République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université 8 Mai 1945 Guelma

Faculté des Mathématiques et de l'Informatique et des Sciences de la Matière
Département de Mathématiques

BROCHURE

Par :

BAHLOUL TAREK

Intitulé

Mathématiques 2

Année Universitaire **2020-2021**

Contents

Introduction

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. Son objectif peut se résumer de la façon suivante: dégager, à partir de données observées sur quelques individus d'une population, des résultats valables pour l'ensemble de la population, résumé clair avec le minimum de perte d'information, permettant de dégager plus facilement un diagnostic.

Il s'agit alors de la statistique descriptive qui recouvre les moyens de présenter ces données et d'en décrire les principales caractéristiques, en les résumant sous forme de tableaux ou de graphiques. Il s'agira ensuite de les interpréter. La description statistique se propose de mettre en évidence certaines permanences ou lois statistiques, qui peuvent éventuellement conduire à des prévisions (élément essentiel de l'étude des séries chronologiques). Une règle qui transforme un ensemble de données en une ou plusieurs valeurs numériques se nomme une statistique, le terme étant cette fois utilisé avec l'article indéfini.

Les méthodes statistiques sont aujourd'hui utilisées dans presque tous les secteurs de l'activité humaine et font partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du biologiste, de l'informaticien $\cdots$ (les statistiques sont partout). Ceci révèle que le monde moderne est presque entièrement tourné vers le quantitatif et le mesurable. D'où l'intérêt de la statistique, discipline relativement récente, mais qui correspond parfaitement à cette orientation du monde moderne.

La théorie des probabilités construire l'outil de base à un ensemble de méthodes objectives permettant d'utiliser des données pour fixer la précision avec laquelle on estime certains paramètres (théorie de l'estimation) ou on teste certaines hypothèses (théorie des tests) .

La brochure est composée d'une introduction et de cinq sections. La première section est introductive est consacré à la définition de la statistique descriptive ainsi que des différents termes qui en constituent le vocabulaire de base.

Dans la deuxième section, nous avons exposé la méthode des moindres carrés pour l'ajustement d'un nuage de points par une droite qui est la fonction analytique la plus simple, mais cette méthode peut se généraliser à un ajustement par d'autres fonctions analytiques. On peut aussi dans certains

cas transformer une des deux variables ou les deux variables avant d'envisager une relation linéaire.

On dispose dans la troisième section d'un modèle statistique paramétrique des qu'on pose qu'une observable est distribuée selon un modèle d'échantillonnage ou seulement le paramètre est inconnu, mais appartient à un espace, de dimension finie, que la littérature scientifique appelle souvent ensemble des états de la nature. Toute conclusion sur une population statistique implique d'une façon ou d'une autre le paramètre du modèle d'échantillonnage choisi pour la représenter.

Quel que soit le paramètre inconnu () a estimer, le mode de raisonnement du statisticien classique est toujours le même. Dans sa tête, () a une valeur unique et son estimation requiert une statistique dont les paramètres dépendent de (). Les données disponibles permettent de calculer un intervalle de confiance correspondant a un risque a fixe. Le paramètre inconnu () est ou n'est pas dans cet intervalle. Aussi, pour décrire son incertitude sur (), le statisticien classique réalise un tour de passe-passe. Il imagine une collection d'échantillons recueillis dans les mêmes conditions et pour chacun d' entre eux, il **calcule** un intervalle de confiance et conclut en disant que $1 - \alpha$ pour cent d'entre-eux contiendraient ().

Dans la quatrième section, certains tests non paramétriques essayons plutôt de caractériser les situations où il est plus (ou moins) avantageux de les utiliser.

Au cours de la cinquième section, nous allons donner la définition d'un certain nombre de termes du vocabulaire utilisé dans un contexte non déterministe et indiquer comment construire le modèle adéquat.

1 Statistique descriptives:

La statistique descriptive est un ensemble de méthodes permettant de décrire, présenter, résumer des données souvent très nombreuses. Ces méthodes peuvent être numériques (tris, élaboration de tableaux, calcul de moyennes, $\cdots$) et/ou mener à des représentations graphiques.

1.1 Définitions

1.1.1 Population

Une population est l'ensemble des éléments auxquels se rapportent les données étudiées. En statistique, le terme $\ll$ population $\gg$ s'applique à des ensembles de toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprises d'un secteur donné $\cdots$)

1.1.2 individu

Dans une population donnée, chaque élément est appelé $\ll$ individu $\gg$ ou $\ll$ unité statistique $\gg$.

1.1.3 échantillon

La collecte d'informations sur une population peut être effectuée sur la totalité des individus ; on parle alors d'enquêtes exhaustives . Lorsque la taille de la population étudiée est élevée, de telles enquêtes sont fort coûteuses ou impossibles, et le cas échéant, leurs résultats souvent très longs à rassembler peuvent être dépassés avant même la fin de l'enquête. C'est la raison pour laquelle on a souvent recours aux enquêtes par sondage qui portent sur une partie de la population appelée échantillon.

1.2 Variables

Chaque individu d'une population peut être décrit selon une ou plusieurs variables qui peuvent être des caractéristiques qualitatives ou prendre des valeurs numériques.

1.2.1 Variable qualitative

Une variable est dite qualitative si ses différentes réalisations (modalités) ne sont pas numériques.

Exemple Le sexe, la situation matrimoniale, la catégorie, socioprofessionnelle $\cdots$

1.2.2 Variable quantitative

Une variable est dite quantitative lorsqu'elle est intrinsèquement numérique. Une variable quantitative peut être une variable statistique discrète ou continue.

1.2.3 Variables statistiques discrètes

Les variables statistiques discrètes sont des variables qui ne peuvent prendre que des valeurs isolées, discrètes

Exemple Le nombre d'enfants d'une famille, le nombre de pétales d'une fleur, le nombre de buts marqués lors d'une rencontre de football $\cdots$

1.2.4 Variables statistiques continues

Les variables statistiques continues peuvent prendre toutes les valeurs numériques possibles d'un ensemble inclus dans $\mathbb{R}$.

Exemple Le revenu, la taille, le taux de natalité $\cdots$

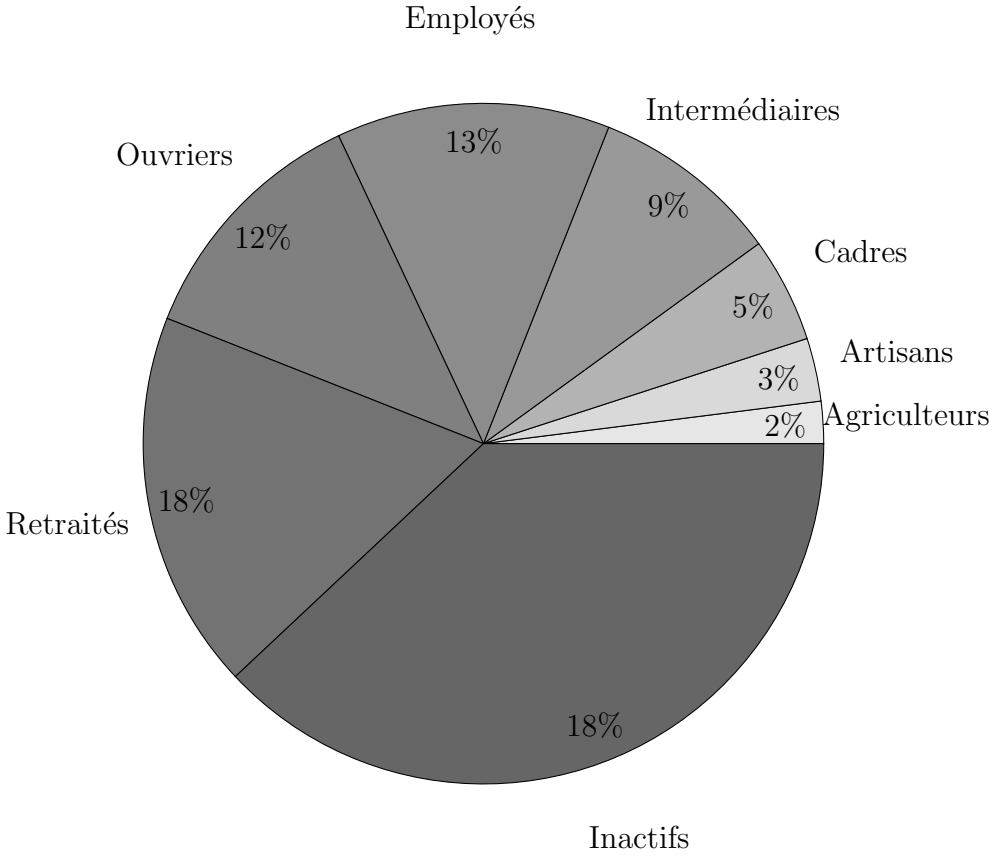
1.3 Représentations graphiques

L'étude d'une population selon une variable sera restreinte au cas des variables quantitatives, car la description d'une population selon une variable qualitative est totalement résumée dans un tableau de pourcentages ou dans un diagramme circulaire, appelé aussi diagramme en $\ll$ camembert $\gg$.

Exemple Voici la répartition, en pourcentage, de la population française, par catégories socioprofessionnelles, pour l'année 1999 :

Catégorie	% (effectif)
Agriculteurs exploitants	2
Artisans, commerçants, chefs d'entreprise	3
Cadres, professions intellectuelles supérieures	5
Professions intermédiaires	9
Employés	13
Ouvriers	12
Retraités	18
Autres sans activité professionnelle	38

Dans ce cas, il est assez naturel de représenter ces données à l'aide d'un diagramme à secteurs. Le disque complet représentera la population française.



1.3.1 diagramme en barres

Ces diagrammes sont des représentations graphiques très ≪ parlantes ≫ visuellement, que l'on peut utiliser aussi bien en calcul des probabilités qu'en

statistique, pour des variables quantitatives continues comme pour des variables quantitatives discrètes.
L'axe des abscisses porte les valeurs de la variable.
L'axe des ordonnées porte les valeurs des fréquences ou des effectifs ou des probabilités.
Les hauteurs des rectangles sont proportionnelles aux effectifs.

Exemple Le tableau suivant donne la distribution du niveau de l'indice de la qualité de l'air ATMO en agglomération parisienne de 2000 à 2006 (en nombre de jours par an).

1. Définir les populations étudiées, l'unité statistique, le caractère étudié et sa nature.

2. Tracez le diagramme en bâtons de la distribution en 2006, et indiquez le mode.

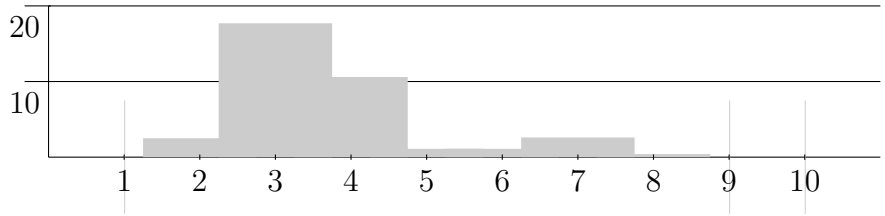
3. Calculez les niveaux annuels moyens de 2000 à 2006.

Niveau	Qualité	2000	2001	2002	2003	2004	2005	2006	Total
1	Très bon	0	0	0	0	0	0	0	0
2	Très bon	8	15	9	15	23	23	25	118
3	Bon	206	190	183	138	186	188	177	1 268
4	Bon	99	97	111	109	96	99	106	717
5	Moyen	36	33	45	47	39	34	11	260
6	Médiocre	13	13	8	30	19	16	26	110
7	Médiocre	2	14	7	16	2	6	11	58
8	Mauvais	2	3	2	10	1	4	4	26
9	Mauvais	0	0	0	0	0	0	0	0
10	Très mauvais	0	0	0	0	0	0	0	0
	Total	366	365	365	365	366	365	365	2 557

Réponses

1. Population : formée de 7 sous-populations associées chacune à une année (2000 à 2006) ; l'ensemble des jours d'une année constitue la population de l'année. Unité statistique : une journée d'une année. Caractère étudié : niveau de l'indice de la qualité de l'air, caractère qualitatif, mais aussi ordinal (les modalités du caractère sont ordonnées).

2. Diagramme en bâtons de la distribution 2006 : mode = niveau 3 .



3. -

Année	2000	2001	2002	2003	2004	2005	2006	2000-2006
Niveau moyen	3.6	3.7	3.7	4.1	3.6	3.6	3.7	3.7

Le niveau moyen a été particulièrement élevé en 2003

1.3.2 Courbe de concentration

Elle est utilisée principalement en statistique économique pour étudier les inégalités de répartition d'une grandeur positive cumulable.

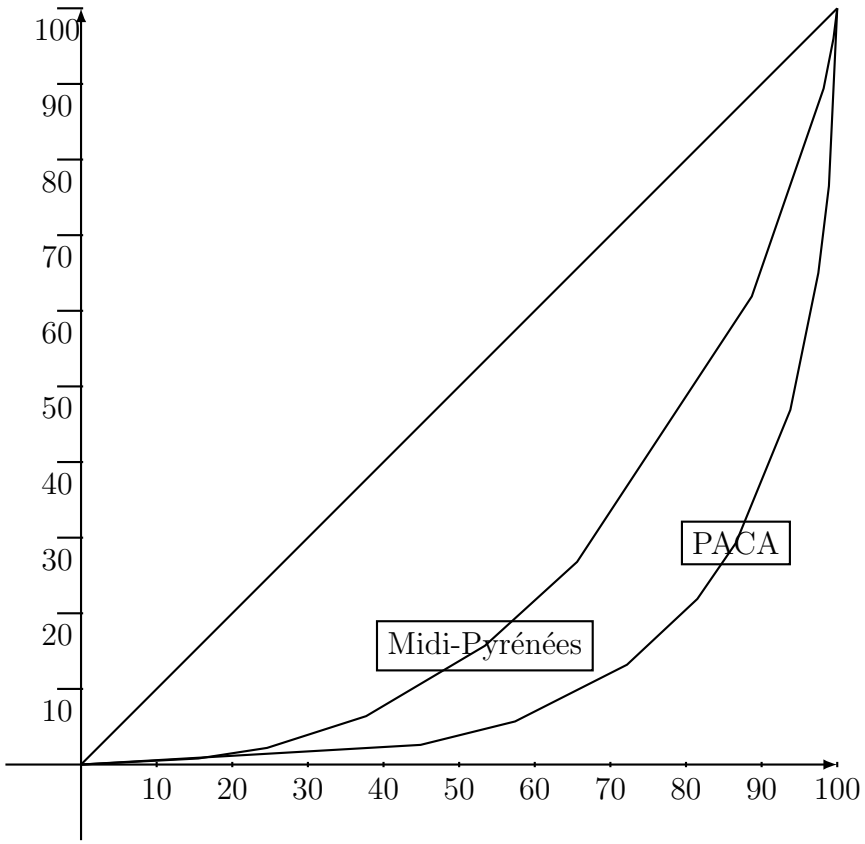
Construction de la courbe de concentration Considérons la distribution des exploitations agricoles par classes de grandeurs des régions Provence-Alpes-Côte d'Azur (PACA) et Midi-Pyrénées en 2005

	Midi-Pyrénées		PACA	
	f_i	Proportion SAU	f_i	Proportion SAU
Moins Mode 5 ha	15.5	0.8	44.9	2.6
5 à moins de 10 ha	9.0	1.4	12.5	3.1
10 à moins de 20 ha	13.2	4.2	14.8	7.6
20 à moins de 35 ha	15.7	9.2	9.3	8.6
35 à moins de 50 ha	12.2	11.1	5.1	7.4
50 à moins de 100 ha	23.1	35.1	7.2	17.6
100 à moins de 200 ha	9.6	27.5	3.7	18.1
200 à moins de 300 ha	1.3	6.6	1.4	11.5
300 ha ou plus	0.5	4.0	1.1	23.5
	100	100	100	100

	Midi-Pyrénées		PACA	
	p_i	q_i	p_i	q_i
Moins Mode 5 ha	15.5	0.8	44.9	2.6
5 à moins de 10 ha	24.6	2.2	57.4	5.7
10 à moins de 20 ha	37.7	6.4	72.2	13.2
20 à moins de 35 ha	53.4	15.7	81.5	21.9
35 à moins de 50 ha	65.6	26.8	86.6	29.3
50 à moins de 100 ha	88.7	61.9	93.8	46.9
100 à moins de 200 ha	98.2	89.4	97.5	65.0
200 à moins de 300 ha	99.5	96.0	98.9	76.5
300 ha ou plus	100	100	100	100

Tableau – Distribution des exploitations agricoles par classes de grandeurs en régions PACA et Midi-Pyrénées

Ceci suggère d'utiliser l'aire, dite aire de concentration, comprise entre la courbe et la bissectrice **OB** comme indicateur d'inégalité



Courbes de concentration des SAU dans les régions PACA et Midi-Pyrénées

On peut comparer la concentration de deux ou plusieurs populations selon un même caractère en représentant sur un même graphique leurs courbes de Lorenz. Les terres agricoles sont plus concentrées dans la région PACA que dans la région Midi-Pyrénées puisque la courbe de Lorenz de la SAU de la région Midi-Pyrénées est incluse dans celle de la région PACA.

