i\right).$$

Or $n_i c_i = \sum_{k=1}^{n_i} x_k$ donc le dernier terme est nul.

$$\sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 = \sum_{k=1}^{n_i} d_i{}^2\left(\frac{k}{n_i+1} - \frac{1}{2}\right)^2$$

$$= \frac{d_i{}^2}{(n_i+1)^2}\sum_{k=1}^{n_i}k^2 + \frac{1}{4}n_i\,d_i{}^2 - \frac{d_i{}^2}{n_i+1}\sum_{k=1}^{n_i}k$$

$$= \frac{d_i{}^2}{(n_i+1)^2}\times\frac{n_i(n_i+1)(2n_i+1)}{6} + \frac{1}{4}n_i\,d_i{}^2 - \frac{d_i{}^2}{n_i+1}\times\frac{n_i(n_i+1)}{2}$$

$$= \frac{n_i\,d_i{}^2(2n_i+1)}{6(n_i+1)} - \frac{1}{4}n_i\,d_i{}^2$$

$$= \frac{n_i\,d_i{}^2}{12}\left(\frac{n_i-1}{n_i+1}\right)$$

Finalement, $n\,\sigma_2{}^2 = \sum_{i=1}^{m} n_i(c_i - \bar{x})^2 + \sum_{i=1}^{m}\frac{n_i\,d_i{}^2}{12}\left(\frac{n_i-1}{n_i+1}\right)$

Soit : $\sigma_2{}^2 = \sigma_1{}^2 + \sum_{i=1}^{m}\frac{n_i\,d_i{}^2}{12n}\left(\frac{n_i-1}{n_i+1}\right)$

D'après ce qui précède, pour la ième classe, on a :

$x_k = a_i + \frac{k\,d_i}{n_i+1}$ où $d_i = a_{i+1} - a_i$.

$$\sum_{k=1}^{n_i}\left(x_k - \bar{x}\right)^2 = \sum_{k=1}^{n_i}\left(x_k - c_i + c_i - \bar{x}\right)^2 =$$

$$= n_i(c_i - \bar{x})^2 + \sum_{k=1}^{n_i}\left(x_k - c_i\right)^2 + 2(c_i - \bar{x})\sum_{k=1}^{n_i}\left(x_k - c_i\right).$$

2.5.2. Regroupement en classe et précision demandée dans les exercices donnés

Lorsque l'on propose, en temps limité, un exercice de statistique, on est conduit à limiter le nombre de données à traiter (pour diminuer les problèmes de saisie et de calcul). Pour cela on regroupe fréquemment les données en un petit nombre de classes. Il est important de veiller à ce que ce regroupement (qui constitue toujours une perte d'information) soit bien compatible avec la précision demandée par la suite et que les réponses aux questions posées puissent être trouvées sans ambiguïté.

Document de J.P. POUGET IA-IPR, académie de Créteil (PDF, 198 Ko)

2.6. Inégalité de Bienaymé-Tchebychev

Soit une série statistique de moyenne m et d'écart type s. Pour tout α réel strictement positif, on note f_α la fréquence des valeurs comprises entre $m - \alpha s$ et $m + \alpha s$ (c'est-à-dire $|X - m| \leq \alpha s$).

Alors $f_\alpha > 1 - \frac{1}{\alpha^2}$

Cette inégalité, bien que médiocre, est valable quelle que soit la série statistique.

Ainsi plus de 75 % des valeurs sont dans $[\, m - 2s \,;\, m + 2s \,]$,

plus de 88 % des valeurs sont dans $[\, m - 3s \,;\, m + 3s \,]$,

plus de 93,75 % des valeurs sont dans $[\, m - 4s \,;\, m + 4s \,]$,

Chapitre 4
Lois discrètes

1. Loi de Bernoulli

Une variable aléatoire X est une variable de Bernoulli si elle ne prend que les valeurs 0 et 1 avec des probabilités non nulles.

$P(X = 1) = p$, $P(X = 0) = 1 - p = q$, avec $p \in]\,0\,;\,1\,[$.

Une variable aléatoire de Bernoulli illustre toute expérience aléatoire n'ayant que deux issues possibles et effectuée une seule fois.

Traditionnellement le « succès » correspond à la valeur 1 et l'« échec » à la valeur 0.

$E(X) = 1.p + 0.(1 - p) = p$

$V(X) = (0 - p)^2 (1 - p) + (1 - p)^2 p = p (1 - p)$

En résumé $E(X) = p$ et $V(X) = pq$.

2. Loi binomiale $\mathcal{B}(n\,;\,p)$

2.1. L'expérience de référence standard

Une urne contient deux catégories de boules : des blanches en proportion p et des noires en proportion $1 - p$. On effectue n tirages successifs d'une boule **avec remise**. On appelle X le nombre de boules blanches obtenues au cours de cette expérience.

2.2. Les résultats de base

Loi de X : $P(X = k) = C_n^k\, p^k (1-p)^{n-k}$, $k = 0;1;\ldots;n$.

Espérance, écart-type.

X peut être considérée comme la somme de n variables de Bernoulli X_i où $X_i = 1$ si la boule tirée au i-ème tirage est blanche et $X_i = 0$ sinon.

On a, pour tout $i \in \{1\,;\,2\,;\,\ldots\,;\,n\}$, $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p = q$.

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

Les variables aléatoires X_i peuvent être considérées comme indépendantes. Donc :

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \left(V(X_i)\right) = npq.$$

2.3. Fréquence binomiale

X suit la loi $\mathcal{B}(n\,;p)$.

La fréquence est la variable $F=\frac{X}{n}$ d'où $E(F)=\frac{1}{n}E(X)=p$ et $V(F)=\frac{1}{n^2}V(X)=\frac{pq}{n}$.

En résumé :

X suit la loi $\mathcal{B}(n\,;p)$.

$P(X=k)=C_n^k\,p^k(1-p)^{n-k}$, $k=0;1;\ldots;n$

$E(X)=np$ $\quad V(X)=npq$ $\quad \sigma(X)=\sqrt{npq}$

$F=\frac{X}{n}$ $\quad E(F)=p$ $\quad V(F)=\frac{pq}{n}$ $\quad \sigma(F)=\sqrt{\frac{pq}{n}}$

3. Loi Hypergéométrique $\mathcal{H}(N\,;\,n\,;\,p)$

3.1.Le modèle

Une urne contient deux catégories de boules : des blanches en proportion p et des noires en proportion $q = 1 - p$. Si N est le nombre de boules dans l'urne, il y a $N\,p$ boules blanches et $N(1 - p)$ boules noires.

On effectue n tirages successifs d'une boule **sans remise**.

On appelle X le nombre de boules blanches obtenues au cours de cette expérience.

On sait que ce type de tirage est équivalent à un tirage exhaustif de n boules.

Ceci est à rapprocher du prélèvement d'un échantillon de n boules dans l'urne.

3.2. La loi

$$\forall k\in[0\,;n],\ P(X=k)=\frac{C_{Np}^k\,C_{N(1-p)}^{n-k}}{C_N^k}\quad ;\quad E(X)=np\quad ;\quad V(X)=npq\times\frac{N-n}{N-1}$$

On peut remarquer que $\mathcal{H}(N\,;n\,;p)$ et $\mathcal{B}(n\,;p)$ ont même espérance mathématique. La variance ne diffère que du **coefficient d'exhaustivité** $\frac{N-n}{N-1}$.

