Première partie

La statistique descriptive

Introduction

Le but de la statistique descriptive est de présenter des données observées ou mesurées (résultats de mesures en laboratoires, chiffres des ventes dans un magasin, tailles d'individus,...) sous une forme telle que l'on puisse aisément en prendre connaissance et que l'on puisse éventuellement mettre en évidence des propriétés remarquables.

Les données considérées ici seront relatives à des *variables numériques* (à valeurs réelles); on distinguera les variables *discrètes*, dont l'ensemble des valeurs possibles est fini ou dénombrable (nombres de camions, résultats d'examens, pourcentages d'abstentions à un vote,...) et les variables *continues*, dont l'ensemble des valeurs possibles est isomorphe à $\mathbb{R}$ (longueur, poids,...).

Chapitre 1

Statistique descriptive à une dimension

1.1 Les distributions de fréquences

Soit n le nombre de valeurs observées (de mesures) d'une variable numérique discrète dont les valeurs positives, rangées dans l'ordre croissant, sont x_1,x_2, . . . ,x_p.

1.1.1 Définitions

- La *fréquence absolue* d'une valeur x_i est le nombre n_i d'observations égales à x_i;

- La *fréquence relative* d'une valeur x_i est le rapport $\frac{n_i}{n}$; (exprimée en %, elle vaut $\frac{100n_i}{n}$);

- La *fréquence* (absolue ou relative) *cumulée* d'une valeur x_i est la somme des fréquences (absolues ou relatives) de cette valeur et des valeurs inférieures;

- La *distribution des fréquences* (absolues ou relatives, cumulées ou non) d'une variable est un tableau contenant les valeurs possibles de cette variable, rangées dans l'ordre croissant, et, pour chacune de ces valeurs, la fréquence (absolue ou relative, cumulée ou non) correspondante.

1.1.2 Groupements des données

Lorsque le nombre de valeurs observées distinctes est grand, on a intérêt à grouper les données. Cela signifie qu'au lieu d'énumérer les valeurs de la variable, on partitionne le domaine de celle-ci en intervalles contigus appelés *classes*.

Dans ce cas :

- la *fréquence absolue* d'une classe C_i est le nombre n_i d'observations appartenant à l'intervalle C_i;
- la *fréquence relative* d'une classe C_i est le rapport $\frac{n_i}{n}$;
- la *fréquence* (absolue ou relative) *cumulée* d'une classe C_i est la somme des fréquences (absolues ou relatives) de cette classe et des classes précédentes;
- la *fréquence* (absolue ou relative) *unitaire* d'une classe C_i est le rapport $\frac{n_i}{\text{longueur de } C_i}$;
- la *distribution groupée des fréquences* d'une variable est un tableau contenant les classes de cette variable et, pour chacune de ces classes, la fréquence correspondante.

Il n'existe pas de loi rigoureuse qui fixe le nombre de classes et leurs longueurs; cela dépend évidemment du problème concret considéré. Le but du groupement des données est d'obtenir des tableaux et des graphiques aisément interprétables, tout en gardant un maximum d'informations. Lorsque les données le permettent, on préfère définir des classes de longueurs égales : de cette façon, la comparaison des fréquences de deux classes a un sens (sinon, il faut utiliser les fréquences unitaires). En pratique, le nombre de classes est rarement supérieur à 20 (pour rendre possible l'interprétation des graphiques) et rarement inférieur à 5 (pour ne pas perdre toute l'information). Les classes peuvent être désignées par leurs extrémités (qui sont des valeurs possibles ou non de la variable) ou par leur milieu si elles ont même longueur (dans ce dernier cas, il ne faut pas perdre de vue que les données ont été groupées pour les graphiques et le calcul des valeurs typiques : cf. plus loin). Les classes extrêmes peuvent ne pas être bornées.

1.1.3 Représentations graphiques

On représente généralement les distributions de fréquence par des graphiques, de façon à visualiser le comportement de la variable étudiée.

Le *diagramme en bâtons* consiste à porter en abscisse les valeurs observées x_i et à tracer en regard de chacune d'elles un segment vertical de longueur égal à sa fréquence (absolue ou relative) non cumulée. Ce type de graphique s'applique aux distributions non groupées.