Exemple Le tableau suivant donne le nombre (en milliers) et la superficie agricole utilisée (SAU, en milliers d'ha) des exploitations agricoles en France métropolitaine par classes de grandeur pour les années 1979, 1988, 2000 et 2005.

	1979		1988	
	Nombre	SAU	Nombre	SAU
Moins de 5 ha	357	677	278	519
5 à moins de 20 ha	410	4 778	279	3 238
20 à moins de 50 ha	347	10 962	288	9 348
50 à moins de 100 ha	114	7 683	128	8 709
100 à moins de 200 ha	29	3 798	37	4 864
200 ha ou plus	6	1 598	7	1 918
Ensemble	1 263	29 496	1 017	28 596

	2000		2005	
	Nombre	SAU	Nombre	SAU
Moins de 5 ha	193	362	132	262
5 à moins de 20 ha	132	1 464	104	1 163
20 à moins de 50 ha	138	4 666	109	3 714
50 à moins de 100 ha	122	8 662	113	8 083
100 à moins de 200 ha	64	8 655	70	9 486
200 ha ou plus	15	4 047	17	4 762
Ensemble	664	27 856	545	27 470

1. Définir la population, l'unité statistique, le caractère étudié et sa nature.

2. Calculez, en pourcentage, les taux annuels moyens de variation du nombre des exploitations agricoles de 1979 à 1988, de 1988 à 2000, de 2000 à 2005. Exprimez le taux annuel moyen de variation de 1979 à 2005 en fonction de ces 3 taux, de quel type de moyenne s'agit-il ?
   Calculez sa valeur.

3. Pour les années 1979, 1988, 2000 et 2005, calculez la SAU moyenne et la SAU moyenne des exploitations de 50 ha ou plus.

4. Pour l'année 2005, représentez l'histogramme de la distribution des exploitations agricoles, ainsi que la courbe de concentration de la SAU.

Réponses

1. Population : les exploitations agricoles de France métropolitaine en 1979, 1988, 2000 et 2005
   Unité statistique : une exploitation agricole de France métropolitaine en 1979, 1988, 2000 et 2005
   Caractère étudié : la taille de la SAU, variable statistique continue.

2. Soit c_1, c_2 et c_3 les taux annuels moyens de variation au cours de chacune des 3 périodes :

$$(1+c_1)^9 = \frac{1017}{1293} = (0,80522)^9 \Rightarrow c_1 \approx -2.4\%$$

$$(1+c_2)^{12} = \frac{664}{1017} = (0,65290)^{12} \Rightarrow c_2 \approx -3.5\%$$

$$(1+c_3)^5 = \frac{545}{664} = (0,96127)^3 \Rightarrow c_3 \approx -3.9\%$$

   Le taux annuel moyen de variation c de 1979 à 2005 est une moyenne géométrique pondérée des 3 taux c_1, c_2 et c_3 :

$$1+c =^{26} \sqrt{(1+c_1)^9(1+c_2)^{12}(1+c_3)^5}$$

$$\Rightarrow 1+c =^{26} \sqrt{\frac{545}{1293}} \approx 0,96819 \Rightarrow c \approx -3.2\%$$

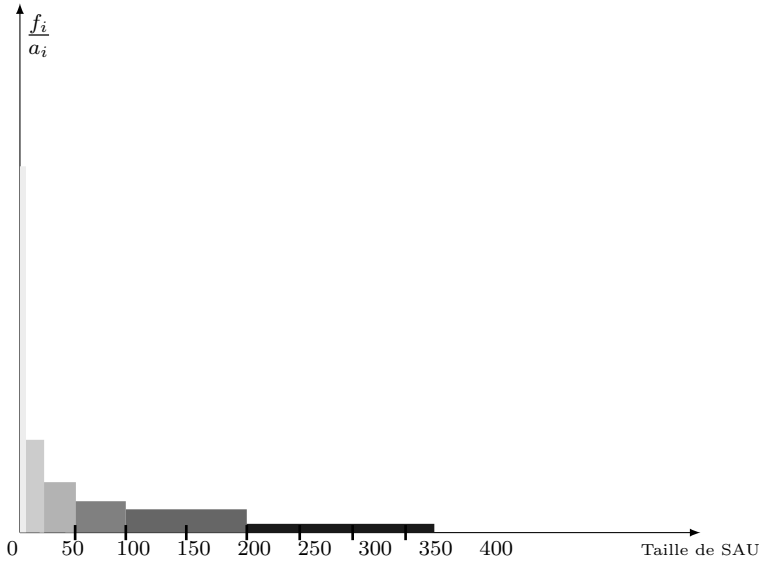
3. -

	1979	1988	2000	2005
SAU moyenne	23	28	42	50
SAU moyenne des exploitations de 50 ha ou plus	88	90	106	112

   Le nombre des exploitations agricoles diminue, la taille moyenne des SAU augmente, ainsi que la taille moyenne des exploitations de 50 ha ou plus.

4. Le centre de la dernière classe étant par hypothèse la SAU moyenne des exploitations de 200 ha ou plus est égale en 2005 à 280 ($= \frac{4762}{17}$). On évalue ainsi la SAU maximum approximativement à 360 ha.
   L'histogramme comporte 6 classes : 6 rectangles de hauteur $\frac{f_i}{a_i}$.

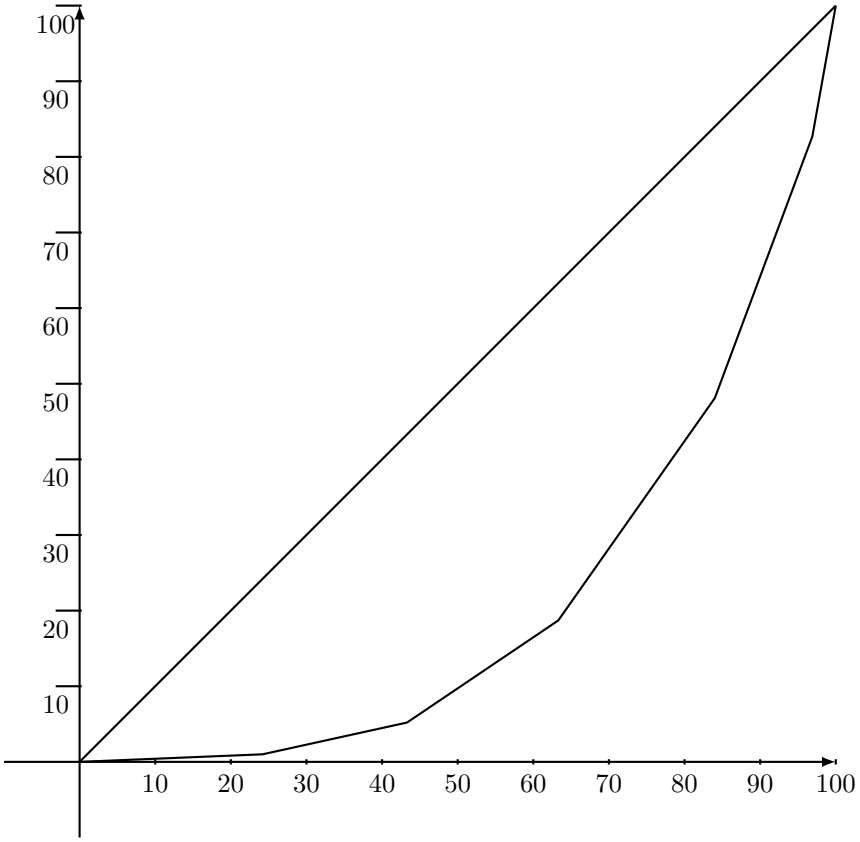
Taille de SAU	[0, 5[	[5, 20[	[20, 50[	[50, 100[	[100, 200[	[200, 360[
f_i	24,2	19,1	20,0	20,7	12,8	3,1
$\frac{f_i}{a_i}$	4,844	1,272	0,667	0,415	0,128	0,019



Histogramme de la distribution des exploitations agricoles

Courbe de concentration

$p_i(\%)$	0	24,2	43,3	63,3	84,0	96,9	100
$q_i(\%)$	0	1,0	5,2	18,7	48,1	82,7	100



1.3.3 Boîte à moustaches ou box-plot

Ce diagramme, introduit par J.W. Tukey, est une représentation synthétique extrêmement efficace des principales caractéristiques d'une variable numérique.
La boîte correspond à la partie centrale de la distribution.
Les moustaches s'étendent de part et d'autre de la boîte.
Représentation graphique symbolique de la médiane et des quartiles d'une distribution statistique.

Les quantiles Les quantiles sont des indicateurs de position.
Le quantile d'ordre α

$$0 \leq \alpha \leq 1,$$

noté x_α, est tel qu'une proportion α des individus ait une valeur du caractère X inférieure ou égale à x_α.
Le quantile $x_{0.5}$ est égal à **la médiane**.
On utilise couramment les quantiles d'ordre $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$. Ils sont ainsi notés et nommés :

- Q_1 = premier quartile = $x_{0.25}$.
- Q_2 = deuxième quartile = médiane = $x_{0.5}$.
- Q_3 = troisième quartile = $x_{0.75}$.

L'intervalle interquartile Les quartiles Q_1,Q_2, Q_3 étant définis par

$$F(Q_1) = 0.25, \quad F(Q_2) = 0.50 \quad et \quad F(Q_3) = 0.75,$$

$\mid Q_3 - Q_1 \mid$ est un indicateur parfois utilisé pour mesurer la dispersion: il est plus robuste que l'étendue.
l'intervalle interquartile mesure l'étendue des 50% de valeurs situées au milieu d'une série de données classées.
Ainsi, la taille de la boîte représente l'étendue interquartile, la position de la médiane est un bon indicateur de la symétrie de la distribution, la taille des lignes de part et d'autre de la boîte traduit la dispersion, et les valeurs éloignées ou extrêmes sont immédiatement repérées.
On représente une boîte de distribution de la façon suivante:

- on trace un rectangle de largeur fixée à priori et de longueur

  $$EIQ = (Q_3 - Q_1),$$

  et on y situe la médiane par un segment positionné à la valeur Q_2, par rapport à Q_3 et Q_1 ; on a alors la boîte,

- on calcule

  $$Q_3 + 1.5 \times EIQ$$

  et

  $$Q_1 - 1.5 \times EIQ$$

  et on cherche :

  1. la dernière observation x_h en deçà de la limite $(Q_3 + 1.5 \times EIQ)$ soit
     $$x_h = \max\{x_i | x_i \leq Q_3 + 1.5 \times EIQ\}$$
  2. la première observation x_b au delà de la limite $(Q_1 - 1.5 \times EIQ)$ soit
     $$x_b = \min\{x_i | x_i \geq Q_1 - 1.5 \times EIQ\}$$

- on trace deux lignes allant des milieux des largeurs du rectangle aux valeurs x_b et x_h.

Exemple Le tableau ci-après donne des caractéristiques des 30 premiers groupes français de l'industrie et des services selon leur chiffre d'affaires en 2001 :

Société	CAHT (millions d' €)	Effectif
TotalFinaElf	105 318	122 025
Carrefour	69 486	382 821
Vivendi Universal	57 360	321 000
PSA Peugeot Citroën	51 663	192 500
France Telecom	43 026	206 184
Suez	42 359	188 050
EDF	40 716	161 738
Les Mousquetaires	37 200	112 000
Renault	36 351	140 417
Saint-Gobain	30 390	173 329
Pinault-Printemps- La Redoute	27 799	115 935
Groupe Auchan	26 200	136 000
Alcatel Alsthom	25 353	99 314
Galec (Leclerc)	25 000	75 000
Alstom	23 453	118 995
Aventis	22 941	91 729
Groupe Casino (Rallye)	21 984	106 736
Bouygues	20 473	126 560
Airbus (EADS)	20 427	2 000
SNCF	20 129	220 747
Vonci	17 172	129 499
La poste	17 028	313 854
Publicis Groupe	16 667	20 592
Michelin	15 775	127 467
Havas	14 950	20 373
Usinor (Arcelor)	14 523	59 516
Groupe Danone	14 470	100 560
Gaz de France	14 357	36 451
L'Oréal (Gespartal)	13 740	49 150
Lafarge	13 698	82 892

1. Définir la population étudiée, l'unité statistique et les caractères étudiés.
2. Calculez la moyenne et l'écart-type du chiffre d'affaires et de l'effectif.
3. Étude du chiffre d'affaires des 30 premiers groupes français.

- Déterminez les trois quartiles.
- Représentez le diagramme branche et feuille de cette distribution.
- Représentez la boîte de distribution.

4. Quel est l'intérêt de chacune de ces deux représentations graphiques comparativement à un histogramme ?

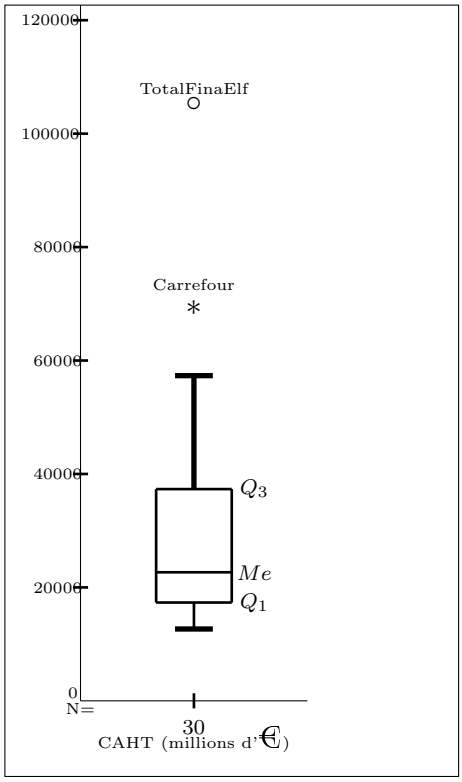
5. Reprendre la question 3 pour l'étude de l'effectif.