3.3. Convergence en loi

On montre que, pour n et p fixés et pour X_N suivant $\mathcal{H}(N\,;n\,;p)$, X_N converge en loi vers une variable X qui suit $\mathcal{B}(n\,;p)$. C'est-à-dire $\lim\limits_{N\to+\infty} P(X_N=k)=P(X=k)$

L'intérêt est énorme : si N est grand et n petit par rapport à N, on peut remplacer la loi hypergéométrique (qui dépend de trois paramètres) par la loi binomiale qui ne dépend que de deux paramètres et pour laquelle il existe des tables.

En pratique, si $n < 0{,}1\,N$, on considère qu'un tirage exhaustif (tirage un à un sans remise) est équivalent à un tirage non exhaustif (tirage un à un avec remise).

4. Loi de Poisson

4.1.Le cadre d'intervention

C'est une « loi limite ». On verra que, sous certaines conditions, une loi de Poisson est limite d'une loi binomiale.

Dans la pratique une telle loi est utilisée pour approcher et décrire des phénomènes où les conditions d'application de la loi binomiale sont réunies (répétitions indépendantes d'une même épreuve dichotomique), où la probabilité du cas favorable est faible et où le nombre d'épreuves est grand.

4.2. La loi

La variable aléatoire X suit la loi de Poisson de paramètre λ ($\lambda>0$) lorsque :

- X a pour ensemble de valeurs $\mathbb{N}$
- $P(X=k)=e^{-\lambda}\dfrac{\lambda^k}{k!}$ $\quad (k=0\,;1\,;\ldots\,;n\,;\ldots)$

Alors : $E(X)=\sum_{k=0}^{+\infty} kP(X=k)=\lambda$ $\qquad V(X)=\sum_{k=0}^{+\infty}(k-\lambda)^2\,P(X=k)=\lambda$

4.3. Extraits des tables de la loi de Poisson fournis lors les épreuves de BTS

$$P(X=k)=e^{-\lambda}.\frac{\lambda^k}{k!} \quad ; \quad E(X)=V(X)=\lambda$$

k\λ	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659
2	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647
3	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494
4	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111
5	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020
6			0,0000	0,0000	0,0000	0,0001	0,0002	0,0003
7						0,0000	0,0000	0,0000
8								
9								
10								

k\λ	1	1,5	2	3	4	5	6	7	8	9	10
0	0,368	0,223	0,135	0,050	0,018	0,007	0,002	0,001	0,000	0,000	0,000
1	0,368	0,335	0,271	0,149	0,073	0,034	0,015	0,006	0,003	0,001	0,000
2	0,184	0,251	0,271	0,224	0,147	0,084	0,045	0,022	0,011	0,005	0,002
3	0,061	0,126	0,180	0,224	0,195	0,140	0,089	0,052	0,029	0,015	0,008
4	0,015	0,047	0,090	0,168	0,195	0,175	0,134	0,091	0,057	0,034	0,019
5	0,003	0,014	0,036	0,101	0,156	0,175	0,161	0,128	0,092	0,061	0,038
6	0,001	0,004	0,012	0,050	0,104	0,146	0,161	0,149	0,122	0,091	0,063
7	0,000	0,001	0,003	0,022	0,060	0,104	0,138	0,149	0,140	0,117	0,090
8	0,000	0,000	0,001	0,008	0,030	0,065	0,103	0,130	0,140	0,132	0,113
9		0,000	0,000	0,003	0,013	0,036	0,069	0,101	0,124	0,132	0,125
10			0,000	0,001	0,005	0,018	0,041	0,071	0,099	0,119	0,125
11				0,000	0,002	0,008	0,023	0,045	0,072	0,097	0,114
12				0,000	0,001	0,003	0,011	0,026	0,048	0,073	0,095
13					0,000	0,001	0,005	0,014	0,030	0,050	0,073
14					0,000	0,000	0,002	0,007	0,017	0,032	0,052
15						0,000	0,001	0,003	0,009	0,019	0,035
16							0,000	0,001	0,005	0,011	0,022
17								0,001	0,002	0,006	0,013
18								0,000	0,001	0,003	0,007
19									0,000	0,001	0,004
20										0,001	0,002
21										0,000	0,001
22											0,000

4.4. Quelques exemples

Exercice 1

Dans une certaine usine il se produit, en moyenne, cinq accidents par an. On suppose que le nombre d'accidents suit une loi de Poisson.

Calculer la probabilité pour qu'il ne dépasse pas sept.

Quelle est la probabilité d'avoir une année sans accident ?

Solution

La variable aléatoire X égale au nombre d'accidents suit une loi de Poisson de paramètre $\lambda = 5$.

$$P(X \leq 7) = e^{-5} \sum_{k=0}^{k=7} \frac{5^k}{k!} \simeq 0{,}87$$

$$P(X = 0) = e^{-5} \simeq 6{,}7 \times 10^{-3}$$

Exercice 2

Pour une femme ayant eu entre 50 et 52 ans en l'an 2000, le nombre d'enfants, noté X, suit une loi de Poisson de paramètre inconnu λ. Un échantillon de 1 000 de ces femmes donne 135 femmes sans enfant.

1. Donner une estimation de λ.
2. Estimer la proportion de ces femmes ayant plus de trois enfants.

Solution

Si on admet que l'échantillon est représentatif de la population, on a $P(X = 0) = e^{-\lambda} \approx 0{,}135$.

Donc $\lambda \approx -\ln(0{,}135) \approx 2$.

$$P(X > 3) = 1 - P(X \leq 3) = 1 - e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{6}\right) \simeq 0{,}145\,.$$

Parmi les femmes qui ont eu entre 50 et 52 ans en l'an 2000, il y en a donc environ 145 sur 1000 qui ont plus de trois enfants.

4.5. Loi binomiale et loi de Poisson

Soit (X_n) une suite de variables aléatoires suivant la loi $\mathcal{B}(n\,;\,p_n)$ avec $\lim\limits_{n \to +\infty} np_n = \lambda$.

Alors (X_n) converge en loi vers une variable de Poisson $\mathcal{P}(\lambda)$.

En pratique :
Soit X une variable aléatoire binomiale de paramètres n et p.
Si $n \geq 30$, $p \leq 0{,}1$ et $np < 15$ alors X suit approximativement la loi de Poisson de paramètre np.

Exemple 1

Une usine fabrique des CD ROM en quantité importante. Une étude statistique a montré que 2 % de ces CD étaient défectueux. Pour effectuer un contrôle de fabrication, on prélève au hasard 150 CD. On note X le nombre de CD défectueux dans cet échantillon.

1. Quelle est la loi de probabilité suivie par la variable aléatoire X ? En préciser les paramètres.
2. Par quelle loi de probabilité peut-on approcher la loi de X ? Calculer P(X > 3)

Solution

1. Soit N est le nombre total de CD produits. L'étude statistique montre que 2% des CD sont défectueux. On prélève un échantillon de 150 CD. Le nombre de CD défectueux dans l'échantillon suit donc la loi Hypergéométrique $\mathcal{H}(N\,;\,150\,;\,0{,}02)$.

Mais d'une part N est grand et d'autre part on peut considérer que 150 est petit par rapport à N (< 0,1 N). Dans ces conditions on peut assimiler le prélèvement des 150 CD à un prélèvement un à un avec remise et donc considérer que les 150 CD sont prélevés indépendamment les un des autres.

X suit donc, à peu de choses près, la loi binomiale de paramètre $n = 150$ et $p = 0{,}02$.