L'*histogramme* se compose de rectangles contigus dont les bases sont les classes de la variable et dont les *aires* sont proportionnelles aux fréquences de ces classes. (Les hauteurs de ces rectangles sont donc les fréquences si les classes ont même longueur et les fréquences unitaires sinon). Ce type de graphique s'applique aux distributions groupées et concerne les fréquences (absolues ou relatives) cumulées ou non.

Le *polygone de fréquences* s'obtient en joignant par des segments de droite les extrémités des segments voisins des diagrammes en bâtons pour les distributions non groupées et (mais ceci est rarement appliqué) les milieux des bases supérieures des rectangles successifs des histogrammes pour les distributions groupées.

Le *polygone de fréquences relatives cumulées* est, pour les *distributions non groupées*, le graphique de la fonction $F^*(x)$ définie comme suit : $\forall x \in \mathbb{R}$,

$$F^*(x) = \begin{cases} 0 & \text{si } x < x_1, \\ \dfrac{n_1 + n_2 + \ldots + n_i}{n} & \text{si } x_i \leq x < x_{i+1} \quad (i = 1,2,\ldots,p-1) \\ 1 & \text{si } x \geq x_p. \end{cases}$$

Cette fonction, appelée *fonction de distribution de la variable*, est une fonction en escalier, non décroissante, continue à droite et variant de 0 à 1.

Pour les distributions groupées, on porte, en regard des extrémités supérieures des classes, des ordonnées égales aux fréquences relatives cumulées de ces classes et on rejoint les points successifs par des segments de droite. Cela revient à associer à chaque x de $\mathbb{R}$ une ordonnée égale au nombre d'observations inférieures ou égale à x (en valeur relative), si on suppose qu'à l'intérieur de chaque classe les valeur observées sont distribuées uniformément. La fonction obtenue s'appelle la fonction de distribution de la variable.

Le *polygone de fréquences absolues cumulées* est le même que celui des fréquences relatives cumulées, les ordonnées étant simplement multipliées par n.

Pour représenter simultanément plusieurs distributions sur un même graphique, on utilisera de préférence les polygones de fréquences relatives cumulées.

1.1.4 Principaux types de distributions de fréquences

Bien que l'allure du graphique d'une distribution puisse être tout à fait quelconque, on distingue 4 types fondamentaux de distributions souvent rencontrées en pratique. La distribution de fréquences (absolues ou relatives) non cumulées d'une variable peut notamment être :

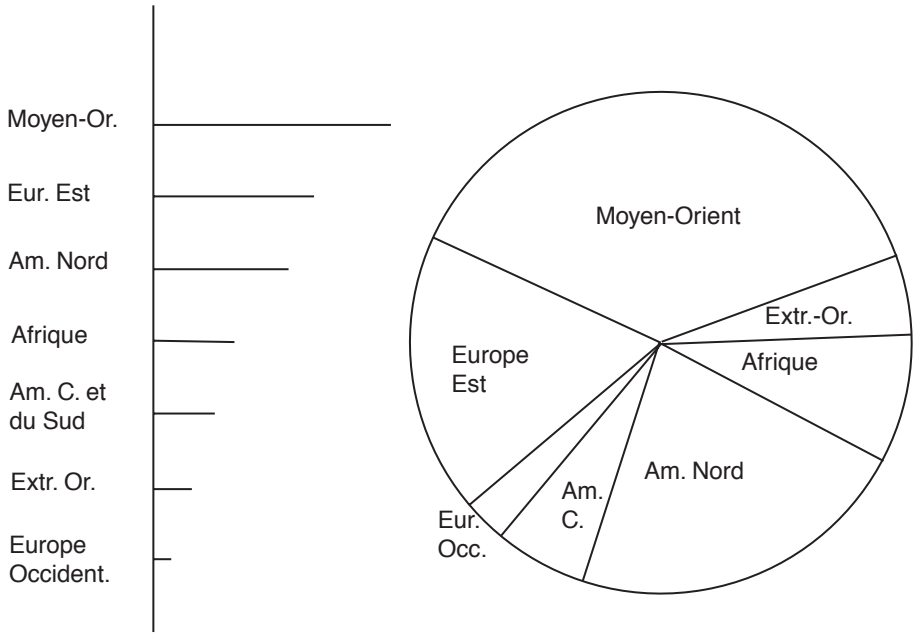
- en cloche (une zone de fréquences maximales entourée de 2 zones où les fréquences diminuent progressivement);
- en u (une zone de fréquences minimales entourée de 2 zones où les fréquences augmentent progressivement);
- en i (les fréquences diminuent de gauche à droite);
- en j (les fréquences augmentent de gauche à droite);