Réponses

1. Population : les 30 premiers groupes français de l'industrie et des services selon leur CAHT en 2001. Unité statistique : un groupe parmi les 30 premiers groupes français de l'industrie et des services selon leur CAHT en 2001. Caractères étudiés : deux caractères quantitatifs, le CAHT en millions d'€et l'effectif

2. CA : $n = 30$
   $\overline{x} = 30000$ millions d'*euro* $s_X = 19729$ millions d'€
   Effectif : $n = 30$
   $\overline{y} \approx 134448 \quad s_y \approx 87248.$

3. - n = 30 $\Rightarrow$ P(Me) = 15,5 $\Rightarrow$ Me = 23 197 millions d'€
     P(Q) = 8 $\Rightarrow$ Q_1 = 16 667 millions d'€et Q_3 = 37 200 millions d'€
   - Graphiques
   - Graphiques

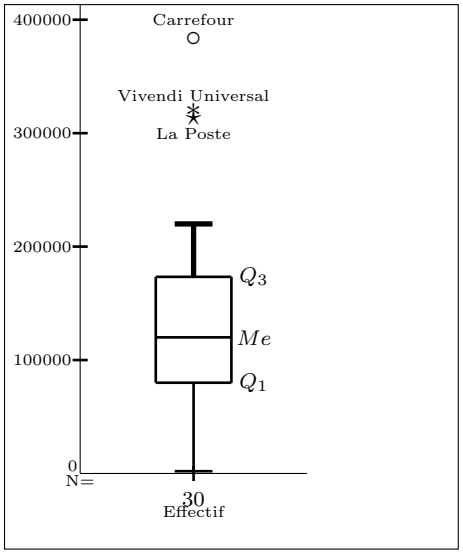


4. Le diagramme ≪ branche et feuille ≫ ne peut s'envisager que pour des distributions de population de taille peu élevée, contrairement à l'histogramme où l'hypothèse d'équirépartition à l'intérieur des classes n'est réaliste qu'avec un effectif suffisant dans chaque classe. Cette représentation permet de plus de ne pas perdre l'information valeur par valeur et aussi d'étiqueter éventuellement les observations. La boîte de distribution met en évidence une valeur éloignée (Carrefour) et une valeur extrême (TotalFinaElf). Cette distribution asymétrique étalée vers les valeurs élevées sera modélisée par la loi de Pareto.

5. La série étant ordonnée selon le CA, il faut maintenant l'ordonner selon l'effectif
   n = 30 ⇒ P(Me) = 15,5 ⇒ Me = 120 510
   P(Q) = 8 ⇒ Q_1 = 82 892 et Q_3 = 173 329

   - Graphiques
   - Graphiques

La boîte de distribution met en évidence trois valeurs éloignées : Carrefour, Vivendi Universal et La Poste.



1.4 Paramètre de dispersion et de position:

1.4.1 Paramètres de positions (Caractéristiques de tendance centrale)

Les plus usitées sont la **médiane**, la **moyenne arithmétique** et le **mode**.

1- **La médiane**

Est la valeurs qui partage la distribution ou la série des valeurs (en statistique) en deux parties de même effectif (50% de l'effectif total). Elle se note le plus souvent M ou Me.
effectif. Elle se détermine soit à partir de la série des valeurs ordonnées, soit à partir de la fonction cumulative.
Si la série statistique présente des observations individualisées en nombre impair

$$2m+1 : x_1 < x_2 < \cdots < x_{2m+1},$$

la médiane Me est la valeur centrale:

$$Me = x_{m+1}.$$

Si la série statistique présente des observations individualisées en nombre pair

$$2m : x_1 < x_2 < \cdots < x_{2m},$$

on peut prendre pour médiane Me est toute valeur de l'intervalle central:

$$Me \text{ au choix entre } x_m \text{ et } x_{m+1}.$$

Le choix est le plus souvent

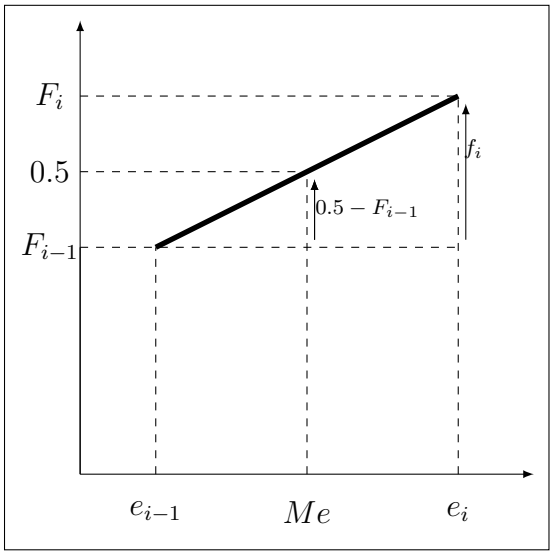
$$\frac{x_m + x_{m+1}}{2}.$$

Lorsque l'on ne connait qu'une répartition en classes on cherche la classe médiane $[e_{i-1}, e_i]$ telle que:

$$F(e_{i-1}) < 0.5 \quad et \quad F(e_i) > 0.5$$

et on détermine Me par interpolation linéaire :

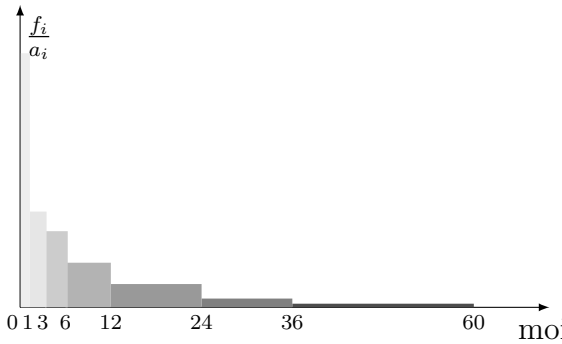
$$Me = e_{i-1} + (e_i - e_{i-1})\frac{0.5 - F_{i-1}}{f_i}.$$



Exemple

	Distribution en pourcentage	
Ancienneté d'inscription	Hommes	Femmes
Moins d'un mois	16,5	16,8
D'un à moins de trois mois	18,6	19,0
De trois à moins de six mois	15,5	15,1
De six mois à moins d'un an	18,5	17,7
D'un à moins de deux ans	18,0	18,5
De deux à moins de trois ans	6,8	7,0
Trois ans ou plus	6,1	5,8
Ensemble	100	100

Tableau – Chômeurs BIT selon le sexe et l'ancienneté de chômage en septembre 2006



Histogramme de la distribution des chômeurs ≪ Femmes ≫ selon l'ancienneté

La classe ≪ Trois ans ou plus ≫ est supposée bornée supérieurement par 5 ans (60 mois).

Pour la distribution de l'ancienneté du chômage des femmes, la médiane appartient à la classe [3 ; 6[

$$Me = 3 + 3 \times \frac{50 - 35.8}{15.1} \simeq 5.8 \quad mois$$

2- **Les moyennes**

2-1-) **La moyenne arithmétique:** On appelle moyenne arithmétique la somme de toutes les données statistiques divisée par le nombre de ces données. La moyenne arithmétique conserve la somme totale des valeurs observées : si on modifie les valeurs de deux observations d'une série statistique tout en conservant leur somme, la moyenne de la série sera inchangée

2-1-1-) **La moyenne arithmétique simple :**

-) de 2 nombres x et y :

$$\overline{x} = \frac{x+y}{2}.$$

-) de n nombres $x_1, x_2, \cdots, x_n$:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Remarque en remarquant que la somme totale

$$n\overline{x}$$

s'obtient en additionnant

$$n_1\overline{x}_1 \quad et \quad n_2\overline{x}_2$$

donc

$$\overline{x} = \frac{1}{n}\left(n_1\overline{x}_1 + n_2\overline{x}_2\right).$$

Exemple

	Distribution en milliers	
Ancienneté d'inscription	Hommes	Femmes
Moins d'un mois	180,3	181,0
D'un à moins de trois mois	203,9	204,9
De trois à moins de six mois	169,3	163,1
De six mois à moins d'un an	202,1	191,1
D'un à moins de deux ans	197,3	199,3
De deux à moins de trois ans	74,5	75,4
Trois ans ou plus	67,1	62,9
Ensemble	1 094,5	1 077,7
Ancienneté moyenne en jours	341	334

Tableau – Chômeurs BIT selon le sexe et l'ancienneté de chômage en septembre 2006

L'ancienneté moyenne d'inscription au chômage pour hommes et femmes réunis en septembre 2006 est égale à

$$\overline{x} = \frac{1}{2172.2}\left(1094,5 \times 341 + 1077,7 \times 334)\right) \simeq 338 \quad jours$$

-) Pour des données réparties en k classes la formule:

$$\Sigma_{i=1}^{k} f_i c_i, \quad c_i = \frac{e_i + e_{i-1}}{2}.$$

2-1-2-) **La moyenne arithmétique pondérée :**

*)

$$\frac{\Sigma_{i=1}^{k} n_i x_i}{\Sigma_{i=1}^{k} n_i}.$$

*)

$$\frac{\Sigma_{i=1}^{k} n_i c_i}{\Sigma_{i=1}^{k} n_i}, \quad c_i = \frac{e_i + e_{i-1}}{2}.$$

2-1-3-) **La moyenne élaguée:**

Exemple : Soit la série de notes d'un élève au cours de l'année $\{12, 13, 11, 14, 2\}$.

Dans certains cas, on retire les valeurs extrêmes et on calcule la moyenne uniquement sur un intervalle de valeurs élagué, le principe est identique quand les données sont groupées par valeurs ou par classes.

$$\frac{\Sigma_{i=1}^{n} x_i}{n} = \frac{12 + 13 + 11 + 14}{4} = 12.5.$$

2-2-) **La moyenne quadratique:**

2-2-1-) **La moyenne quadratique simple:**

$$\sqrt{\frac{\Sigma_{i=1}^{n} x_i^2}{n}}.$$

2-2-2-) **La moyenne quadratique pondérée:**

$$\sqrt{\frac{\Sigma_{i=1}^{k} n_i x_i^2}{\Sigma_{i=1}^{k} n_i}}.$$

2-3-) **La moyenne géométrique:** C'est la moyenne applicable à des mesures de grandeurs dont la croissance est géométrique ou exponentielle

2-3-1-) **La moyenne géométrique simple:**

$$(\Pi_{i=1}^{n} x_i)^{\frac{1}{n}}.$$

Exemple Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie suivante, ils aient été multipliés par 4 ; le coefficient multiplicateur moyen par décennie est égal à :

$$\sqrt{2 \times 4} = \sqrt{8} \simeq 2.83.$$

La moyenne arithmétique (= 3) n'est pas égale au coefficient demandé.
Prenons, par exemple, un salaire de 300 €au début de la première décennie,il sera de $300 \times 2 \times 4 = 2400$ € au bout des vingt ans, ce qui équivaut à $300 \times (2.83)^2$, soit un coefficient multiplicateur moyen de 2,83 par décennie.

2-3-2-) **La moyenne géométrique pondérée:**

$$\left(\Pi_{i=1}^{k} x_i^{n_i}\right)^{\frac{1}{\Sigma_{i=1}^{k} n_i}}.$$

2-4-) **La moyenne harmonique :** La moyenne harmonique est l'inverse de la moyenne arithmétique des inverses des valeurs.

2-4-1-) **La moyenne harmonique simple:**

$$\frac{n}{\Sigma_{i=1}^{n} \frac{1}{x_i}}.$$

2-4-2-) **La moyenne harmonique pondérée:**

$$\frac{\Sigma_{i=1}^{k} n_i}{\Sigma_{i=1}^{k} \frac{n_i}{x_i}}.$$

Exemple On achète des dollars une première fois pour 100 € au cours de 1,23 € le dollar, une seconde fois pour 100 € au cours de 0,97 € le dollar.
Le cours moyen du dollar pour l'ensemble de ces deux opérations est égal à :

$$\frac{200}{\frac{100}{1.23} + \frac{100}{0.97}} \simeq 1.085.$$

La moyenne arithmétique (= 1,1) ne représente pas le cours moyen du dollar.

Exemple Dans un atelier, le coût horaire de la main d'œuvre est de 8 € (base 35 h par semaine). Une heure supplémentaire revient à 10 €, et le service de paie indique que le coût total des heures supplémentaires représente 30 % du coût total de la main d'œuvre. Calculez le coût horaire moyen et indiquez le type de moyenne utilisée.

Réponses Appelons x le coût total de la main d'œuvre :
coût horaire moyen $= \frac{coût\ \ total}{nombre\ \ total\ \ d'heures}$

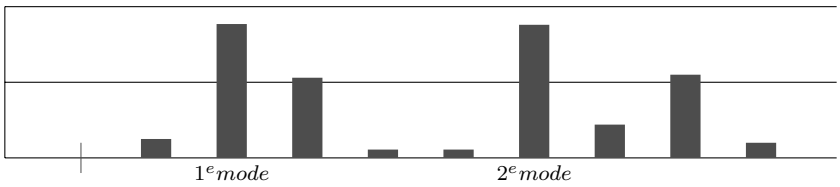
$$= \frac{x}{\frac{0.7x}{8} + \frac{0.3x}{10}} \simeq 8.51.$$

moyenne harmonique pondérée.

La moyenne harmonique peut être utilisée lorsqu'il est possible d'attribuer un sens réel aux inverses des données en particulier pour les taux de change, les taux d'équipement, le pouvoir d'achat, les vitesses. Elle est notamment utilisée dans les calculs d'indices.

3-**Le mode**

Le mode d'une série statistique est en principe la valeur la plus fréquente, *i.e.* la valeur de la variable pour laquelle l'effectif est maximal, classe correspondant au pic de l'histogramme pour une variable continue. Sa détermination est malaisée et dépend du découpage en classes.
Le mode, historiquement l'un des premiers paramètres de position utilisés, est un peu moins employé aujourd'hui.



Exemple de distribution bimodale d'une variable discrète

1.4.2 Paramètres de dispersion (Caractéristiques de dispersion)

Plus encore que la tendance centrale, la dispersion est la notion clé en statistique car si tous les individus avaient la même valeur il n' y aurait plus de raisonnement statistique

L'étendue ou intervalle de variation

Dépendante des valeurs extrêmes c'est un indicateur instable

$$w = x_{max} - x_{min}.$$

C'est-à-dire: L'étendue d'une série statistique est l'écart entre ses valeurs extrêmes.

L'intervalle interquartile

Indicateur de dispersion attaché à une variable aléatoire réelle.