2. n est grand ($n \geq 30$), p est faible ($p \leq 0{,}1$) et $np = 3$ est inférieur à 15. On peut donc utiliser la loi de Poisson de paramètre 3 comme approximation.

$$P(X>3) \simeq 1-P(X=0)-P(X=1)-P(X=2)-P(X=3)$$
$$\simeq 1-\mathrm{e}^{-3}\left(1+3+\frac{9}{2}+\frac{27}{6}\right) \simeq 0{,}35$$

Exemple 2

On suppose qu'une urne contient 1 boule blanche et 99 boules noires.

On effectue n tirages successifs d'une boule avec remise.

Déterminer n pour que la probabilité de tirer au moins une fois la boule blanche soit supérieure ou égale à 0,95.

Solution

Soit X la v.a. égale au nombre de fois où on tire la boule blanche au cours de n tirages. X suit B(n ; 0,01). $P(X \geq 1) = 1 - P(X = 0) = 1 - 0{,}99^n$

Si on veut que P(X ≥ 1) ≥ 0,95, il faut $(0{,}99)^n \leq 0{,}05$, soit $n \geq \dfrac{\ln(0{,}05)}{\ln(0{,}99)}$

i.e. $n \geq 298{,}1$ et donc $n \geq 299$.

Il faut donc effectuer 299 tirages au moins pour être sûr, à 95 %, d'avoir au moins une boule blanche.

Calcul approché

n est grand, p est faible. On essaye d'approcher X par une variable de Poisson de paramètre $\dfrac{n}{100}$. $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\frac{n}{100}}$.

Donc, pour avoir P(X ≥ 1) ≥ 0,95, il faut $e^{-\frac{n}{100}} \leq 0{,}05$, soit $n \geq -100.\ln(0{,}05)$

Donc $n \geq 299{,}6$ et par suite $n \geq 300$

4.6. Processus de Poisson

Soit T une période de temps que l'on subdivise en n intervalles d'égale amplitude Δt. On a donc $T = n\,\Delta t$.

¤ Si, à l'intérieur de chacun de ces intervalles, la probabilité qu'un événement A se produise est constante et égale à p,

¤ Si, de plus, on admet que l'événement A ne peut se produire qu'au plus une fois à l'intérieur de chaque intervalle,

on dit alors que la réalisation de l'événement A est un **processus de Poisson**.

Exercice

Un standard téléphonique reçoit, en moyenne, 2 appels par minute. Les appels sont répartis au hasard dans le temps.

1. Expliquer pourquoi le fait de recevoir un appel téléphonique peut être considéré comme un processus de Poisson. Préciser le paramètre de cette loi.

2. Quelle est la loi de probabilité régissant le nombre d'appels reçus en 4 minutes ?

Calculer la probabilité pour que ce nombre d'appel dépasse 10.

Solution

1. On peut fractionner la minute T en intervalles d'une seconde Δt. Alors $n = 60$ et $\Delta t = 1$.

On admet alors que, chaque seconde, la probabilité de recevoir un appel est constante : $p = \frac{2}{60} = \frac{1}{30}$.

On admet aussi que, chaque seconde, il ne peut se produire au plus qu'un appel et que, d'une seconde sur l'autre, les appels sont indépendants.

Le fait de recevoir un appel est alors un processus de Poisson.

Si X désigne le nombre d'appels reçus en une minute, on peut considérer qu'à chaque seconde :

¤ le standard reçoit un appel avec la probabilité $p = \frac{1}{30}$ et qu'il n'en reçoit pas avec la probabilité $q = \frac{29}{30}$

¤ X suit la loi binomiale $\mathcal{B}\left(60\,;\frac{1}{30}\right)$ d'espérance mathématique 2. Cette loi peut être approchée par une loi de Poisson de paramètre 2.

2. Sur une période de quatre minutes, le partage en 240 secondes conduit à la loi binomiale $\mathcal{B}\left(240\,;\frac{1}{30}\right)$ d'espérance mathématique 8, ce qui permet encore une approximation par la loi de Poisson de paramètre 8.

Si Y est la variable aléatoire qui suit la loi de Poisson de paramètre 8, on lit dans la table que : $P(Y \geq 10) = 1 - P(Y \leq 9 \approx 0{,}283\,38$

NB : Si on utilise directement la loi binomiale $\mathcal{B}\left(240\,;\frac{1}{30}\right)$, on obtient 0,281 24

Chapitre 5
Lois continues

1. Rappel

Voir page 11 pour les rappels sur les variables à densité.

RAPPEL

Une variable X est absolument continue s'il existe une fonction f définie sur $\mathbb{R}$ telle que :

- f est positive sur $\mathbb{R}$,
- f est continue sur $\mathbb{R}$ sauf peut-être en un nombre fini de points où elle admet une limite à droite et une limite à gauche,
- $\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t = 1$,
- La fonction de répartition F de X est liée à f par : $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$.

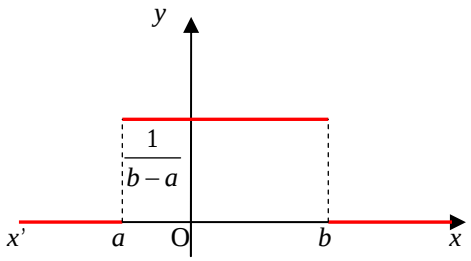
On dit que f est **une densité** de X. Abusivement, f est appelée **loi de X.**

$$\mathrm{P}([a\,;b]) = \int_a^b f(t)\,\mathrm{d}t \qquad E(X) = \mu = \int_{-\infty}^{+\infty} t\ f(t)\,\mathrm{d}t \qquad V(X) = \int_{-\infty}^{+\infty} (t-\mu)^2\ f(t)\,\mathrm{d}t$$

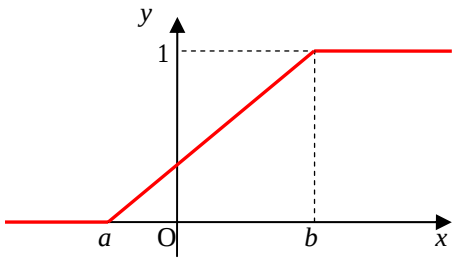
2. Loi uniforme

Une v. a. X suit la loi continue uniforme sur [a ; b] ($a \neq b$) si, et seulement si, X a pour densité de probabilité la fonction f définie par $\begin{cases} \forall x \in [\,a\,;b\,], \quad f(x) = \dfrac{1}{b-a} \\ \forall x \in \mathbb{R} - [\,a\,;b\,], \quad f(x) = 0 \end{cases}$.

Cette loi est notée $\mathcal{U}([\,a\,;b\,])$.



Densité de probabilité



Fonction de répartition

$$\mathrm{E(X)} = \frac{a+b}{2} \qquad \mathrm{V(X)} = \frac{(b-a)^2}{12}$$

Soit X et Y deux v.a. telles que Y = $(b-a)$ X + a

X suit la loi uniforme sur [0 ; 1] si et seulement si Y suit la loi uniforme sur [a ; b].