1.1.5 Représentations graphiques pour des variables nominales

Une *variable nominale* est une variable dont les valeurs ne peuvent pas être rangées dans un ordre. Pour ce type de variable, la notion de fréquence cumulée n'a pas de sens et aucun des graphiques vus précédemment n'est possible. On doit donc se contenter, dans ce cas, de tableaux associant à chaque valeur de la variable la fréquence de cette valeur et de graphiques sous forme de bâtons de longueurs proportionnelles aux fréquence ou de régions planes d'aires proportionnelles aux fréquences, comme dans l'exemple qui suit.

Afrique	Amérique du Nord	Amérique Centrale et du Sud	Europe Occidentale	Europe de l'Est, U.R.S.S., Chine	Moyen Orient	Extrême-Orient Australie
9,7%	18,2%	7,8%	1,3%	21,6%	37,2%	4,2%

Production de pétrole dans le monde en 1976



1.1.6 Exemples

a) **Données non groupées**

La "série statistique" ci-dessous donne le nombre de voitures vendues au cours du dernier mois par chacun des 40 distributeurs de la marque x.

7 , 1 , 5 , 12 , 3 , 2 , 1 , 4 , 7 , 5 ,
6 , 4 , 1 , 8 , 10 , 3 , 3 , 6 , 5 , 2
5 , 8 , 2 , 6 , 0 , 7 , 4 , 4 , 6 , 8
5 , 5 , 4 , 7 , 8 , 10 , 6 , 5 , 3 , 3

Variable étudiée: nombre de voitures vendues au cours du dernier mois. Nombre de valeurs observées : $n = 40$

Valeurs de la variable	Fréqu. absolues	Fréqu. relatives	Fréq. rel. en %	Fréq. abs. cumulées	Fréq. rel. cumulées	Fréq. rel cum. en %
0	1	1/40	2,5	1	1/40	2,5
1	3	3/40	7,5	4	4/40	10
2	3	3/40	7,5	7	7/40	17,5
3	5	5/40	12,5	12	12/40	30
4	5	5/40	12,5	17	17/40	42,5
5	7	7/40	17,5	24	24/40	60
6	5	5/40	12,5	29	29/40	72,5
7	4	4/40	10	33	33/40	82,5
8	4	4/40	10	37	37/40	92,5
9	0	0	0	37	37/40	92,5
10	2	2/40	5	39	39/40	97,5
11	0	0	0	39	39/40	97,5
12	1	1/40	2,5	40	1	100