Les quartiles Q_1, Q_2 et Q_3 étant définis par $F(Q_1) = 0.25$ $F(Q_2) = 0.5$ et $F(Q_3) = 0.75$.
$\mid Q_3 - Q_1 \mid$ est un indicateur parfois utilisé pour mesurer la dispersion: il est plus robuste que l'étendue.

La variance et l'écart-type

La variance s^2 est définie par :

$$s^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \overline{x})^2.$$

Et

$$s^2 = \frac{1}{n}\Sigma_{i=1}^{n}x_i^2 - \overline{x}^2,$$

la variance est le principal indicateur numérique de dispersion.
L'écart-type s s'exprime dans la même unité que la variable étudiée.

Les notations ne sont pas universelles.

2 Méthode des moindres carrés:

Si $((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n))$ est le nuage de points, et si $f = f_{a,b,\cdots}$ est la forme connue ou présumée de la relation mathématique, dépendant des paramètres $a, b, \cdots$, la méthode standard (due à Gauss et Legendre) s'appelle la méthode des moindres carrés (méthode qui minimise les écarts quadratiques entre modèle et observations): elle consiste à déterminer les valeurs des paramètres $a, b, \cdots$, comme celles qui minimisent la somme des carrés des écarts

$$\Sigma_{n=1}^{\infty} \left(y_i - f(x_i)\right)^2.$$

On cherche donc il ajuster au nuage des points (x_i, y_i) une droite d'équation $y^* = b + ax = f(x)$ de telle sorte que $\Sigma_{i=1}^{n} \left(y_i - y_i^* = f(x_i)\right)^2$ soit minimal. La méthode élémentaire de détermination de a et b est la suivante:

$$\Sigma_{i=1}^{n} \left(y_i - (b + ax_i)\right)^2 = F(a, b)$$
$$F(a, b) = \Sigma_{i=1}^{n} \left[(y_i - ax_i)^2 - 2b\left(y_i - ax_i\right) + b^2\right]$$
$$F(a, b) = \Sigma_{i=1}^{n} \left[(y_i - ax_i)^2 - 2b\left(y_i - ax_i\right)\right] + nb^2$$

donc

$$\frac{\partial}{\partial b} F(a, \hat{b}) = \Sigma_{i=1}^{n} \left[-2\left(y_i - ax_i\right)\right] + 2n\hat{b}$$

Ce minimum est atteint pour

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0.$$

Alors

$$\Sigma_{i=1}^{n} \left[-2\left(y_i - ax_i\right)\right] + 2n\hat{b} = 0$$
$$n\hat{b} = \Sigma_{i=1}^{n} \left(y_i - ax_i\right)$$
$$\hat{b} = \frac{1}{n} \Sigma_{i=1}^{n} \left(y_i - ax_i\right) \Rightarrow \hat{b} = \overline{y} - a\overline{x}.$$

Et

$$F(a, \hat{b}) = \Sigma_{i=1}^{n} \left(y_i - (\hat{b} + ax_i)\right)^2$$
$$F(a, \hat{b}) = \Sigma_{i=1}^{n} \left[y_i - (\overline{y} - a\overline{x} + ax_i)\right]^2$$
$$F(a, \hat{b}) = \Sigma_{i=1}^{n} \left[(y_i - \overline{y}) - a\left(x_i - \overline{x}\right)\right]^2$$
$$F(a, \hat{b}) = \Sigma_{i=1}^{n} \left(y_i - \overline{y}\right)^2 - 2a\Sigma_{i=1}^{n} \left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right) + a^2 \Sigma_{i=1}^{n} \left(x_i - \overline{x}\right)^2$$

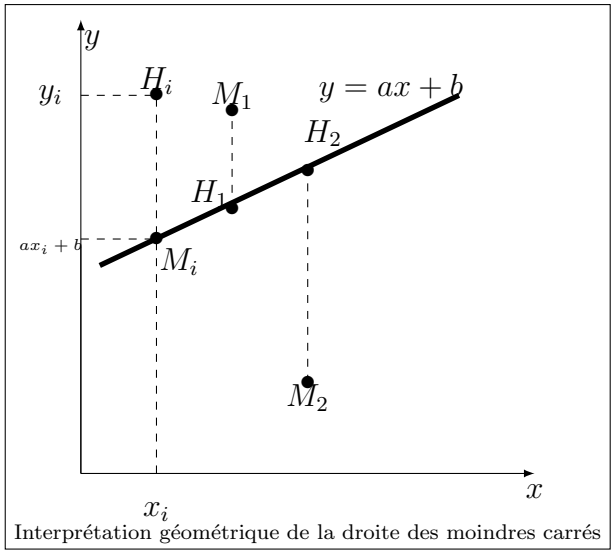
ce qui peut encore s'écrire :

$$F(a, \hat{b}) = n\left[var(Y) - 2acov(X, Y) + a^2 var(X)\right]$$

$$\frac{\partial}{\partial a}F(\hat{a},\hat{b}) = 0 \Rightarrow -2cov(X,Y) + 2\hat{a}var(X) = 0$$

$$\frac{\partial}{\partial a}F(\hat{a},\hat{b}) = 0 \Rightarrow \hat{a} = \frac{var(X)}{cov(X,Y),}$$

par conséquence la droite des moindres carrés passe par le point de coordonnées $(\overline{x},\overline{y})$ qu'on appelle parfois le centre de gravité ou point moyen du nuage.



Interprétation géométrique de la droite des moindres carrés

2.1 Droite de régression:

On s'intéresse à une statistique ayant deux dimensions que nous désignons par les variables X et Y. La notion de courbe de régression est un concept général qui va nous permettre de mettre en évidence au moyen d'un graphique s'il existe une relation entre ces deux variables et quelle est la nature de cette relation.
La courbe de régression est en fait un tracé que l'on fait passer entre les observations d'un nuage de points. Le plus souvent, on essaie de tracer une droite que l'on désigne alors par droite de régression ou, plus simplement par l'expression droite de tendance.
X Quoique la très grande majorité des relations réelles entre variables ne soient pas linéaires, c'est néanmoins l'ajustement linéaire qui est retenu dans de nombreux cas, pour trois raisons :
1) L'ajustement linéaire est beaucoup plus simple à traiter mathématiquement.
2) Beaucoup de relations sont approximativement linéaires si l'on prend un intervalle de variation suffisamment petit.

3) Certaines relations peuvent être rendues linéaires par un changement de variable approprié (généralement une transformation logarithmique).

2.1.1 Droite de régression linéaire:

Ce cas, le plus important dans la pratique, le point moyen est le point qui a pour coordonnées la moyenne de X et la moyenne de Y. On l'appelle aussi le centre de gravité.
La droite de régression est une droite qui passe par le point moyen. C'est aussi la droite qui minimise la somme des carrés des écarts des observations. Une fois connue, l'équation de cette droite permet de résumer la série et de faire des prévisions.

2.1.2 Utilité de la droite de régression

La droite de régression sert d'abord à vérifier l'existence d'une relation linéaire et la nature de celle-ci.
La droite de régression sert ensuite à faire des prévisions. Ainsi, nous pouvons utiliser l'équation de la droite de régression pour calculer des valeurs de Y associées à une valeur de X que l'on se donne.

2.2 Ajustement par des fonctions de puissances (courbe sigmoïde) :

Le problème se pose alors, connaissant un ensemble de points expérimentaux (un nuage , dans le vocabulaire statistique), de trouver la forme mathématique de la liaison et les bons coefficients (ou paramètres) numériques. Lorsque le phénomène sera déterministe perturbé aléatoirement, le nuage de points sera très voisin de la courbe y = f(x) que l'on cherche à déterminer. Lorsque le phénomène sera intrinsèquement aléatoire, le nuage sera plus dispersé et la courbe y = f(x) sera la courbe qui passera au mieux au milieu du nuage des points. La courbe ainsi déterminée, dans un cas comme dans l'autre, sera appelée courbe d'ajustement ou, dans le cas particulier affine, droite d'ajustement.
On considère deux grandeurs numériques x (*covariable*) et y intervenant dans l'étude d'un phénomène. On suppose qu'il existe une liaison entre ces deux grandeurs, et que cette liaison est aléatoire (qu'il s'agisse d'un phénomène déterministe perturbé par exemple par des erreurs de mesure, ou d'un phénomène intrinsèquement aléatoire). Le plus souvent, cette liaison sera représentée – ou sera présumée pouvoir être représentée – par une relation mathématique simple, par exemple : $y = bx^P$ (liaison puissance).

Les liaisons puissance est souvent ramenée aux cas linéaire par passage aux logarithmes : $y = bx^P$ devient $lny = lnb + plnx$.
Ce type d'ajustement convient à la description du cycle de vie de certains produits.

3 Statistiques paramétriques:

Soit X une v.a. associée à un certain phénomène aléatoire observable de façon répétée. Notre objectif est ≪d'estimer≫ certaines caractéristiques d'intérêt de sa loi (la moyenne, la variance, la fonction de répartition, la fonction de densité, etc.) sur la base d'une série d'observations.
nous souhaitons donner un ensemble de valeurs plausibles pour θ essentiellement sous forme d'un intervalle.

3.1 Intervalles de confiance:

3.1.1 Exemple introductif

Un industriel commande un lot de tiges métalliques qu'il ne peut utiliser que si leur longueur est comprise entre $23,60mm$ et $23,70mm$. Ces tiges ont été fabriquées par une machine qui, lorsqu'elle est réglée à la valeur m, produit des tiges dont la longueur peut être considérée comme une v.a. X de loi normale $N(m,\sigma)$, où l'écart type σ est une caractéristique de la machine, de valeur connue, ici $\sigma = 0,02mm$. Compte tenu de la symétrie de la distribution normale, la proportion de tiges utilisables par l'industriel sera maximale si le réglage a été effectué à $m_0 = 23,65mm$. Ne connaissant pas cette valeur, à la réception d'un lot de tiges l'industriel prélève au hasard n tiges dont il mesure les longueurs $X_1, \cdots, X_n$ pour se faire une idée de la valeur du paramètre de réglage m. Il calcule la moyenne des longueurs observées et ayant obtenu la valeur $\overline{x_n} = 23,63$ il en conclut que, s'il est peu réaliste de croire que la valeur de m est exactement $23,63mm$, elle doit malgré tout être très proche de cette valeur moyenne observée sur l'échantillon. Il lui paraît raisonnable d'aboutir à une conclusion de la forme ≪ il y a 95 chances sur 100 que la valeur de m soit comprise entre $23,63-a$ et $23,63+b$ ≫. Le problème consiste alors à fixer des valeurs précises pour a et b et on conçoit bien qu'elles doivent dépendre des ≪ chances ≫ que l'on a attribué à cet intervalle de contenir effectivement la vraie valeur de m. L'intervalle ainsi obtenu s'appellera intervalle de confiance et sa probabilité qui a permis de le déterminer, niveau de confiance. La longueur de cet intervalle sera bien sûr proportionnelle à ce niveau de confiance. On peut par exemple toujours fournir un intervalle qui contient avec certitude le paramètre en le choisissant suffisamment large ; mais dans ce cas, cet intervalle ne nous renseigne en aucune façon sur la vraie valeur du paramètre. Il faut donc arriver à un compromis entre un intervalle pas trop grand et une probabilité assez élevée de contenir le paramètre.
Pour une famille quelconque de lois de probabilité $(P_\theta; \theta \in \Theta)$ on peut donner

la définition suivante.

3.1.2 Définition

Un intervalle de confiance pour le paramètre θ , de niveau de confiance $1-\alpha \in]0,1[$, est un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur du paramètre θ.
Dans l'exemple précédent, nous avions abouti à un intervalle de la forme qui correspond à la réalisation d'un événement devant se

$$\overline{x_n} - a < m < \overline{x_n} + b$$

produire avec une probabilité fixée $1-\alpha$. La détermination des valeurs de a et b va donc se faire à partir de la valeur $1-\alpha$ de la probabilité, fixée par le statisticien, à partir de la condition qui s'écrit ici :

$$1-\alpha = P\{\overline{X}_n - a < m < \overline{X}_n + b\}$$

qui est équivalente à :

$$1-\alpha = P\{-b < m - \overline{X}_n < a\}.$$

On choisit $b = a$ et on utilise la variable centrée et réduite pour déterminer la valeur de a qui vérifie la condition :

$$1-\alpha = P\{-\frac{a}{\sigma/\sqrt{n}} < \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} < \frac{a}{\sigma/\sqrt{n}}\}.$$

Si Φ est la f.r. de la loi $\mathcal{N}(0,1)$, alors a est solution de :

$$1-\alpha = \Phi\left(\frac{a}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-a}{\sigma/\sqrt{n}}\right) = 2\Phi(\frac{a}{\sigma/\sqrt{n}}) - 1$$

ou

$$1-\frac{\alpha}{2} = \Phi\left(\frac{a}{\sigma/\sqrt{n}}\right),$$

soit

$$\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \frac{a}{\sigma/\sqrt{n}}.$$

Pour un niveau de confiance de $0,95$, soit $\alpha = 0.05$, et pour une taille d'échantillon $n = 100$, le fractile d'ordre $0,975$ de la loi $\mathcal{N}(0,1)$ a pour valeur $1,96$ et on en déduit $a = 0,004$, d'où l'intervalle :

$$23,626 < m < 23,634.$$

3.1.3 Principe

La méthode des intervalles de confiance est la suivante:
Soit T un estimateur (Un estimateur est une statistique, c'est-à-dire une variable aléatoire fonction d'un échantillon) de θ (on prendra évidemment le meilleur estimateur possible), dont on connait la loi de probabilité pour chaque valeur de θ.
Étant donné une valeur θ_0 de θ, on détermine un intervalle de probabilité de niveau $1-\alpha$ (coefficient de confiance) pour T, c'est -à-dire deux bornes t_1 et t_2 telles que :

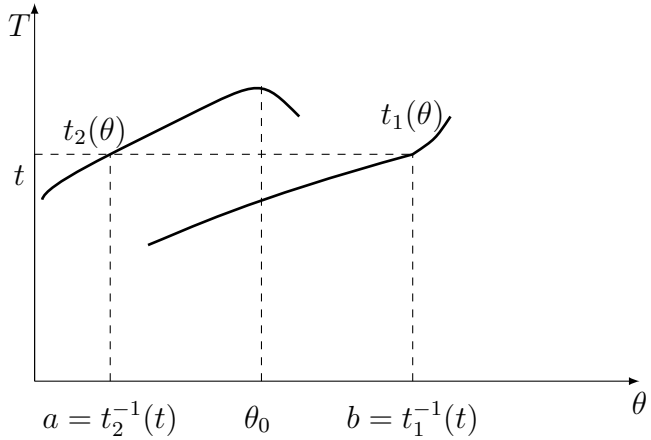
$$P(t_1 < T < t_2 \mid_{\theta=\theta_0}) = 1 - \alpha.$$

Ces bornes dépendent évidemment de θ_0.
On choisit dans la plupart des cas un intervalle de probabilité à risques symétriques $\frac{\alpha}{2}$ et $\frac{\alpha}{2}$.
On adopte alors la règle de décision suivante: soit t la valeur observée de T:

- si $t \in [t_1, t_2]$ on conserve θ_0 comme valeur possible de θ ;

- si $t \notin [t_1, t_2]$ on élimine θ_0,

On répète cette opération pour toutes les valeurs de θ.
On peut traduire graphiquement celte méthode dans un plan (θ, T) où l'on trace $t_2(\theta), t_1(\theta)$,



3.2 Test d'égalité des moyennes et d'égalité des variances de deux échantillons:

3.2.1 Test d'égalité des moyennes de deux échantillons

Descriptif du test On considère deux variables quantitatives X_1 et X_2 (qui mesurent la même caractéristique, mais dans deux populations différentes). On suppose que X_1 a pour moyenne théorique μ_1 et pour variance (σ_1^2 et X_2 a pour moyenne théorique μ_2 et pour variance σ_2^2. À partir des estimations calculées sur deux échantillons de tailles respectives n_1 et n_2 issus des deux populations, on veut comparer μ_1 et μ_2.
Les hypothèses du test sont

$$\mathcal{H}_0 : \mu_1 = \mu_2$$

contre

$$\mathcal{H}_1 : \mu_1 \neq \mu_2$$

$$\mu_1 < \mu_2$$

$$\mu_1 > \mu_2.$$

Sous $\mathcal{H}_0 \ll$ hypothèse zéro (ou nulle) $\gg$, la statistique de test est :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\backsim \mathcal{T}(n_1 + n_2 - 2)$$

avec ici

$$\hat{\sigma} = \frac{\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)}{n_1 + n_2 - 2},$$

$\hat{\sigma}_1$ et $\hat{\sigma}_2$ les estimateurs des variances des deux populations.