3. Loi normale ou loi de Laplace-Gauss

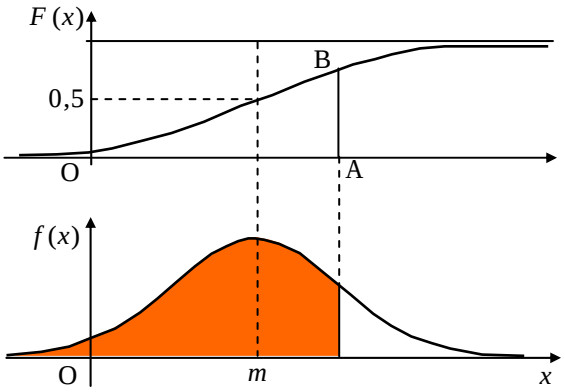
3.1. Définition et premières propriétés

Soit deux réels m et σ avec $\sigma > 0$.

Une variable aléatoire X suit une loi normale de paramètres m et σ notée $\mathcal{N}(m\ ;\sigma)$ si et seulement si X a pour densité la fonction f définie par : $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$. On a alors E(X) = m et V(X) = σ.

La courbe représentative de f est appelée « courbe en cloche ».

La figure suivante représente f et sa fonction de répartition F. La longueur du segment [AB] est égale à l'aire du domaine grisé.



3.2. Loi normale centrée réduite

Une variable aléatoire T suit la loi normale centrée réduite si elle suit la loi normale $\mathcal{N}(0\ ;1)$. Sa fonction densité de probabilité, f, est alors définie par : $f(t)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}$.

Son espérance mathématique est E(T) = 0 et sa variance V(T) = 1.

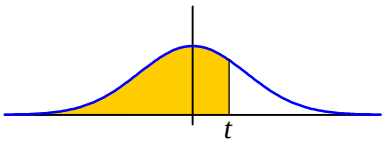
Si on note Π sa fonction de répartition, on a : $\Pi(t)=P(T<t)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,\mathrm{d}x$

Une variable aléatoire X suit la loi normale $\mathcal{N}(m\ ;\sigma)$ si et seulement si $T=\frac{X-m}{\sigma}$ suit la loi normale $\mathcal{N}(0\ ;1)$.

3.3. Les tables de la loi normale centrée réduite

Extraits de la table de la fonction intégrale de la loi normale centrée réduite $\mathcal{N}$ (0 , 1).

$$\Pi(t) = P(X \leq t) = \int_{-\infty}^{t} f(x)\ dx$$



t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500 0	0.504 0	0.508 0	0.512 0	0.516 0	0.519 9	0.523 9	0.527 9	0.531 9	0.535 9
0.1	0.539 8	0.543 8	0.547 8	0.551 7	0.555 7	0.559 6	0.563 6	0.567 5	0.571 4	0.575 3
0.2	0.579 3	0.583 2	0.587 1	0.591 0	0.594 8	0.598 7	0.602 6	0.606 4	0.610 3	0.614 1
0.3	0.617 9	0.621 7	0.625 5	0.629 3	0.633 1	0.636 8	0.640 6	0.644 3	0.648 0	0.651 7
0.4	0.655 4	0.659 1	0.662 8	0.666 4	0.670 0	0.673 6	0.677 2	0.680 8	0.684 4	0.687 9
0.5	0.691 5	0.695 0	0.698 5	0.701 9	0.705 4	0.708 8	0.712 3	0.715 7	0.719 0	0.722 4
0.6	0.725 7	0.729 1	0.732 4	0.735 7	0.738 9	0.742 2	0.745 4	0.748 6	0.751 7	0.754 9
0.7	0.758 0	0.761 1	0.764 2	0.767 3	0.770 4	0.773 4	0.776 4	0.779 4	0.782 3	0.785 2
0.8	0.788 1	0.791 0	0.793 9	0.796 7	0.799 5	0.802 3	0.805 1	0.807 8	0.810 6	0.813 3
0.9	0.815 9	0.818 6	0.821 2	0.823 8	0.826 4	0.828 9	0.831 5	0.834 0	0.836 5	0.838 9
1.0	0.841 3	0.843 8	0.846 1	0.848 5	0.850 8	0.853 1	0.855 4	0.857 7	0.859 9	0.862 1
1.1	0.864 3	0.866 5	0.868 6	0.870 8	0.872 9	0.874 9	0.877 0	0.879 0	0.881 0	0.883 0
1.2	0.884 9	0.886 9	0.888 8	0.890 7	0.892 5	0.894 4	0.896 2	0.898 0	0.899 7	0.901 5
1.3	0.903 2	0.904 9	0.906 6	0.908 2	0.909 9	0.911 5	0.913 1	0.914 7	0.916 2	0.917 7
1.4	0.919 2	0.920 7	0.922 2	0.923 6	0.925 1	0.926 5	0.927 9	0.929 2	0.930 6	0.931 9
1.5	0.933 2	0.934 5	0.935 7	0.937 0	0.938 2	0.939 4	0.940 6	0.941 8	0.942 9	0.944 1
1.6	0.945 2	0.946 3	0.947 4	0.948 4	0.949 5	0.950 5	0.951 5	0.952 5	0.953 5	0.954 5
1.7	0.955 4	0.956 4	0.957 3	0.958 2	0.959 1	0.959 9	0.960 8	0.961 6	0.962 5	0.963 3
1.8	0.964 1	0.964 9	0.965 6	0.966 4	0.967 1	0.967 8	0.968 6	0.969 3	0.969 9	0.970 6
1.9	0.971 3	0.971 9	0.972 6	0.973 2	0.973 8	0.974 4	0.975 0	0.975 6	0.976 1	0.976 7
2.0	0.977 2	0.977 8	0.978 3	0.978 8	0.979 3	0.979 8	0.980 3	0.980 8	0.981 2	0.981 7
2.1	0.982 1	0.982 6	0.983 0	0.983 4	0.983 8	0.984 2	0.984 6	0.985 0	0.985 4	0.985 7
2.2	0.986 1	0.986 4	0.986 8	0.987 1	0.987 5	0.987 8	0.988 1	0.988 4	0.988 7	0.989 0
2.3	0.989 3	0.989 6	0.989 8	0.990 1	0.990 4	0.990 6	0.990 9	0.991 1	0.991 3	0.991 6
2.4	0.991 8	0.992 0	0.992 2	0.992 5	0.992 7	0.992 9	0.993 1	0.993 2	0.993 4	0.993 6
2.5	0.993 8	0.994 0	0.994 1	0.994 3	0.994 5	0.994 6	0.994 8	0.994 9	0.995 1	0.995 2
2.6	0.995 3	0.995 5	0.995 6	0.995 7	0.995 9	0.996 0	0.996 1	0.996 2	0.996 3	0.996 4
2.7	0.996 5	0.996 6	0.996 7	0.996 8	0.996 9	0.997 0	0.997 1	0.997 2	0.997 3	0.997 4
2.8	0.997 4	0.997 5	0.997 6	0.997 7	0.997 7	0.997 8	0.997 9	0.997 9	0.998 0	0.998 1
2.9	0.998 1	0.998 2	0.998 2	0.998 3	0.998 4	0.998 4	0.998 5	0.998 5	0.998 6	0.998 6