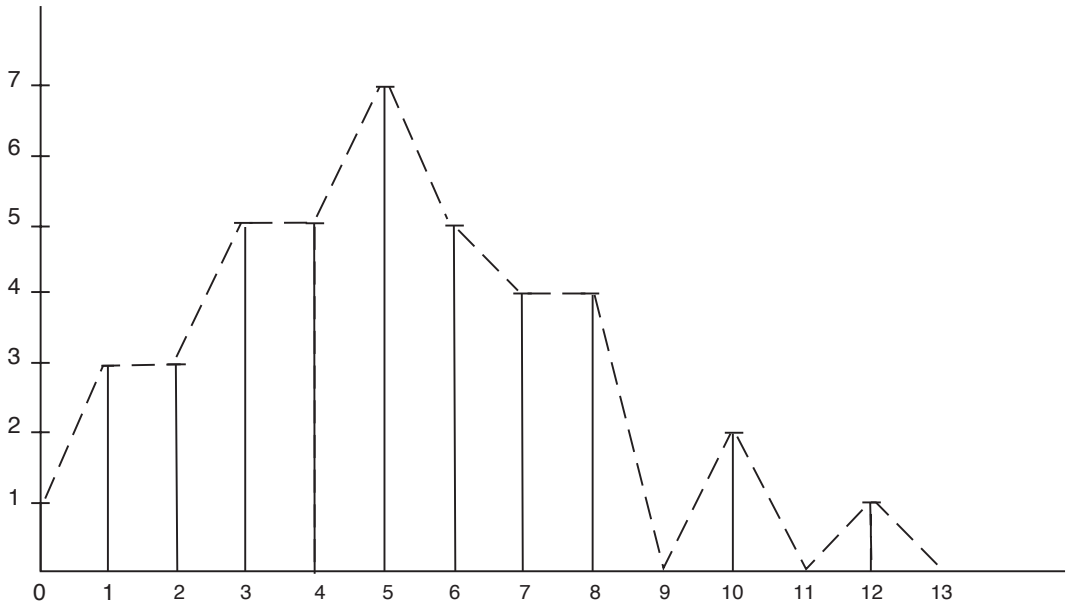
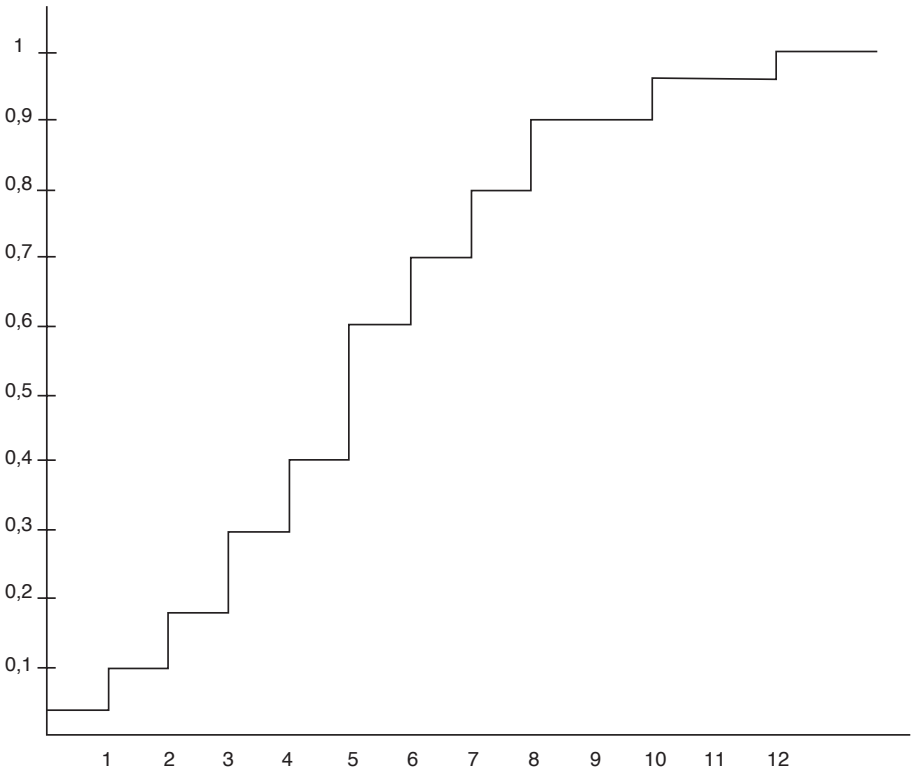


Diagramme en bâtons des fréquences absolues

En pointillé : polygone de fréquences absolues



Polygone des fréquences relatives cumulées (diagramme de la *fonction de distribution)*