Conditions de validité: Normalité des variables X_1 et X_2 et variances égales.

Remarque La décision devient alors fortement dissymétrique : ou bien l'on rejette $\mathcal{H}_0$ (risque de première espèce ou risque d'erreur), ou bien l'on ne rejette pas $\mathcal{H}_0$ (risque de seconde espèce ou risque de manque de puissance).

3.2.2 Test d'égalité des variances de deux échantillons

Descriptif du test Ce test est souvent utile comme préalable à d 'autres tests comme celui de la comparaison de deux moyennes dans le cas de faibles effectifs. En effet , dans ce cas, la statistique n'est pas la même suivant que les variances de X_1 (variable concernant le premier échantillon) et de X_2 (variable concernant le second échantillon) peuvent être considérées comme égales ou non. Les hypothèses du test sont

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2$$

versus ((l' hypothèse l'alternative)

$$\mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2$$

$$\sigma_1^2 < \sigma_2^2$$

$$\sigma_1^2 > \sigma_2^2.$$

La statistique de test sous $\mathcal{H}_0$ est:

$$T = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2}$$

$$\backsim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Conditions de validité: Normalité de X_1 et X_2.

Remarque: On permute éventuellement les notations X_1 et X_2 pour que $\hat{\sigma}_1^2$ soit la plus grande des deux variances observées.

Exemple Supposons que deux machines A et B produisent le même type de produit, mais la machine A fournit un produit plus cher de qualité supérieure. La qualité d'un produit se mesure à une entité aléatoire qui est de loi $\mathcal{N}(5; 1)$ pour la machine A et $\mathcal{N}(4; 1)$ pour la machine B, et ne diffère donc que par la moyenne. Un client achète le produit le plus cher par lots de 10 et désire développer un test pour contrôler qu'un lot donné provient bien de la machine A. Comme accuser le producteur à tort peut avoir de graves conséquences, il doit limiter le risque correspondant et tester

$$\mathcal{H}_0 : \mu = 5$$

contre

$$\mathcal{H}_1 : \mu = 4$$

à un niveau 0,05 par exemple. Il semble naturel d'utiliser comme statistique de test la moyenne $\overline{X}$ du lot. Sous $\mathcal{H}_0$ sa loi est $\mathcal{N}(5; \frac{1}{10})$ et l'on a alors l'intervalle de probabilité 0,95 :

$$[5 - \frac{1,96}{\sqrt{10}}, 5 + \frac{1,96}{\sqrt{10}}],$$

soit

$$[4,38; 5,62].$$

D'où une règle de décision simple :

- accepter $\mathcal{H}_0$ si la réalisation $\overline{x}$ (moyenne du lot considéré) de $\overline{X}$ est dans
$$[4,38; 5,62],$$
- rejeter sinon.

Il est possible de calculer la puissance de ce test puisque la loi de $\overline{X}$ est connue sous $\mathcal{H}_1$: c'est la loi $\mathcal{N}(4; \frac{1}{10})$. Le risque de deuxième espèce vaut :

$$\beta = P(4,38 < \overline{X} < 5,62 \mid \mathcal{H}_1)$$
$$= P(\frac{4,38 - 4}{\frac{1}{\sqrt{10}}} < Z < \frac{5.62 - 4}{\frac{1}{\sqrt{10}}})$$

avec

$$Z \rightsquigarrow N(0; 1)$$
$$\beta = P(1,20 < Z < 5,12) \simeq 0,115.$$

D'où une puissance d'environ 0,885. Notons que l'on peut obtenir un test plus puissant en prenant comme région d'acceptation l'intervalle de probabilité 0,95:

$$[\frac{5 - 1,645}{\sqrt{10}}; +\infty[$$

où -1,645 est le quantile d'ordre 0,05 de la loi $\mathcal{N}(0; 1)$, soit

$$[4,48; +\infty[.$$

En effet :

$$\beta = P(4,38 < \overline{X} \mid \mathcal{H}_1)$$
$$\beta = P(\frac{4,38 - 4}{\frac{1}{\sqrt{10}}} < Z)$$
$$\beta = P(1,52 < Z) \simeq 0,064,$$

ce qui donne une puissance de 0,936. Intuitivement on sent bien que, dans le premier test, il est peu pertinent de borner la zone d'acceptation vers le haut car cela conduit à rejeter l'hypothèse nulle pour de très grandes valeurs de $\overline{x}$, au-delà de 5,62.

4 Tests non paramétriques:

On introduit souvent une distinction entre tests paramétriques et tests non paramétriques. Lorsque l'hypothèse $\mathcal{H}_0$ à tester nécessite une hypothèse préalable sur la loi de probabilité (ce sera le plus souvent une hypothèse de normalité) et implique des paramètres de cette loi, on dit que le test est paramétrique. Pour autant, la plupart des tests paramétriques sont robustes, i.e. supportent (dans des limites raisonnables que l'on peut préciser) que la loi réelle s'écarte de la loi nominale du test. Lorsque l'hypothèse $\mathcal{H}_0$ ne nécessite aucune hypothèse préalable sur la loi de probabilité, on dit que le test est non paramétrique.

4.1 Tests d'adéquation du khi-deux:

4.1.1 Tests d'adéquation

L'examen de la loi de probabilité empirique associée à un échantillon dont la loi parente est inconnue permet de choisir parmi les lois usuelles celle qui lui $\ll$ ressemble $\gg$ le plus. Si notre choix s'oriente vers une certaine loi P de fonction de répartition $(f.r.)F$, on pourra retenir l'hypothèse que l'échantillon provient de cette loi si la distance entre la $f.r.$ théorique F et la $f.r.$ empirique F_n est faible. Ayant fait le choix d'une certaine distance d entre fonctions de répartition, on se fixera une règle de décision qui s'énonce ainsi : $\ll$ Si l'événement $d(F_n, F) < C$ est réalisé, alors je retiens l'hypothèse qu'il s'agit d'un échantillon de la loi de $f.r.F \gg$. On peut cependant se tromper en rejetant cette hypothèse alors que F est bien la $f.r.$ des variables de l'échantillon; cette erreur se produit avec une probabilité qui est $\alpha = P\{d(F_n, F) > C\}$. Si on veut que ce risque d'erreur soit faible, on fixera une valeur α faible à cette probabilité (par exemple 5 ou 1) et cette valeur permettra alors de préciser la valeur de la constante C qui apparaît dans la règle de décision, si on connaît la loi de probabilité de la $v.a.$ $d(F_n, F)$. Nous aurons ainsi réalisé un test d'adéquation, ou d'ajustement, entre une loi théorique donnée et une loi empirique associée à un échantillon d'observations. La fixation du risque α déterminera alors la valeur du seuil d'acceptation, ou seuil critique C. Nous allons présenter maintenant deux tests, associés à deux distances entre $f.r.$, permettant de déterminer la loi approchée de la $v.a.$ $d(F_n, F)$ pour toute $f.r.F$, le premier étant plutôt destiné aux lois discrètes et le second réservé aux lois continues.

Test du khi-deux: Ce test est à retenir si les données sont discrètes, avec des valeurs possibles notées x_i , de probabilité p_i pour $1 \leq i \leq k$, ou

si les données individuelles ne sont pas fournies, mais ont été réparties en classes (a_i, a_{i+1}) dont les fréquences théoriques sont calculées à partir de la loi théorique postulée :

$$p_i = P\{X \in (a_i, a_{i+1})\} = F(a_{i+1}) - F(a_i)$$

Si N_i est le nombre (aléatoire) d'observations x_i , ou appartenant à la classe (a_i, a_{i+1}) , nous allons le comparer à l'effectif théorique qui est np_i . La distance euclidienne classique entre F_n , représentée par les k effectifs observés N_i , et la $f.r.F$, représentée par les k effectifs théoriques np_i , serait

$$\Sigma_{i=1}^{k}(N_i - np_i)^2.$$

Cependant, comme cette distance ne permet pas de déterminer la loi asymptotique de cette $v.a.$, on préfère retenir une autre distance. Cette dernière sera déterminée à partir de la remarque que les $v.a.N_i$ suivent des lois binômiales de paramètres n et p_i et que les variables centrées $\frac{(N_i - np_i)}{\sqrt{np_i}}$ convergent vers la loi $N(0, \sqrt{1-p_i})$. On retient donc la distance :

$$d(f_n, F) = \Sigma_{i=1}^{k}\frac{(N_i - np_i)^2}{np_i}$$

et cette somme de carrés de $v.a.$ centrées qui sont asymptotiquement normales et liées par la relation $\Sigma_{i=1}^{k}(N_i - np_i) = 0$ converge vers une loi χ^2_{k-1} . La valeur de C est alors déterminée approximativement, en utilisant cette loi asymptotique, comme le fractile d'ordre $1-\alpha$ de la loi du khi-deux à $1-k$ degrés de liberté. Cette approximation est justifiée si n est assez grand et p_i pas trop petit, avec comme règle empirique $np_i \geq 5$. Si ce n'est pas le cas à cause d'une valeur de pi trop petite on doit regrouper des classes (ou des valeurs) contiguës. Pour le calcul de la distance, il est préférable d'utiliser la formule développée :

$$d(F_n, F) = \Sigma_{i=1}^{k}\frac{N_i^2}{np_i} - n.$$

Les tests du khi-deux sont habituellement rangés parmi les tests non paramétriques, c'est clair si l'on considère le test du khi-deux d'indépendance (ou d'homogénéité), c'est moins évident si l'on considère le test du khi-deux d'ajustement $\cdots$.
Lorsque les caractères sont qualitatifs l'étude de la corrélation se fait par un test statistique développé par Karl PEARSONS et appelé test d'indépendance du Khi deux. Pour introduire ce test, considérons l'exemple suivant:

4.1.2 Exemple

On a observé pendant deux heures le nombre de voitures arrivées par minute à un poste de péage. Si X est la $v.a.$ représentant le nombre de voitures arrivant dans une minute à ce poste de péage, on fait l'hypothèse qu'elle suit une loi de Poisson. Le tableau ci-après contient les valeurs observées xi de cette variable et le nombre d'observations correspondantes N_i . Pour calculer les probabilités théoriques

$$p_i = P(X = x_i)$$

il faut spécifier complètement la loi, c'est-à-dire indiquer le paramètre de cette loi de Poisson. Le calcul de la moyenne empirique donne $\overline{x} = 3.7$ et la variance empirique a pour valeur 4.41. On retient alors la valeur entière 4 comme paramètre de cette loi de Poisson. Les valeurs de np_i sont alors obtenues par lecture de la table 4 et arrondies à l'entier le plus proche, en vérifiant bien que le total est égal à $n = 120$; par exemple

$$nP(X = 3) = 120 \times 0.1954 = 23.448$$

est arrondi à 23.

x_i	0	1	2	3	4	5	6	7	8	9	10	11
N_i	4	9	24	25	22	18	6	5	3	2	1	1
np_i	2	9	18	23	23	19	13	7	3	2	1	0

Les valeurs des effectifs théoriques inférieures à 5 nécessitent de regrouper les deux premières et les quatre dernières valeurs, ramenant à 8 le nombre de valeurs retenues, soit $8 - 1 - 1 = 6$ degrés de liberté puisqu'un paramètre a été estimé. Le fractile d'ordre 0.95 de la loi du khi-deux à 6 degrés de liberté est $C = 12.6$. On obtient

$$d(F_n, F) = 7.14$$

ce qui conduit donc à accepter l'hypothèse que X suit une loi de Poisson de paramètre 4.

4.1.3 Exemple:

100 consommateurs sont questionnés sur leurs préférences à l'égard de 4 variétés d'un produit (A, B, C et D). On leur demande : Parmi ces 4 produits, quel est celui que vous préférez . Ces consommateurs sont groupés en deux catégories, les moins de 20 ans et les plus de 20 ans, afin de déterminer si l'âge a une influence sur la préférence.

Produits	Moins de 20 ans	Plus de 20 ans	Total
A	10	15	25
B	10	25	35
C	15	5	20
D	20	0	20
Total	55	45	100

Si l'âge n'a aucune influence sur le choix, les 2 premières colonnes devraient être proportionnelles à la troisième. On va donc calculer deux colonnes fictives, mais proportionnelles à la troisième, afin d'avoir les effectifs qui correspondent à une indépendance de l'âge sur le choix.
Dans la formule ci-après, la fréquence des plus de 20 ans est $\frac{45}{100}$. Celle des moins de 20 ans : est $\frac{55}{100}$.
N_i est l'effectif théorique correspondant à une répartition homogène. Enfin, n_i est l'effectif observé.