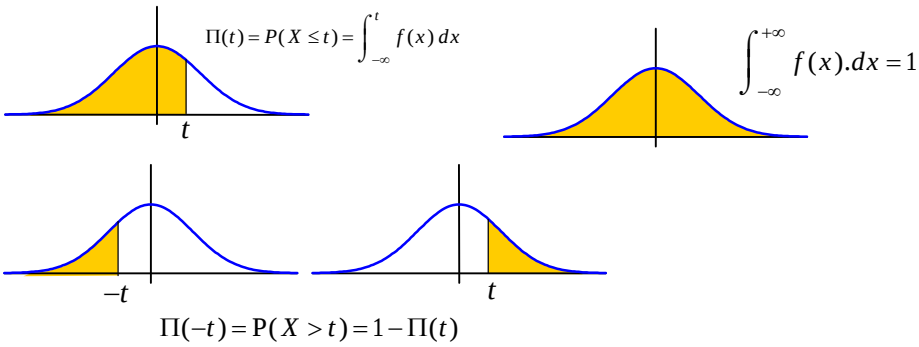
Table pour les grandes valeurs de *t* :

t	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.8	4.0	4.5
$\Pi(t)$	0.998 65	0.999 03	0.999 31	0.999 52	0.999 66	0.999 77	0.999 841	0.999 928	0.999 968	0.999 997

Ces tables sont construites uniquement pour t positif, mais les deux propriétés graphiques suivantes :

- la courbe est symétrique par rapport à l'axe des ordonnées
- l'aire de la surface comprise entre la courbe et l'axe des abscisses est égale à 1

permettent d'effectuer les calculs dans tous les cas.



$$\Pi(-t) = \mathrm{P}(X > t) = 1 - \Pi(t)$$

Quelques résultats importants :

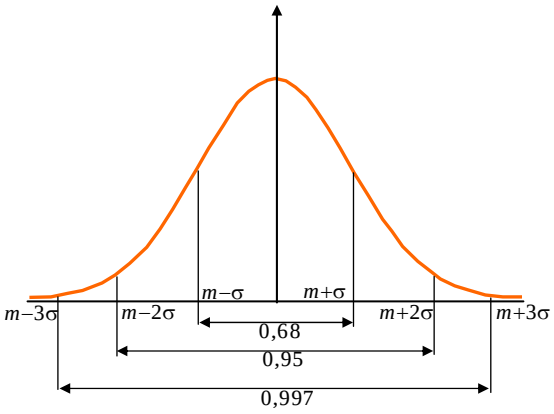
$\forall t_1 \in \mathbb{R}, \forall t_2 \in \mathbb{R},$ avec $t_1 < t_2$,	$\mathrm{P}(t_1 \le \mathrm{T} \le t_2) = \Pi(t_2) - \Pi(t_1)$
$\forall t \in \mathbb{R}$,	$\mathrm{P}(\mathrm{T} > t) = 1 - \Pi(t)$
$\forall t \in [\,0\,;+\infty\,[$	$\mathrm{P}(T \le -t) = 1 - \Pi(t)$
$\forall t \in [\,0\,;+\infty\,]$	$\mathrm{P}(-t \le T \le t) = 2\Pi(t) - 1$

Cette dernière relation est souvent utilisée.

Il convient de retenir deux valeurs importantes :

$$\mathrm{P}(-1{,}96 \le T \le 1{,}96) = 0{,}95$$
$$\mathrm{P}(-2{,}58 \le T \le 2{,}58) = 0{,}99$$

Plus généralement on a :



3.4. Quelques exercices d'application

Exercice 1

Une usine fabrique en grande série un certain type de pièces cylindriques. On appelle X la v.a. qui, à chaque pièce tirée au hasard, associe sa longueur et Y la v.a. qui associe son diamètre.

On suppose que X et Y sont indépendantes et suivent des lois normales de moyennes respectives $\bar{x}$ = 8,55 cm et $\bar{y}$ = 5,20 cm et d'écarts types respectifs σ_x = 0,05 cm et σ_y = 0,05 cm.

1. Déterminer, à 10^{-3} près les probabilités P(8,45 < X < 8,70) et P(5,07 < Y < 5,33.

2. Une pièce est conforme si : 8,45 < X < 8,70 et 5,07 < Y < 5,33.

   a. Calculer le pourcentage de pièces non conformes à la sortie de la chaîne.

   b. Les machines nécessitent-elles un réglage, sachant que le pourcentage de pièces non conformes ne peut dépasser 1 % ?

Solution

X suit la loi normale $\mathcal{N}(8{,}55\ ;\ 0{,}05)$. Donc la v.a. T, définie par $T = \dfrac{X - 8{,}55}{0{,}05}$, suit la loi normale centrée réduite.

$8{,}45 \leq X \leq 8{,}70$ équivaut à $\dfrac{8{,}45 - 8{,}55}{0{,}05} \leq T \leq \dfrac{8{,}70 - 8{,}55}{0{,}05}$, donc à $-2 \leq T \leq 3$.

La probabilité cherchée est donc $\Pi(3) - \Pi(-2)$. Or $\Pi(-2) = 1 - \Pi(2)$. Donc :

$P(8{,}45 \leq X \leq 8{,}70) = \Pi(2) + \Pi(3) - 1 = 0{,}976$

De même, pour calculer $P(5{,}07 \leq Y \leq 5{,}33)$, on utilise la variable normale centrée réduite $T' = \dfrac{Y - 5{,}20}{0{,}05}$.

$5{,}07 \leq Y \leq 5{,}33$ équivaut à $\dfrac{5{,}07 - 5{,}20}{0{,}05} \leq T' \leq \dfrac{5{,}33 - 5{,}20}{0{,}05}$, donc à $-2{,}6 \leq T' \leq 2{,}6$.

$P(5{,}07 \leq Y \leq 5{,}33) = 2\,\Pi(2{,}6) - 1 = 0{,}991$

2. a. Soit D l'événement « la pièce n'est pas conforme ».

$D = (8{,}45 \leq X \leq 8{,}70)$ **et** $(5{,}07 \leq Y \leq 5{,}33)$

Or les v.a. X et Y sont indépendantes.

Donc $P(D) = P(8{,}45 \leq X \leq 8{,}70) \times P(5{,}07 \leq Y \leq 5{,}33) = 0{,}967$

Par suite la probabilité qu'une pièce ne soit pas conforme est $1 - 0{,}967 = 0{,}033$

Cette probabilité est supérieure à 1 %.

Il faut régler les machines...

Exercice 2

Une entreprise spécialisée dans la production de matériel optique fabrique des lentilles en grande série. On a mesuré la vergence x, exprimée en dioptries, de 1 000 lentilles du même type et on a obtenu la série statistique des mesures x_i suivantes avec les effectifs n_i correspondants.

x_i	1,975	1,980	1,985	1,990	1,995	2,000
n_i	8	27	67	118	176	200
x_i	2,005	2,010	2,015	2,020	2,025	
n_i	180	122	64	28	10	

1. Donner la moyenne $\bar{x}$ ainsi que l'écart type σ de cette série.
Représenter cette série à l'aide d'un diagramme en bâtons.

2. À chaque lentille de la production, on associe sa vergence x, exprimée en dioptries. On définit ainsi une v.a. X. L'allure du diagramme précédent amène à considérer que X suit la loi normale $\mathcal{N}(2\ ;\ 0{,}01)$. Une lentille est déclarée acceptable lorsque $1{,}98 < x < 2{,}02$. Elle est déclarée défectueuse dans le cas contraire.

Calculer la probabilité pour qu'une lentille de la production soit défectueuse.

3. Un réglage de machine permet de modifier l'écart type sans changer la moyenne. Dans cette question X suit donc une loi normale N(2 ; σ')

Déterminer σ' pour que la probabilité d'obtenir d'obtenir une lentille défectueuse soit inférieure ou égale à 0,01.