b) **Données groupées**

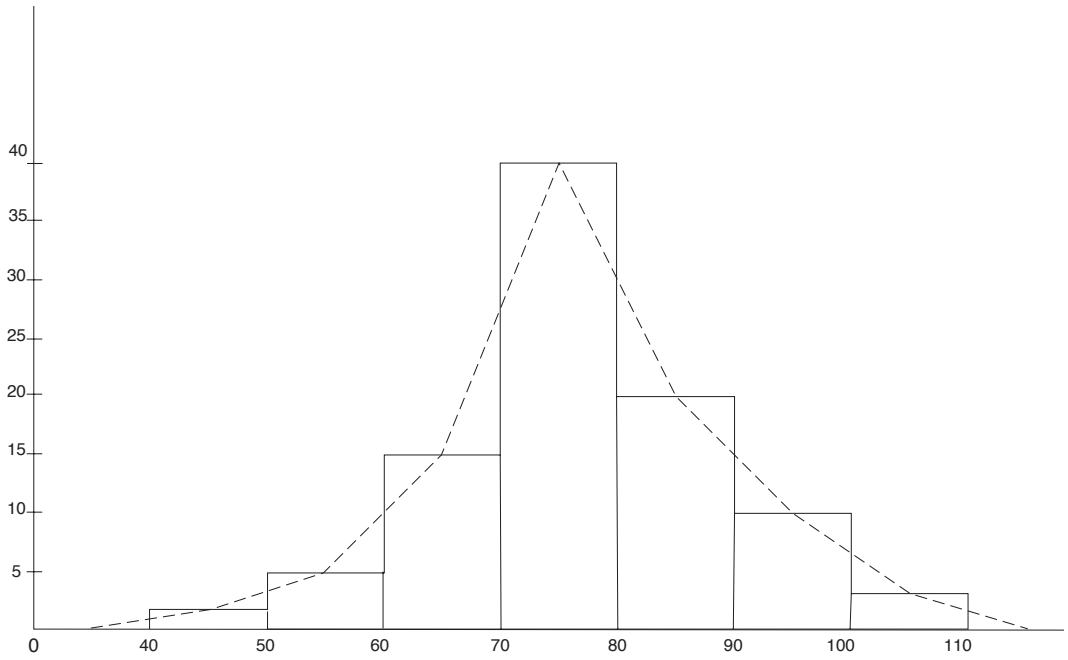
Le responsable du stock d'un atelier a noté, au cours de 98 jours de travail, le nombre de boulons d'un certain type utilisés dans cet atelier. Les données brutes sont données ci-dessous.

72	51	56	95	68	66	77	81	83	75
41	79	92	78	85	55	104	76	80	61
65	70	83	92	88	59	75	75	81	69
71	96	101	87	65	74	68	73	78	68
73	86	84	51	85	75	79	90	68	71
75	74	81	64	88	78	77	66	91	75
69	73	82	75	76	71	74	96	72	74
102	74	80	82	86	78	87	61	80	78
48	68	71	66	59	92	77	76	81	70
85	77	68	82	78	75	91	77		

Ces données, groupées en classes de longueur 10, donnent les distributions de fréquences suivantes.

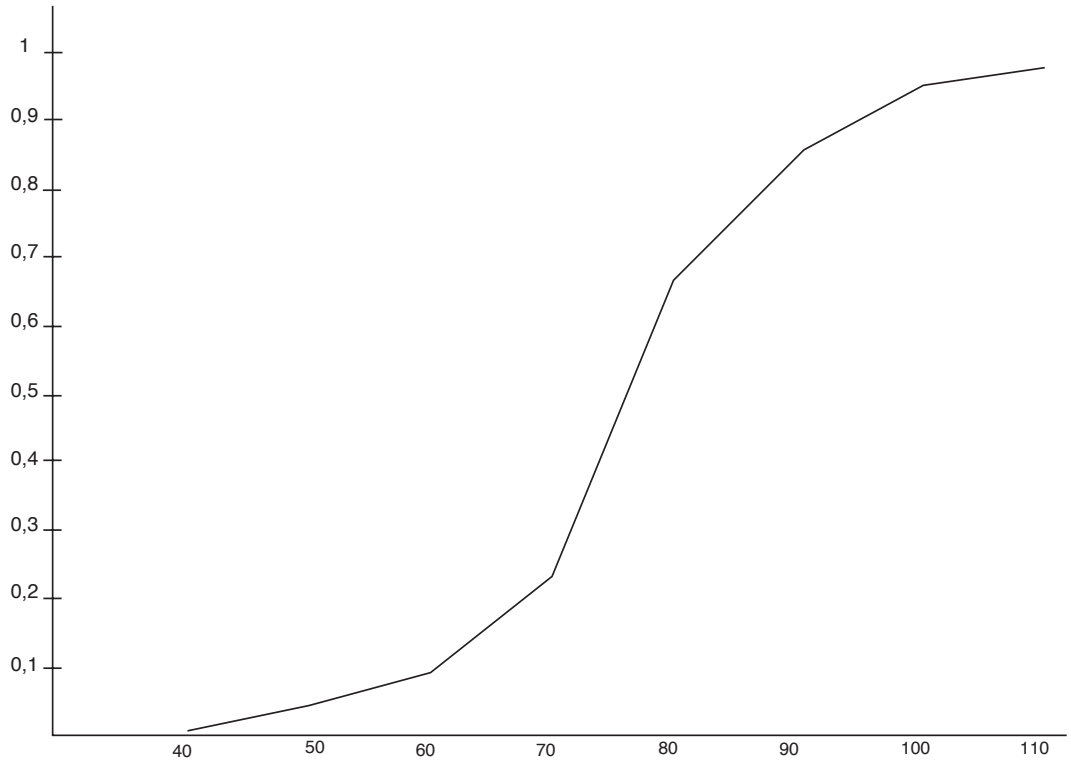
Classes	Fr. abs.	Fr. rel. en %	Fr. abs. cumul.	Fr. rel. cumul.
40-49	2	2,04	2	2,04
50-59	6	6,12	8	8,16
60-69	16	16,33	24	24,49
70-79	40	40,82	64	65,31
80-89	22	22,45	86	87,76
90-99	9	9,18	95	96,94
100-109	3	3,06	98	100