-20 ans

Produits	N_i	n_i	$N_i - n_i$	$(N_i - n_i)^2$	$\frac{(N_i - n_i)^2}{N_i}$
A	$\frac{55}{100} \times 25$ =13.75	10	3.75	14.0625	1.02272
B	19.25	10	9.25	85.5625	4.448
C	11	15	-4	16	1.4545
D	11	20	-9	81	7.3636

+20 ans

Produits	N_i	n_i	$N_i - n_i$	$(N_i - n_i)^2$	$\frac{(N_i - n_i)^2}{N_i}$
A	$\frac{45}{100} \times 25 = 11.25$	15	-3.75	14.0625	1.25
B	15.75	25	-9.25	85.5625	5.4325
C	9	5	4	16	1.7777
D	9	0	9	81	9

Par définition :

$$\chi^2 = \Sigma_{i=1}^{n} \frac{(N_i - n_i)^2}{N_i}$$

En appliquant cette définition aux données du tableau , on obtient :

$$\chi^2 = 31.74$$

Une fois que l'on connaît le khi-carré calculé, on doit le comparer avec la valeur du khi deux issue de la distribution du khi-carré (voir le tableau ci-dessous). Ici, le nombre de $\ll$ degrés de liberté $\gg$ est égal à [8 (nombre d'observations) moins 2 (nombres de variables)], ce qui donne 6. Ensuite, nous devons choisir la probabilité de fiabilité du test : 5% de chances de se

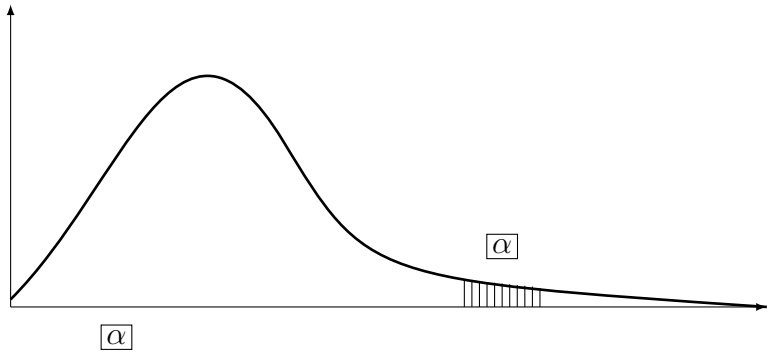
tromper (deuxième colonne), 1% (troisième colonne) et 1 pour 1000 (quatrième colonne). Si nous choisissons $P = 0.05$, nous avons donc :

$$\chi^2_{0.05} = 12.59 < 31.74$$

Ce qui nous permet de conclure que la répartition des préférences est suffisamment différente d'une répartition homogène pour qu'on puisse raisonnablement se fier à l'idée que l'âge a une influence sur le choix du produit (avec 5% de chances de nous tromper).

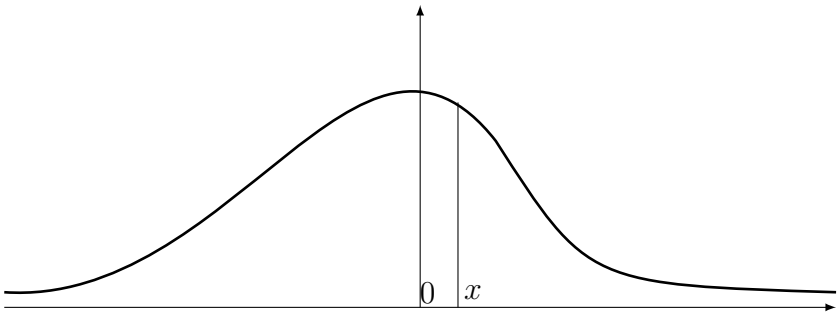
degrés de liberté	P=0.05	P=0.01	P=0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	**12.59**	**16.81**	**22.46**
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.19

4.1.4 Loi du khi-deux (loi dérivée de la loi normale)



ν	0.995	0.975	0.95	0.9	0.1	0.05	0.025	0.005
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	7,879
2	0,010	0,051	0,103	0,211	4,605	5,991	7,378	10,597
3	0,072	0,216	0,352	0,584	6,251	7,815	9,348	12,838
4	0,207	0,484	0,711	1,064	7,779	9,488	11,143	14,860
5	0,412	0,831	1,145	1,610	9,236	11,070	12,832	16,750
6	0,676	1,237	1,635	2,204	10,645	12,592	14,449	18,548
7	0,989	1,690	2,167	2,833	12,017	14,067	16,013	20,278
8	1,344	2,180	2,733	3,490	13,362	15,507	17,535	21,955
9	1,735	2,700	3,325	4,168	14,684	16,919	19,023	23,589
10	2,156	3,247	3,940	4,865	15,987	18,307	20,483	25,188
11	2,603	3,816	4,575	5,578	17,275	19,675	21,920	26,757
12	3,074	4,404	5,226	6,304	18,549	21,026	23,337	28,300
13	3,565	5,009	5,892	7,041	19,812	22,362	24,736	29,819
14	4,075	5,629	6,571	7,790	21,064	23,685	26,119	31,319
15	4,601	6,262	7,261	8,547	22,307	24,996	27,488	32,801
16	5,142	6,908	7,962	9,312	23,542	26,296	28,845	34,267
17	5,697	7,564	8,672	10,085	24,769	27,587	30,191	35,718
18	6,265	8,231	9,390	10,865	25,989	28,869	31,526	37,156
19	6,844	8,907	10,117	11,651	27,204	30,144	32,852	38,582
20	7,434	9,591	10,851	12,443	28,412	31,410	34,170	39,997
21	8,034	10,283	11,591	13,240	29,615	32,671	35,479	41,401
22	8,643	10,982	12,338	14,041	30,813	33,924	36,781	42,796
23	9,260	11,689	13,091	14,848	32,007	35,172	38,076	44,181
24	9,886	12,401	13,848	15,659	33,196	36,415	39,364	45,558
25	10,520	13,120	14,611	16,473	34,382	37,652	40,646	46,928
26	11,160	13,844	15,379	17,292	35,563	38,885	41,923	48,290
27	11,808	14,573	16,151	18,114	36,741	40,113	43,195	49,645
28	12,461	15,308	16,928	18,939	37,916	41,337	44,461	50,994
29	13,121	16,047	17,708	19,768	39,087	42,557	45,722	52,335
30	13,787	16,791	18,493	20,599	40,256	43,773	46,979	53,672
40	20,707	24,433	26,509	29,051	51,805	55,758	59,342	66,766
50	27,991	32,357	34,764	37,689	63,167	67,505	71,420	79,490
60	35,534	40,482	43,188	46,459	74,397	79,082	83,298	91,952
70	43,275	48,758	51,739	55,329	85,527	90,531	95,023	104,215
80	51,172	57,153	60,391	64,278	96,578	101,879	106,629	116,321
90	59,196	65,647	69,126	73,291	107,565	113,145	118,136	128,299
100	67,328	74,222	77,929	82,358	118,498	124,342	129,561	140,170

4.1.5 La loi normale centrée réduite



$$\boxed{\Pi(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt}$$

x	0,00	0,01	0,02	0,03	0,04
0, 0	0.5000	0,5040	0,5080	0,5120	0,5160
0,1	0,5398	0,5438	0,5478	0,5517	0,5557
0,2	0,5793	0,5832	0,5871	0.5910	0,5948
0,3	0,6179	0,6217	0,6255	0,6293	0,6331
0.4	0.6554	0,6591	0,6628	0,6664	0,6700
0.5	0,6915	0.6950	0,6985	0,7019	0,7054
0,6	0.7257	0,7290	0,7324	0,7357	0,7389
0.7	0,7580	0,7611	0,7642	0,7673	0,7704
0,8	0,7881	0.7910	0,7939	0,7967	0,7995
0,9	0,8159	0,8186	0,8212	0,8238	0,8264
1,0	0,8413	0,8438	0,8461	0,8485	0,8508
1.1	0,8643	0,8665	0,8686	0,8708	0,8729
1,2	0,8849	0.8869	0.8888	0,8907	0,8925
1.3	0,9032	0,9049	0,9066	0,9082	0,9099
1.4	0,9192	0,9207	0.9222	0,9236	0,9251
1.5	0,9332	0.9345	0,9357	0.9370	0,9382
1.6	0,9452	0,9463	0,9474	0,9484	0,9495
1.7	0.9554	0,9564	0,9573	0.9582	0,9591
1.8	0,9641	0,9649	0.9656	0,9664	0,9671
1.9	0,9713	0,9719	0,9726	0,9732	0,9738

x	0,05	0,06	0,07	0,08	0,09
0, 0	0,5199	0,5239	0.5279	0,5319	0.5359
0,1	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5987	0.6026	0,6064	0,6103	0,6141
0,3	0,6368	0,6406	0,6443	0,6480	0,6517
0.4	0,6736	0,6772	0,6808	0,6844	0,6879
0.5	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7422	0,7454	0.7486	0,7517	0,7549
0.7	0.7734	0,7764	0,7794	0,7823	0,7852
0,8	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8531	0.8554	0.8577	0,8599	0,8621
1.1	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8944	0,8962	0,8980	0,8997	0,9015
1.3	0,9115	0,9131	0,9147	0,9162	0,9177
1.4	0,9265	0.9279	0.9292	0.9306	0,9319
1.5	0,9394	0,9406	0,9418	0,9429	0,9441
1.6	0,9505	0,9515	0,9525	0,9535	0,9545
1.7	0,9599	0,9608	0.9616	0.9625	0,9633
1.8	0,9678	0,9686	0,9693	0,9699	0,9706
1.9	0,9744	0,9750	0,9756	0,9761	0,9767

x	0,00	0,01	0,02	0,03	0,04
2	0,9772	0.9779	0,9783	0,9788	0,9793
2.1	0.9821	0,9826	0,9830	0.9834	0,9838
2.2	0.9861	0,9864	0,9868	0,9871	0,9875
2.3	0.9893	0,9896	0,9898	0.9901	0,9904
2.4	0.9918	0,9920	0,9922	0,9925	0,9927
2.5	0,9938	0,9940	0,9941	0,9943	0,9945
2.6	0,9953	0.9955	0,9956	0,9957	0.9959
2.7	0,9965	0,9966	0,9967	0.9968	0,9969
2.8	0,9974	0.9975	0,9976	0,9977	0,9977
2.9	0,9981	0,9982	0,9982	0,9983	0,9984

x	0,05	0,06	0,07	0,08	0,09
2	0,9798	0,9803	0,9808	0,9812	0,9817
2.1	0,9842	0,9846	0.9850	0,9854	0,9857
2.2	0,9878	0,9881	0,9884	0,9887	0,9890
2.3	0,9906	0.9909	0.9911	0.9913	0.9916
2.4	0,9929	0,9931	0,9932	0,9934	0,9936
2.5	0,9946	0,9948	0,9949	0,9951	0.9952
2.6	0,9960	0,9961	0,9962	0,9963	0,9964
2.7	0,9970	0,9971	0,9972	0,9973	0.9974
2.8	0.9978	0,9979	0,9979	0,9980	0,9981
2.9	0,9984	0,9985	0,9985	0,9986	0,9986

4.2 Test de comparaison de deux échantillons indépendants:

4.2.1 Test de Wilcoxon

Ce test repose sur l'idée que si l'on mélange les deux séries de valeurs et qu'on ordonne le tout par valeurs croissantes on doit obtenir un mélange homogène. Test d'hypothèse non paramétrique utilisé pour comparer les distributions de deux échantillons statistiques. Aussi appelé $\ll$ test de la somme des rangs $\gg$, il fonctionne, non pas à partir des valeurs précises observées, mais à partir des rangs de ces valeurs interclassées. Si les variables aléatoires X et Y dont proviennent respectivement les deux échantillons ont même loi, elles ont en particulier même espérance mathématique, et c'est très souvent comme test de l'hypothèse dérivée

$$\mu_X = \mu_Y$$

(deux espérances mathématiques) que le test de **Wilcoxon** est utilisé. L'hypothèse (réellement testée)

$$\mathcal{H}_0 : X \textit{ et } Y \textit{ ont même loi}$$

a pour conséquence immédiate la symétrie

$$P(X \leq Y) = P(X \geq Y)$$

(si les lois sont continues, on a par surcroît

$$P(X = Y) = 0,$$

et donc

$$P(X \leq Y) = P(X \geq Y) = \frac{1}{2}.$$

La mise en œuvre du test de Wilcoxon est une simple exploitation de cette égalité des probabilités symétriques.

- Données. Deux séries :

$$(x_1, x_2, \cdots, x_{n_X}), \qquad (y_1, y_2, \cdots, y_{n_Y}).$$

- Hypothèse réellement testée.

$$\mathcal{H}_0 : X \textit{ et } Y \textit{ ont même loi}$$

contre $\mathcal{H}_1$ alternative.

- Hypothèse dérivée.

$$\mathcal{H}_0 : \mu_X = \mu_Y$$

contre

$$\mathcal{H}_1 : \mu_X \neq \mu_Y.$$

- Déroulement technique du test

1. On classe les
$$n_X + n_Y$$
valeurs observées par ordre croissant.
2. On calcule la somme W_X des rangs des valeurs de la variable X (s'il y a des ex æquo, on leur attribue le rang moyen).
3. On calcule la valeur observée de la variable de test :
$$W = \frac{\mid W_X - \frac{n_X(n_X+n_Y+1)}{2} \mid}{\sqrt{\frac{n_Y n_X(n_X+n_Y+1)}{12}}}$$

Les valeurs de référence de la variable de test sont à lire, soit dans des tables spécifiques pour les petites valeurs de n_X et n_Y , soit dans la table de la loi normale (centrée réduite), pour le risque bilatéral α.

- Conditions et précautions:

- Il n'y a aucune condition sur la loi commune à X et Y ;
- par contre, la loi normale (centrée réduite) est la loi limite pour la variable de test, ce qui induit une condition de taille si l'on ne dispose pas de table spécifique ; il est classique de demander

$$n_X \textit{ et } n_Y \geq 10$$

pour pouvoir se référer à la table de la loi normale.

4.2.2 Exemple:

On veut comparer les performances de deux groupes d'élèves à des tests d'habileté manuelle.
On choisit aléatoirement 8 individus du premier groupe et 10 du deuxième. Les performances en minutes sont les suivantes:

Groupe 1: 22	31	14	19	24	28	27	28		
Groupe 2: 25	13	20	11	23	16	21	18	17	26

On réordonne les 18 observations par ordre croissant. Les résultats du premier groupe sont soulignés:

Observations: 11	13	14	16	17	18	<u>19</u>	20	21
Rangs: 1	2	3	4	5	6	7	8	9

22	23	24	25	26	<u>27</u>	<u>28</u>	<u>28</u>	<u>31</u>
10	11	12	13	14	15	16	17	18

La somme des rangs des individus du premier groupe est:

$$W_X = 3 + 7 + 10 + 12 + 15 + 16 + 17 + 18 = 98$$

Si $\mathcal{H}_0$ était vraie:

$$\frac{n_X(n_X + n_Y + 1)}{2} = \frac{8(8 + 10 + 1)}{2} = 76,$$

$$\frac{n_X n_Y(n_X + n_Y + 1)}{12} = \frac{8 \times 10(8 + 10 + 1)}{12} = 126.7 = (11.25)^2.$$

Comme

$$\frac{98 - 76}{11.25} = 1.76,$$

on peut rejeter $\mathcal{H}_0$ avec $\alpha = 0.10$ et conclure à une plus grande 11.25 rapidité des élèves du groupe 2.

5 Probabilités:

La notion essentielle introduite étant bien sûr celle de probabilité, avec la notion d'indépendance d'événements qui lui est associée et qui joue un rôle très important en statistique. La représentation formelle du modèle probabiliste sous jacent est presque toujours absente dans un problème concret de statistique. Cependant, cette formalisation rigoureuse est indispensable pour obtenir les outils théoriques nécessaires à la résolution d'un tel problème statistique.