Solution

1. Les résultats arrondis au centième sont $\bar{x} = 2{,}00$ et $\sigma = 0{,}01$

2. On admet que X suit la loi normale N(2 ; 0,01). Une lentille est déclarée acceptable si l'événement $1{,}98 \leq X \leq 2{,}02$ est réalisé.

La v.a. T définie par $T = \dfrac{X-2}{0{,}01}$ suit la loi normale centrée réduite.

$1{,}98 \leq X \leq 2{,}02$ équivaut à $-2 \leq T \leq 2$. On sait que $P(-2 \leq T \leq 2) = 2\,\Pi(2) - 1$.

La table donne $\Pi(2) = 0{,}977\,2$. Donc $P(1{,}98 \leq X \leq 2{,}02) = 0{,}954\,4$

La probabilité pour qu'une lentille soit défectueuse est $1 - 0{,}954\,4$ soit, au centième le plus proche, 0,05.

3. Si X suit la loi normale N(2 ; σ') alors la variable T définie par $T = \dfrac{X-2}{\sigma'}$ suit la loi normale centrée réduite.

On veut que la probabilité d'obtenir une lentille défectueuse soit inférieure ou égale à 0,01, donc que la probabilité d'obtenir une lentille acceptable soit strictement supérieure à 0,99 : $P(1{,}98 \leq X \leq 2{,}02) > 0{,}99$.

Cette inégalité s'écrit :

$$P(|X-2| \leq 0{,}02) > 0{,}99, \text{ soit } P\left(|T| \leq \frac{0{,}02}{\sigma'}\right) > 0{,}99$$

On en déduit que $2\Pi\left(\dfrac{0{,}02}{\sigma'}\right) - 1 > 0{,}99$ soit $\Pi\left(\dfrac{0{,}02}{\sigma'}\right) > 0{,}995$.

La table de la loi normale centrée réduite donne $\Pi(2{,}575) = 0{,}995$.

Alors $\dfrac{0{,}02}{\sigma'} = 2{,}575$, soit $\sigma' = 0{,}00776\ldots$ Au millième le plus proche, $\sigma' = 0{,}008$.

d'après BTS Génie Optique

Chapitre 6
Théorèmes de convergence

I. La convergence en loi

On a déjà rencontré une convergence en loi lors de l'approximation d'une loi binomiale par une loi de Poisson. Ce problème se place dans un cadre plus général où on souhaite remplacer la loi d'une variable aléatoire par une loi d'usage plus simple

I.1. Définition

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{P}(\Omega), p)$, F_n leurs fonctions de répartition, et X une variable aléatoire définie sur ce même espace, de fonction de répartition F.

On dit que la suite (X_n) converge en loi vers X si, et seulement si, en tout point x où F est continue, on a $\lim\limits_{n \to +\infty} F_n(x) = F(x)$

> RAPPEL
> Soit (X_n) une suite de variables aléatoires binomiales de paramètres n et p_n telle que $\lim\limits_{n \to +\infty} n\, p_n = \lambda$.
> Alors (X_n) converge en loi vers une variable de Poison de paramètre λ.
>
> En pratique :
>
> Soit X une v.a. suivant la loi binomiale $\mathcal{B}(n\,;\,p)$.
> Si $n \geq 30$, $p \leq 0{,}1$ et $np < 15$, alors X suit approximativement la loi de Poisson de paramètre np.

I.2. Le théorème de Moivre - Laplace

On considère une suite (X_n) de variables binomiales $\mathcal{B}(n\,;\,p)$, p étant un paramètre fixé. Alors $T_n = \dfrac{X_n - np}{\sqrt{npq}}$ converge en loi vers $\mathcal{N}(0\,;\,1)$.

En pratique, pour $n \geq 30$, $np \geq 15$ et $npq \geq 5$ (Carnec), $\mathcal{B}(n\,;\,p)$ suit approximativement $\mathcal{N}(np\,;\sqrt{npq})$.

Les conditions changent suivant les auteurs :

Saporta :	n assez grand, $np > 5$ et $nq > 5$
Grais :	$npq \geq 9$
Le programme :	$n > 30$ et $0{,}3 < p < 0{,}7$
Un document du GTD :	$n > 30$, $np > 5$ et $nq > 5$

...

Attention à la correction de continuité !

Prenons $\mathcal{B}(100\ ;\ 0{,}5)$. $npq = 25$.

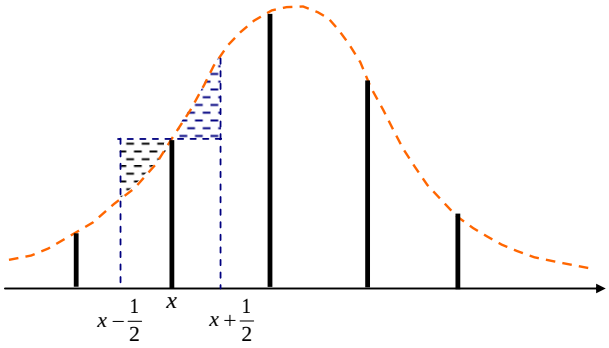
L'approximation de $T = \dfrac{X - 50}{5}$ par $\mathcal{N}(0\ ;\ 1)$ est justifiée.

On s'intéresse à $P(X \leq 52)$. La valeur exacte est 0,6914.

Par l'approximation de la loi normale on obtient :

$X \leq 52 \Leftrightarrow T \leq 0{,}4$ donc $P(X \leq 52) = \Pi(0{,}4)$

Soit $P(X \leq 52) =$ 0,6554. Erreur non négligeable !!



On corrige de la façon suivante : $P(X \leq x) \approx P\left(T \leq \dfrac{x + 0{,}5 - np}{\sqrt{npq}}\right)$

$P(X = x) \approx P\left(\dfrac{x - 0{,}5 - np}{\sqrt{npq}} \leq T \leq \dfrac{x + 0{,}5 - np}{\sqrt{npq}}\right)$

$P(y \leq X \leq x) \approx P\left(\dfrac{y - 0{,}5 - np}{\sqrt{npq}} \leq T \leq \dfrac{x + 0{,}5 - np}{\sqrt{npq}}\right)$

Pour l'exemple précédent, $P(X \leq 52) \approx \Pi(0{,}5) =$ 0,6915.

Exemple 1

On lance une pièce de monnaie « honnête » 1 000 fois. Quelle est la probabilité d'obtenir au moins 548 piles ?

Solution

X désigne le nombre de piles obtenus. On cherche $P(X \geq 548)$.

$P(X \geq 548) = 1 - P(X \leq 547)$.

On peut approcher par la loi normale. $n = 1\,000$; $np = 500$; $npq = 250$

Soit $\sqrt{npq} = 5\sqrt{10} \approx 15{,}8$.

$T = \dfrac{X - 500}{5\sqrt{10}}$ suit approximativement $\mathcal{N}(0\ ;1)$

$P(X \leq 547) \approx P\left(T \leq \dfrac{547{,}5 - 500}{5\sqrt{10}}\right)$

$P(T \leq 3) \approx 0{,}998\,65$ D'où $P(X \geq 548) \approx 0{,}001\,35$ (une chance sur 1 000)

Remarque

1. La valeur exacte de $P(X \geq 548$ est $1 - 0{,}998\,92 = 0{,}001\,18$
2. $P(X \in [\ m - 3\,\sigma\ ;\ m + 3\,\sigma\ ]) \approx 0{,}997$

Exemple 2

On organise un QCM de 100 questions. Pour chaque question il y a trois réponses possibles dont une et une seule est exacte.