5.1 Vocabulaire de base:

5.1.1 Ensemble fondamental

Le résultat d'une expérience aléatoire s'appelle **événement**, La quantification des $\ll$ chances $\gg$ qu'un tel événement a de se réaliser correspond à la notion intuitive de probabilité. Pour réaliser cette quantification, il est nécessaire de décrire au préalable, très précisément, l'ensemble des résultats possibles, appelés événements élémentaires. Cet ensemble expérimental s'appelle **ensemble fondamental** (ou univers) et est noté traditionnellement Ω.
Chaque élément $\omega \in \Omega$ représente donc un événement élémentaire, et toute partie $A \subset \Omega$ (ou $A \in \mathcal{P}(\Omega)$) sera un événement. Parfois on dit que Ω est l'ensemble des éventualités possibles et les événements élémentaires sont alors les singletons, c'est-à-dire les ensembles réduits à un seul élément ω , qui sont effectivement en toute rigueur des événements, puisque appartenant à $\mathcal{P}(\Omega)$, ce qui n'est pas le cas du point ω.
L'ensemble Ω dépend évidemment de l'expérience considérée, mais aussi du choix de celui qui construit le modèle, et par là présente donc un certain arbitraire.
Cet ensemble Ω peut être fini ou infini, continu ou discret.

5.2 Probabilités élémentaires:

5.2.1 Définition

On appelle probabilité P sur $(\Omega, \mathcal{A})$ une application $P : \mathcal{A} \to [0, 1]$ telle que :

-) On associera à une épreuve aléatoire un ensemble non vide de parties de Ω, noté $\mathcal{A}$, qui vérifiera :

i) pour tout $A \in \mathcal{A}$ alors $\overline{A} \in \mathcal{A}$;

ii) pour tout $A \in \mathcal{A}$ et tout $B \in \mathcal{A}$alors $A \cup B \in \mathcal{A}$.

$$\Updownarrow$$

ii′) pour tout $A \in \mathcal{A}$ et tout $B \in \mathcal{A}$alors $A \cap B \in \mathcal{A}$.

-) $P(\Omega) = 1$;

-) pour toute suite A_n d'événements incompatibles, soit soit $A_n \in \mathcal{A}$ avec $A_m \cap A_n = \phi$ pour $m \neq n$:

$$P\left(\cup_{n=0}^{\infty} A_n\right) = \Sigma_{n=0}^{\infty} P(A_n).$$

Une probabilité est donc une application qui à un événement va associer un nombre. Le triplet $(\Omega, \mathcal{A}, P)$ s'appelle un espace probabilisé. Comme conséquences de la définition on déduit les propriétés suivantes.

P_1) $P(\phi) = 0$.

P_2) $P(\overline{A}) = 1 - P(A)$.

P_3) $A \subset B \Rightarrow P(A) \leq P(B)$.

P_4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

5.2.2 Cas où Ω est fini

$$P(A) = \Sigma\{p_i / \omega_i \in A\}$$

.

5.3 Probabilités conditionnelles:

On considère l'espace probabilisé $(\Omega, \mathcal{A}, P)$ et un événement particulier B de $\mathcal{A}$ tel que $P(B) > 0$. La connaissance de la réalisation de B modifie la probabilité de réalisation d'un événement élémentaire, puisque l'ensemble des résultats possibles est devenu B et non plus Ω.
Nous avons évoqué en introduction le lien particulier entre l'information disponible, le contrôle des facteurs déterminants d'un phénomène et l'importance de sa partie aléatoire, donc de sa probabilité de réalisation. Nous allons retrouver ceci au travers de la notion de probabilité conditionnelle.
La connaissance d'une information complémentaire sur le déroulement de l'épreuve équivaut à la modification des probabilités définies sur les éléments de $\mathcal{A}$. En effet, cette information acquise n'est autre qu'une condition désormais

supposée réalisée quel que soit le résultat de l'expérience aléatoire.
On remarque que l'ensemble fondamental a été modifié sous des conditions, ainsi que la mesure de probabilité P. Cette modification s'appelle un conditionnement, car elle correspond à la prise en compte d'une condition supplémentaire sur la réalisation de l'épreuve aléatoire. On est ainsi conduit à définir les probabilités conditionnelles

5.3.1 Définition

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et soit $C \in \mathcal{A}$ un événement particulier, appelé condition, de probabilité non nulle. Pour tout événement $A \in \mathcal{A}$, on appelle probabilité conditionnelle de A sachant C, notée $P(A \mid C)$, la quantité :

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)}.$$

5.3.2 Définition

$$A \text{ et } C \text{ indépendants}$$

$$\Updownarrow$$

$$P(A \cap C) = P(A) \times P(C)$$

Notons encore que si A et C sont deux événements indépendants, alors :

$$P(A \mid C) = P(A \mid \overline{C}) = P(A)$$

$$P(C \mid A) = P(C \mid \overline{A}) = P(C)$$

Nous donnerons donc d'abord le résultat connu sous le nom de théorème de Bayes, pour examiner ensuite le débat sur la notion de probabilité.

5.3.3 Théorème de Bayes

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé, et soient $A_1, A_2, \cdots, A_n$ un ensemble d'événements deux à deux incompatibles vérifiant

$$\cup_{k=1}^{n} A_k = \Omega$$

(on dit que les A_k forment un système complet d'événements). Pour tout événement C, on a alors :

$$P(A_i \mid C) = \frac{P(C \mid A_i) \times P(A_i)}{\Sigma_{k=1}^{n} P(C \mid A_i) \times P(A_i)}$$

pour $i = 1, 2, \cdots, n$

Exemple Pour un système de crédit à la clientèle on distingue trois types de dossiers : les dossiers aboutissant en contentieux, les dossiers à difficultés temporaires ou légères et les dossiers sans difficultés de paiement. On a évalué sur la base d'expériences antérieures les proportions respectives des trois catégories à $\frac{1}{5}$, $\frac{3}{10}$ et $\frac{1}{2}$. D'autre part, on dispose pour chaque dossier d'un score d'appréciation global du client rapporté à l'une des deux modalités suivantes : élevé ou bas. Enfin, on sait que 90% des dossiers en contentieux correspondaient à un score bas, que 60% des dossiers à difficultés légères correspondaient à un score bas, et que 85% des dossiers sans difficultés correspondaient à un score élevé. Si on tire un dossier au hasard pour lequel le score est bas, quelle est la probabilité qu'il ait abouti en contentieux ? (resp. qu'il n'ait donné lieu à aucune difficulté de paiement ? qu'il ait engendré des difficultés légères ?) Les trois événements $A_1 =\ll$ aboutir en contentieux $\gg$, $A_2 =\ll$ difficultés légères $\gg$ et $A_3 =\ll$ aucune difficulté $\gg$ forment un système complet. On dispose des probabilités a priori:

$$P(A_1) = 0.2 \quad P(A_2) = 0.3 \quad P(A_3) = 0.5$$

ainsi que des probabilités conditionnelles pour les événements
$C =\ll$ score bas $\gg$ et $\overline{C} =\ll$ score élevé $\gg$

$$P(C \mid A_1) = 0.9 \quad P(C \mid A_2) = 0.6 \quad P(C \mid A_3) = 0.15$$

d'où :

$$P(C) = P(C \cap A_1) + P(C \cap A_2) + P(C \cap A_3)$$
$$= P(C \mid A_1) \times P(A_1) + P(C \mid A_2) \times P(A_2) + P(C \mid A_3) \times P(A_3)$$
$$= 0.435$$

On en déduit :

$$P(A_1 \mid C) = \frac{P(A_1 \cap C)}{P(C)} = \frac{P(C \mid A_1) \times P(A_1)}{P(C)} = 0.414$$

ainsi que :

$$P(A_2 \mid C) = 0.414 \quad P(A_3 \mid C) = 0.172$$

5.4 Variables aléatoires discrètes:

Le concept de variable aléatoire formalise la notion de grandeur variant selon le résultat d'une expérience aléatoire.

5.4.1 Variables aléatoires à une dimension

Définition Étant donné un espace probabilisé $(\Omega, \mathcal{A}, P)$, une variable aléatoire ($v.a.$ en abrégé) est une application X définie sur l'ensemble fondamental Ω et à valeurs réelles :

$$X : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto X(\omega)$$

on définit $P_X(B)$ par:

$$P_X(B) = P\left([\omega \mid X(\omega) \in B]\right)$$
$$= P\left(\left[X^{-1}(B)\right]\right)$$

Exemple On jette deux dés non pipés ; l'ensemble fondamental associé à cette expérience aléatoire est formé de 36 événements élémentaires équiprobables:

$$\Omega = (\{1,1\}; \{1,2\}; \{2,1\}; \cdots ; \{6,6\})$$

Si on s'intéresse à la somme des points marqués par les deux dés, on définira sur cet espace probabilisé une $v.a.X$ égale à cette somme ; l'ensemble de ses valeurs possibles est :

$$\{2,3,4,5,6,7,8,9,10,11,12\}$$

Pour obtenir la probabilité d'une valeur quelconque de X, il suffit de dénombrer les événements élémentaires de Ω qui réalisent cette valeur; ainsi :

$$P(X=4) = P\left(\{1,3\} \cup \{2,2\} \cup \{3,1\}\right)$$
$$= P(\{1,3\}) + P(\{2,2\}) + P(\{3,1\}) = \frac{3}{36} = \frac{1}{12}$$

Définition On dit que la variable aléatoire X est :

- discrète finie si l'ensemble $X(\Omega)$ est fini, discrète infinie si l'ensemble $X(\Omega)$ est infini dénombrable,
- continue si l'ensemble $X(\Omega)$ est un intervalle de $\mathbb{R}$ non réduit à un point (ou une réunion d'intervalles de $\mathbb{R}$).

Définition On appelle $v.a.$ discrète définie sur $(\Omega, \mathcal{A})$ une application

$$X : \Omega \longrightarrow \mathbb{R}$$

telle que $X(\Omega)$ est dénombrable (en général $X(\Omega)$ est fini ou $X(\Omega) \subset \mathbb{N}$ ou $X(\Omega) \subset \mathbb{Z}$ et dans tous les cas $X(\Omega)$ est en correspondance bijective avec $\mathbb{N}$) et telle que pour tout x réel :

$$X^{-1}(x) = \{\omega \in \Omega / X(\omega) = x\} \in \mathcal{A}$$

ce qui exprime tout simplement que $X^{-1}(x)$ est un événement.

Cas particulier

Variable indicatrice: Soit $A \in \mathcal{A}$ un événement quelconque ; on appelle $v.a.$ **indicatrice** de cet événement A, la $v.a.$ définie par :

$$X(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in A \\ 0 & \text{si} \quad \omega \in \overline{A} \end{cases}$$

et notée $X = \mathbf{1}_A$. Ainsi :

$$P_X(X = 1) = P\{\omega / \omega \in A\} = P(A)$$

$$P_X(X = 0) = P\{\omega / \omega \in \overline{A}\} = 1 - P(A).$$

5.4.2 Fonction de répartition

La fonction de répartition est l'instrument de référence pour définir de façon unifiée la loi de probabilité d'une variable aléatoire qu'elle soit discrète ou continue. Si cette fonction est connue, il est possible de calculer la probabilité de tout intervalle et donc, en pratique, de tout événement. C'est pourquoi c'est elle qui est donnée dans les tables des lois de probabilité.

Définition On appelle fonction de répartition de la $v.a.X$, la fonction F définie pour x réel par :

$$F(x) = P_X\{X < x\} = P\{\omega \in \Omega / X(\omega) < x\}.$$

Si par exemple X prend les valeurs

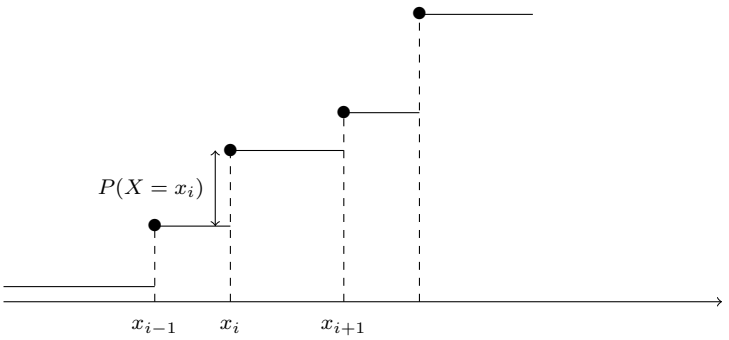
$$x_1 < x_2 < \cdots < x_n,$$

on aura

$$F(x) = 0$$

pour $x \leq x_1$, puis le graphe de F présentera un saut en chaque point x_i , jusqu'à la valeur

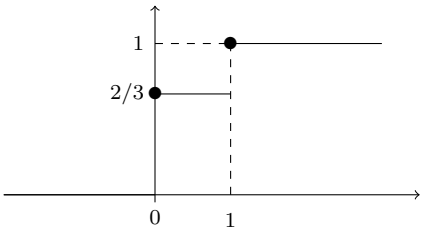
$$F(x) = 1 \quad pour \quad x > x_n.$$



Exemple On considère la loi de Bernouilli de paramètre 1/3 :

$$P(X = 1) = \frac{1}{3}, \quad P(X = 0) = 2/3$$

La représentation graphique de sa fonction de répartition est



On peut déduire de F les probabilités individuelles par:

$$p_i = F(x_{i+1}) - F(x_i) \quad pour \quad 1 \leq i \leq n-1$$

et

$$p_n = 1 - F(x_n).$$

5.4.3 Moments d'une v.a. discrète

Espérance mathématique

Définition On appelle espérance mathématique de la $v.a.X$ la quantité, si elle existe :

$$E(X) = \Sigma_{i \in \mathbb{N}} p_i x_i$$

Exemple Pour une $v.a.$ **indicatrice** :

$$E(X) = 0 \times P_X(X = 0) + 1 \times P_X(X = 1) = P(A) = p.$$

Remarque À l'origine des probabilités, la quantité $E(X)$ a été introduite pour traduire la notion de gain moyen, ou espérance de gain, la $v.a.X$ représentant la valeur du gain à un certain jeu.

Propriétés Les propriétés de l'opérateur espérance mathématique sont celles du signe somme.

$$1) \quad E(X + a) = E(X) + a \quad 2) \quad E(aX) = aE(X), \quad a \in \mathbb{R}$$

$$3) \quad E(X+Y) = E(X)+E(Y) \quad 4) \quad E(\lambda X+\mu Y) = \lambda E(X)+\mu E(Y), \quad \lambda, \mu \in \mathbb{R}$$

Remarque Si g est une fonction continue quelconque, alors :

$$E[g(X)] = \Sigma_{i\in\mathbb{N}} p_i g(x_i).$$

Variance Il s'agit d'un indicateur mesurant la dispersion des valeurs xi que peut prendre la $v.a.X$, autour de la moyenne en probabilité $E(X)$ et défini par:

$$\begin{aligned} V(X) &= \Sigma_{i\in\mathbb{N}} p_i \left(x_i - E(X)\right)^2 \\ &= E\left[X - E(X)\right]^2 = \sigma_X^2, \end{aligned}$$

σ_X^2 désignant alors l'écart type de X qui s'exprime dans les mêmes unités de mesure que la variable.

Exemple Pour une $v.a.$ **indicatrice** :

$$V(X) = E(X - p)^2 = p \times (1 - p)^2 + (1 - p) \times (0 - p)^2 = p(1 - p).$$

Cet indicateur de dispersion vient compléter l'information sur la distribution, fournie par la valeur moyenne.