Trouver un entier k tel que si un candidat a au moins k réponses justes, il y a moins de 5% de chances que toutes les réponses soient dues au hasard.

Solution

On note X la variable aléatoire donnant le nombre de réponses exactes par un candidat répondant à toutes les questions au hasard. Il s'agit de trouver un entier k tel que : $P(X \geq k) < 0,05$ ou $P(X \leq k-1) \geq 0,95$.

X suit $\mathcal{B}(100 ; 1/3)$. Les conditions d'approximation par la loi normale sont vérifiées et $np = \frac{100}{3}$, $\sqrt{npq} = \frac{10}{3}\sqrt{2}$.

$T = \dfrac{X - \frac{100}{3}}{\frac{10}{3}\sqrt{2}}$ suit approximativement $\mathcal{N}(0 ; 1)$ et

$P(X \leq k-1) \approx P\left(T \leq \dfrac{k-1+0,5-np}{\sqrt{npq}}\right) \geq 0,95 = \Pi(1,645)$

D'où $\dfrac{k-1+0,5-np}{\sqrt{npq}} \geq 1,645$ soit $k \geq 41,59$ donc $k \geq 42$.

Les « fourchettes » d'une fréquence binomiale

Soit X_n une variable aléatoire qui suit $\mathcal{B}(n ; p)$. On suppose les conditions d'approximation par la loi normale vérifiées ;

On pose $F_n = \frac{X}{n}$ la fréquence associée. Alors : $P\left(|F_n - p| \leq \frac{1}{\sqrt{n}}\right) > 0,95$

Fourchette au niveau 0,95

Démonstration

$T_n = \dfrac{X_n - np}{\sqrt{npq}}$ converge en loi vers $\mathcal{N}(0 ; 1)$. Donc $P(|T_n| \leq 1,96) \approx 0,95000435$

$|T_n| \leq 1,96 \Leftrightarrow \left|\dfrac{X_n - np}{\sqrt{npq}}\right| \leq 1,96 \Leftrightarrow |F_n - p| \leq \dfrac{1,96}{\sqrt{n}}\sqrt{pq}$

Or $pq = p(1-p) \leq 0,25$ donc $|F_n - p| \leq \frac{1}{\sqrt{n}}$. Ainsi $|T_n| \leq 1,96 \Rightarrow |F_n - p| \leq \frac{1}{\sqrt{n}}$ et

$P\left(|F_n - p| \leq \frac{1}{\sqrt{n}}\right) \geq P(|T_n| \leq 1,96) \approx 0,95000435 > 0,95$ cqfd.

On a, de même, les autres fourchettes :

$P\left(|F_n - p| \leq \dfrac{0,825}{\sqrt{n}}\right) > 0,90$; $P\left(|F_n - p| \leq \dfrac{1,5}{\sqrt{n}}\right) > 0,99$

Pour $n = 1000$ on obtient : $P(|F_{1000} - p| \leq 2,61\%) > 90\%$

$P(|F_{1000} - p| \leq 3,2\%) > 95\%$

$P(|F_{1000} - p| \leq 4,75\%) > 99\%$

I.3. Convergence de la loi de Poisson vers la loi normale

Soit une suite (X_n) de variables de Poisson $\mathcal{P}(n\lambda)$ où λ est un réel positif fixé.

Alors $T_n = \dfrac{X_n - n\lambda}{\sqrt{n\lambda}}$ converge en loi vers $\mathcal{N}(0\ ;\ 1)$ lorsque n tend vers $+\infty$.

L'approximation est satisfaisante dès que $n\lambda > 15$.

I.4. Le théorème central-limite

Soit $X_1, X_2, \ldots, X_n$ des variables aléatoires mutuellement indépendantes ayant la même loi de moyenne m et d'écart type σ.

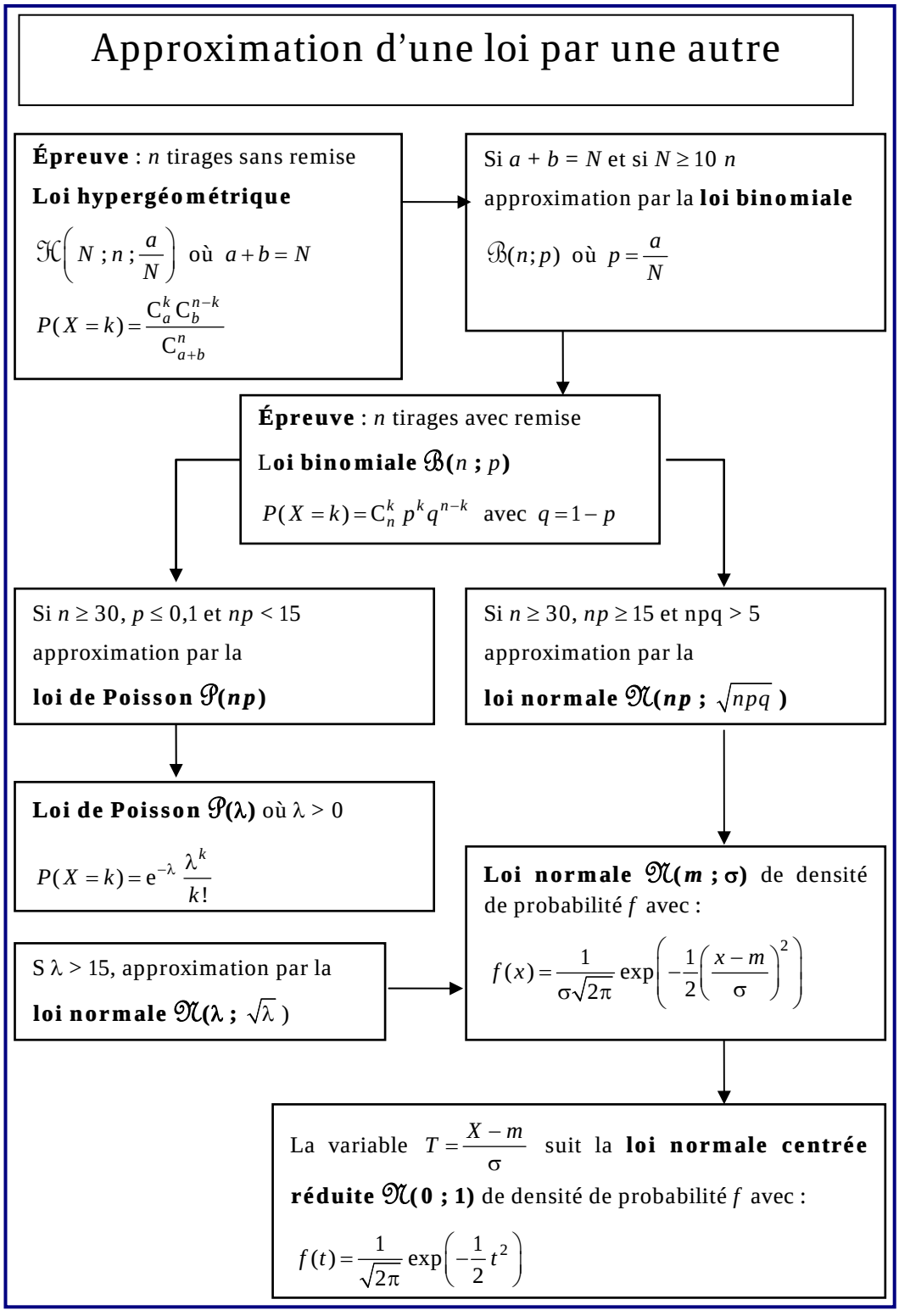
On pose $\overline{X}_n = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$

Alors $(\overline{X}_n)$ converge en loi vers $\mathcal{N}(m\ ;\ \dfrac{\sigma}{\sqrt{n}})$

Pour $\boldsymbol{n \geq 30}$, $\overline{X}$ suit approximativement $\mathcal{N}(m\ ;\ \dfrac{\sigma}{\sqrt{n}})$.