Exemple Considérons les deux distributions suivantes :

X	2	4	6
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

Y	-4	3	33
	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Elles ont comme valeurs moyennes :

$$E(X) = \frac{1}{2} + 1 + 3 = 4.5 \quad et \quad E(Y) = -2 + 1 + \frac{11}{2} = 4.5$$

donc même centre de distribution. Par contre :

$$E(X^2) = 1 + 4 + 18 = 23 \quad et \quad E(Y^2) = 8 + 3 + 121 \times \times \frac{3}{2} = \frac{385}{2}$$

d'où

$$V(X) = \frac{11}{4} \quad et \quad V(Y) = \frac{689}{4},$$

valeur très supérieure qui indique une dispersion de Y autour de sa moyenne beaucoup plus grande que celle de $X \neq$.

Propriétés

$P_1)$ $V(X) \geq 0$ $P_2)$ $V(a+X) = V(X)$ P_3 $V(a \times X) = a^2 V(X)$, $a \in \mathbb{R}$

$$P_4) \quad V(X) = E(X^2) - E^2(X)$$

Si X et Y sont deux $v.a.$ indépendantes, alors :

$$P_5) \quad V(X + Y) = V(X) + V(Y)$$

dans le cas général :

$$P_5') \quad V(X + Y) = V(X) + V(Y) + 2cov(X, Y)$$

On définit la covariance par

$$cov(X, Y) = E(XY) - E(X)E(Y).$$

Exemple La demande journalière X d'un bien fabriqué par une entreprise est une $v.a.$ qui suit la loi suivante :

$$P(X = 0) = \frac{1}{6}, \quad P(X = 1) = \frac{1}{6}, \quad P(X = 2) = \frac{1}{2}, \quad P(X = 3) = \frac{1}{6}.$$

On suppose que le profit, fonction de la demande et du coût, vérifie la relation:

$$\Pi(X) = p.X - C,$$

p étant le prix unitaire du bien fixé à 600 €, C étant le coût supposé indépendant de la demande et égal à 800€.

1. Calculez l'espérance et l'écart-type du profit. Quelle est la signification de l'espérance du profit ?

2. Déterminez la fonction de répartition du profit et tracez son graphe.

Réponses

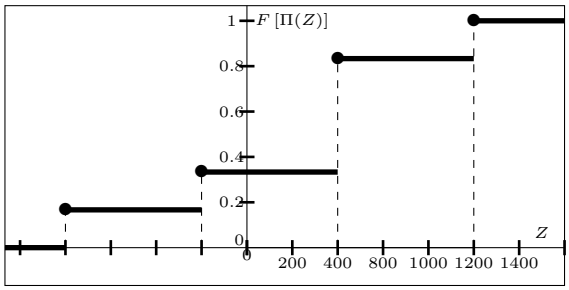
1.
$$E(X) = \frac{5}{3} = 1.667, \quad \sigma_X = 0.943.$$
$$E\left[\Pi(X)\right] = E\left[p.X - C\right] = pE(X) - C = 600 \times \frac{5}{3} - 800 = 200.$$
$$\sigma_{\Pi(X)} = p \times \sigma_X = 600 \times 0.943 = 565.68.$$

2. Loi de probabilité du profit :

valeur de X	0	1	2	3
valeur de Π	- 800	- 200	400	1 000
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$

Fonction de répartition du profit : fonction en escalier, continue à droite, les points de discontinuité correspondant aux valeurs possibles du profit.

Z	$<$ - 800	[- 800 , - 200[	[- 200, 400[	[400, 1 000[	$\geq$ 1 000
valeur de $F\left[\Pi(Z)\right]$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{5}{6}$	1



5.5 Variables aléatoires continues:

5.5.1 Définition

On appelle $v.a.$ réelle définie sur $(\Omega, \mathcal{A})$ une application

$$X : \Omega \longrightarrow \mathbb{R}$$

telle que pour tout intervalle $I \subset \mathbb{R} \doteq$ on ait :

$$X^{-1}(I) = \{\omega \in \Omega / X(\omega) \in I\} \in \mathcal{A}.$$

Il suffit en fait de vérifier que pour tout réel x :

$$X^{-1}(]-\infty, x[) \in \mathcal{A}$$

5.5.2 Loi de probabilité

Elle est déterminée par la fonction de répartition F, définie pour tout x réel par:

$$\begin{aligned} F(x) &= P_X(X < x) \\ &= P\{X^{-1}(]-\infty, x[)\} \\ &= P\{\omega \in \Omega / X(\omega) < x\} \end{aligned}$$

5.5.3 Propriétés de la fonction de répartition

- P_1) Elle est croissante au sens large.
- P_2) Elle prend ses valeurs entre 0 et 1 :

  $$0 \leq F(x) \leq 1$$

  avec

  $$\lim_{x \to -\infty} F(x) = 0$$

  et

  $$\lim_{x \to +\infty} F(x) = 1$$

- P_3) Elle est continue à gauche :

  $$\lim_{h \to 0^+} F(x-h) = F(x)$$

- P_4) La probabilité de l'intervalle $[a, b[$, pour $a < b$, se calcule par :

  $$P_X(a \leq X < b) = F(b) - F(a).$$

5.5.4 Loi continue

Si la fonction F est continue, *i.e.* continue à droite, on dit que X est une variable aléatoire réelle continue. Dans ce cas, pour tout réel x :

$$P_X(X = x) = 0$$

Exemple Considérons la $f.r.F$ définie par :

$$F(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ \frac{x}{4} & \text{si} \quad 0 < x \leq 1 \\ \frac{x}{2} & \text{si} \quad 1 < x \leq 2 \\ 1 & \text{si} \quad 2 < x \end{cases}$$

elle n'est pas continue en 1 car :

$$\lim_{h \to 0^+} F(1-h) = \frac{1}{4} = F(1) \neq F(1) = \lim_{h \to 0^+} F(1+h) = \frac{1}{2}$$

5.5.5 Loi absolument continue

La valeur moyenne de la probabilité d'un intervalle de longueur $h > 0$ est :

$$\frac{1}{h} P_X(x \leq X \leq x+h) = \frac{F(x+h) - F(x)}{h}$$

et représente donc une densité moyenne, si F admet une dérivée f :

$$\lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = F'(x) = f(x),$$

la fonction f étant appelée **densité de probabilité** de la $v.a.X$

$$F(x) = \int_{-\infty}^{x} f(\tau) d\tau$$

Propriétés

- P_1) Une densité est positive:

$$f \geq 0;$$

- P_2) Une densité est intégrable sur $\mathbb{R}$, d'intégrale égale à un :

$$\int_{-\infty}^{+\infty} f(\tau) d\tau = 1;$$

- P_3) La probabilité d'un intervalle s'obtient en intégrant la densité sur cet intervalle:

$$P_X\{x \in [x_1, x_2]\} = \int_{x_1}^{x_2} f(\tau) d\tau.$$

5.6 Moments d'une $v.a.$ absolument continue

5.6.1 Espérance mathématique

Définition Elle est définie par:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

5.6.2 Variance

Définition Elle est définie par:

$$V(X) = E\left[X - E(X)\right]^2 = \int_{-\infty}^{+\infty} \left[x - E(X)\right]^2 f(x) dx$$

$$= E(X^2) - E^2(X) = \sigma^2(X).$$

Exemple Soit la fonction

$$f(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ cx(1-x) & \text{si} \quad 0 \leq x \leq 1 \\ 0 & \text{si} \quad 1 \geq x \end{cases}$$

Pour quelle valeur de c est-ce une densité de probabilité ? Déterminer alors la fonction de répartition de cette loi et sa médiane.

Réponses On doit avoir d'abord

$$f(x) \geq 0 \quad pour \quad x \in [0,1]$$

soit $c > 0$ puisque

$$x(1-x) \geq 0 \quad sur \quad [0,1].$$

Puis :

$$\int_0^1 cx(1-x)dx = c \times \frac{1}{6}$$

doit être égal à 1. D'où c = 6. La fonction de répartition vaut, pour $x \in [0,1]$:

$$6\int_0^1 x(1-x)dx = -2x^3 + 3x^2.$$

Plus généralement:

$$F(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ -2x^3 + 3x^2 & \text{si} \quad 0 \leq x \leq 1 \\ 1 & \text{si} \quad 1 \geq x \end{cases}$$

On vérifie que F est continue partout. La médiane est la valeur de x telle que $F(x) = \frac{1}{2}$. L'équation n'est pas simple à résoudre mais l'on peut constater que le graphe de $f(x)$ est symétrique par rapport à $x = \frac{1}{2}$ qui est donc la médiane (et la moyenne).

Exemple Soit la loi de Pareto de paramètres strictement positifs a et θ, dont la fonction de densité est :

$$f(x) = \begin{cases} 0 & \text{si} \quad x < a \\ \frac{\theta}{a}\left(\frac{a}{x}\right)^{\theta+1} & \text{si} \quad a \leq x \end{cases}$$

1. Calculer la moyenne et la variance de cette loi. Quand ces moments existent ils ?
   Généraliser à l'existence d'un moment d'ordre quelconque.

2. Montrer que sa fonction génératrice des moments n'existe pas.

Réponses On a :

$$E(X) = \int_a^{+\infty} x \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx$$

$$E(X) = \theta a^{\theta} \int_a^{+\infty} \left(\frac{1}{x}\right)^{\theta} dx.$$

L'intégrale ci-dessus ne convergeant que si $\theta > 1$, la moyenne n'existe qu'à cette condition et vaut alors :

$$E(X) = \theta a^{\theta} \left[\frac{x^{-\theta+1}}{-\theta+1}\right]_a^{+\infty}$$

$$E(X) = \theta a^{\theta} \left[-\frac{a^{-\theta+1}}{-\theta+1}\right]$$

$$E(X) = a\frac{1}{1-\frac{1}{\theta}}.$$

Puis :

$$E(X^2) = \int_a^{+\infty} x^2 \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx$$

$$E(X^2) = \theta a^{\theta} \int_a^{+\infty} \left(\frac{1}{x}\right)^{\theta-1} dx.$$

L'intégrale ci-dessus ne convergeant que si $\theta > 2$, $E(X^2)$ et la variance n'existent qu'à cette condition. Alors :

$$E(X^2) = \theta a^{\theta} \left[\frac{x^{-\theta+2}}{-\theta+2}\right]_a^{+\infty}$$

$$E(X^2) = \theta a^{\theta} \left[-\frac{a^{-\theta+2}}{-\theta+2}\right]$$

$$E(X^2) = a^2\frac{1}{1-\frac{2}{\theta}}.$$

Ainsi :

$$V(X) = E(X^2) - [E(X)]^2$$

$$V(X) = a^2\frac{1}{1-\left(\frac{2}{\theta}\right)} - \left[a\frac{1}{1-\left(\frac{1}{\theta}\right)}\right]^2$$

$$V(X) = a^2\left[\frac{1}{1-\left(\frac{2}{\theta}\right)} - \frac{1}{\left[1-\left(\frac{1}{\theta}\right)\right]^2}\right]$$

$$V(X) = a^2 \left[\frac{\left[1 - \left(\frac{1}{\theta}\right)\right]^2}{\left[1 - \left(\frac{2}{\theta}\right)\right]\left[1 - \left(\frac{1}{\theta}\right)\right]^2} - \frac{1 - \left(\frac{2}{\theta}\right)}{\left[1 - \left(\frac{1}{\theta}\right)\right]^2\left[1 - \left(\frac{2}{\theta}\right)\right]} \right]$$

$$V(X) = a^2 \left[\frac{\left[1 + \left(\frac{1}{\theta}\right)^2 - \left(\frac{2}{\theta}\right)\right]}{\left[1 - \left(\frac{1}{\theta}\right)\right]^2\left[1 - \left(\frac{2}{\theta}\right)\right]} - \frac{1 - \left(\frac{2}{\theta}\right)}{\left[1 - \left(\frac{1}{\theta}\right)\right]^2\left[1 - \left(\frac{2}{\theta}\right)\right]} \right]$$

$$V(X) = a^2 \frac{1}{\left[1 - \left(\frac{1}{\theta}\right)\right]^2\left[1 - \left(\frac{2}{\theta}\right)\right]}.$$

Généralisons à

$$\mu_r = E(X_r)$$

avec $r > 2$, r entier :

$$E(X) = \int_a^{+\infty} x^r \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx$$

$$E(X) = \theta a^\theta \int_a^{+\infty} \left(\frac{1}{x}\right)^{\theta-r+1} dx.$$

L'intégrale ne converge que si $\theta - r + 1 > 1$, soit $\theta > r$.

Notons qu'il en va de même pour

$$\mu'_r = E\left[(X - \mu)^r\right]$$

qui s'exprime en fonction de

$$\mu_r, \mu_{r-1}, \cdots, \mu_2, \mu.$$

Or si μ_r existe, les moments d'ordres inférieurs existent. Pour $\theta > r$ on a donc :

$$E(X^r) = \theta a^\theta \left[\frac{x^{-\theta+r}}{-\theta + r} \right]_a^{+\infty}$$

$$E(X^r) = \theta a^\theta \left[-\frac{a^{-\theta+r}}{-\theta + r} \right]$$

$$E(X^r) = a^r \frac{1}{1 - \left(\frac{r}{\theta}\right)}.$$

La condition $\lambda > r$ laisse entendre que la fonction génératrice permettant pour θ fixé d'obtenir tous les moments ne peut exister, ce que nous vérifions directement en calculant:

$$\Psi(t) = E(e^{tX}) = \int_a^{+\infty} e^{tx} \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx$$

$$E(X) = \theta a^{\theta} \int_{a}^{+\infty} \frac{e^{tx}}{(x)^{\theta-r+1}} dx.$$

Or, quel que soit $\theta \in \mathbb{R}$,

$$\lim_{x \longrightarrow +\infty} \frac{e^{tx}}{(x)^{\theta-r+1}} \longrightarrow +\infty \quad si \quad t > 0$$

et l'intégrale ne peut converger.
Donc Ψ n'est définie dans aucun voisinage de 0, condition nécessaire pour la définition de la fonction génératrice des moments afin que

$$\Psi'(0), \Psi''(0), \; etc$$

puissent exister.

References

[1] S. Gilbert , *Probabilités analyse des données et statistique*,Editions Teclmip. Paris, (2006).

[2] J.P. Lecoutre, *Statistique et probabilités. Cours et exercices corrigés*, Dunod, (2016) .

[3] B. Goldfarb, C. Pardoux *Introduction à la méthode statistique, manuel et exercices corrigés*, Dunod, Paris, (2011).

[4] D. François , *Les Probabilités et la statistique de A à Z*, Dunod, Paris, (2007).

[5] M. Lejeune, *Statistique La théorie et ses applications*, Springer-Verlag France, Paris, (2010).

[6] J.J. Boreux, E. Parent, J. Bernier, *Pratique du calcul bayesien*, Springer-Verlag France, Paris, (2010).

[7] R. Rakotomalala, *Comparaison de populations Tests non paramétriques*,Université Lumière Lyon 2 (2008).

[8] T. Gérard,D. Jacques, *TikZ pour l'impatient*,(2017).

http://math.et.info.free.fr/TikZ/bdd/TikZ-Impatient.pdf,30/12/2018:15:16

[9] https://en.wikibooks.org/wiki/LaTeX/Tables,12:23,31/12/2018.

[10] https://www.developpez.net/forums/d1592510/autres-langages/autres-langages/latex/tableaux-graphiques-images-flottants/tikz-tracer-portion-courbe/,15:50,01/01/2019.