Les deux théorèmes de convergence précédents sont des cas particuliers de ce théorème plus général.

I.5. Résumé


Approximation d'une loi par une autre
Épreuve : n tirages sans remise
Loi hypergéométrique
$\mathcal{H}\left(N ; n ; \frac{a}{N}\right)$ où $a+b=N$
$P(X=k)=\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}$
Si $a + b = N$ et si $N \geq 10\ n$
approximation par la loi binomiale
$\mathcal{B}(n;p)$ où $p=\frac{a}{N}$
Épreuve : n tirages avec remise
Loi binomiale $\mathcal{B}(n ; p)$
$P(X=k)=C_n^k p^k q^{n-k}$ avec $q=1-p$
Si $n \geq 30$, $p \leq 0{,}1$ et $np < 15$
approximation par la
loi de Poisson $\mathcal{P}(np)$
Si $n \geq 30$, $np \geq 15$ et npq > 5
approximation par la
loi normale $\mathcal{N}(np ; \sqrt{npq})$
Loi de Poisson $\mathcal{P}(\lambda)$ où $\lambda > 0$
$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$
S $\lambda > 15$, approximation par la
loi normale $\mathcal{N}(\lambda ; \sqrt{\lambda})$
Loi normale $\mathcal{N}(m ; \sigma)$ de densité de probabilité f avec :
$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$
La variable $T=\frac{X-m}{\sigma}$ suit la loi normale centrée réduite $\mathcal{N}(0 ; 1)$ de densité de probabilité f avec :
$f(t)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\right)$

2. La convergence en probabilité

2.1. Définition

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{P}(\Omega), p)$ et X une variable aléatoire définie sur ce même espace.

On dit que la suite (X_n) converge en probabilité vers X si, et seulement si, quel que soit ε strictement positif : $\lim_{n \to +\infty} P(|X_n - X| < \varepsilon) = 1$

2.2. L'inégalité de Bienaymé-Tchébychev

Soit X une variable aléatoire de moyenne m et d'écart type σ.

Pour tout réel t strictement positif, $P(|X - m| \leq t) > 1 - \frac{\sigma^2}{t^2}$.

2.3. Le théorème de Bernoulli

Soit X une variable aléatoire qui suit $\mathcal{B}(n ; p)$.

On pose $F_n = \frac{X}{n}$ la fréquence associée.

Alors pour tout $\varepsilon > 0$, $P(|F_n - p| \leq \varepsilon) > 1 - \frac{pq}{n\varepsilon^2}$

C'est l'inégalité précédente avec $t = \varepsilon$ et $\sigma^2 = \frac{pq}{n}$.

Corollaire :

Soit X une variable aléatoire qui suit $\mathcal{B}(n ; p)$.

On pose $F_n = \frac{X}{n}$ la fréquence associée.

Alors, pour tout $\varepsilon > 0$, $\lim_{n \to +\infty} P(|F_n - p| \leq \varepsilon) = 1$

Autrement dit (F_n) converge en probabilité vers la variable aléatoire certaine p.

> Ce corollaire (version « allégée » de la loi faible des grands nombres) justifie le point de vue des « fréquentistes » qui attribuent comme probabilité d'un événement une valeur autour de laquelle la fréquence d'apparition de cet événement se stabilise lorsque le nombre d'expériences indépendantes devient très grand.

2.4. Un exemple d'application

Trouver un entier n_0 à partir duquel P($|F_n - 1/6| \leq 0{,}01$) > 0,95 :

– avec Bienaymé Tchebychev.

– avec la « fourchette » .

– avec la loi normale.

On lance un dé parfait n fois. On note F_n la fréquence de sortie de l'as.

Solution

1. $P\left(\left|F_n - \frac{1}{6}\right| \leq 10^{-2}\right) > 1 - \frac{\frac{1}{6} \times \frac{5}{6}}{n \times 10^{-4}}$

d'où $1 - \frac{5 \times 10^{-4}}{36n} \geq 0{,}95$ ou $\frac{5 \times 10^{-4}}{36n} \leq 5 \times 10^{-2}$ soit $n \geq \frac{10^6}{36}$

$n_0 = 27\,778$

2. Fourchette (conditions d'approximation supposées satisfaites)

$P\left(\left|F_n - \frac{1}{6}\right| \leq \frac{1}{\sqrt{n}}\right) > 0{,}95.$

Il suffit de prendre $\frac{1}{\sqrt{n}} \leq 10^{-2}$ soit $n_0 = 10^4$

3. Approximation par la loi normale

On suppose $n > 30$. Ainsi les conditions sont vérifiées.

nF_n suit $\mathcal{B}\left(n ; \frac{1}{6}\right)$ et $T_n = \frac{nF_n - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}$ suit approximativement $\mathcal{N}(0 ; 1)$.

$T_n = \frac{F_n - \frac{1}{6}}{\frac{1}{6}\sqrt{\frac{5}{n}}}$

$P\left(\left|F_n - \frac{1}{6}\right| \leq 10^{-2}\right) \approx P\left(|T_n| \leq \frac{6 \times 10^{-2}}{\sqrt{\frac{5}{n}}}\right)$

$P\left(|T_n| \leq \frac{6 \times 10^{-2}}{\sqrt{\frac{5}{n}}}\right) > 0{,}95 = P(|T_n| \leq 1{,}96)$ d'où $\frac{6 \times 10^{-2}}{\sqrt{\frac{5}{n}}} > 1{,}96$

soit $\frac{\sqrt{n}}{\sqrt{5}} > \frac{196}{6}$ donc $n > \frac{5 \times 196^2}{36} \approx 5\,325{,}55$

$n_0 = 5\,336$

Comparer les trois méthodes ! !

Annexe :
Sources et Bibliographie

Des statistiques à la pensée statistique (2001) :IREM de Montpellier.

Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?, (1998), IREM de Lyon.

Enseigner la statistique au lycée : des enjeux aux méthodes (2001), brochure n°112 de la commission Inter-IREM Lycées Technologiques.

Probabilités, analyse des données et statistique (1990) par G. SAPORTA, éditions Technip.

Le jeu de la science et du hasard (1994) par Daniel SCHWARTZ, éditions Champs/Flammarion.

Statistiques au collège par J.C. GIRARD, revue Repères n° 23 d'avril 1996.

Itinéraires en statistiques et probabilités par CARNEC, DAGOURY, SEROUX et THOMAS, éditions Ellipses (2000).

Exercices ordinaires de probabilités (1992) par G. FRUGIER, éditions Ellipses.

Cours de probabilités et statistiques (1987) par LEBOEUF, ROQUE, GUEGAND, éditions Ellipses.

Stage de Statistiques (2001) par P. TERRACHER, IREM de Bordeaux.

Les statistiques au collège (1997) par Yves OLIVIER, IA-IPR de Mathématiques.

Un modèle de connaissance incontournable : la Statistique par J.L. PIEDNOIR, Inspecteur Général de mathématiques