Ce tableau permet de voir que le nombre de boulons utilisés chaque jour se situe ”en général” entre 60 et 90 et ”pratiquement toujours” entre 40 et 110 et qu'il faut, ”en moyenne ”, environ 75 boulons par jour. Pour avoir des informations plus précises et pour quantifier les expressions ”en général”et ”pratiquement toujours”, on utilisera le calcul des probabilités et l'inférence statistique (cf. plus loin).



Histogramme des fréquences absolues

En pointillé : polygone de fréquences absolues



Polygone des fréquences relatives cumulées (diagramme de la *fonction de distribution)*

1.2 Valeurs typiques des distributions de fréquences

Au lieu de décrire complètement une distribution de fréquences par un tableau ou un graphique, on peut songer à résumer ses caractéristiques essentielles au moyen de quelques paramètres bien choisis. On distingue essentiellement les paramètres de position (pour caractériser l'ordre de grandeur des observations), les paramètres de dispersion (pour caractériser la variabilité des observations), les paramètres de dissymétrie et d'aplatissement. De manière générale, on préférera calculer les valeurs de ces paramètres directement à partir des observations; on évitera donc, lorsque c'est possible, le groupement des données en classes car cette opération introduit nécessairement une certaine imprécision dans le calcul des paramètres.

1.2.1 Les paramètres de position

a) **La moyenne arithmétique**

Définition :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i,$$

où x_i sont les valeurs observées (pour les distributions non groupées) ou les milieux des classes (pour les distributions groupées).

Propriétés :

1) si on fait subir une transformation linéaire aux x_i, leur moyenne arithmétique subit la même transformation : $x'_i = ax_i + b, \forall i \Rightarrow \bar{x'} = a\bar{x} + b$ (démonstration immédiate); on utilise parfois cette propriété pour simplifier le calcul de $\bar{x}$;
2) si deux séries statistiques exprimées dans les mêmes unités ont respectivement comme moyennes $\bar{x}_1$ et $\bar{x}_2$, alors la série obtenue en les réunissant a comme moyenne

$$\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2},$$

où n_1 et n_2 sont les effectifs des 2 séries (démonstration immédiate).

La moyenne arithmétique étant la plus utilisée, nous l'appellerons dorénavant moyenne.

b) **La moyenne géométrique**

Définition :

$$\bar{x}_G = \left[\prod_{i=1}^{p} x_i^{n_i} \right]^{\frac{1}{n}} \qquad (x_i > 0)$$

Propriétés :

1) le logarithme de $\bar{x}_G$ est la moyenne (arithmétique) des logarithmes des x_i;
2) la moyenne géométrique est inférieure ou égale à la moyenne arithmétique.

c) **La moyenne harmonique**

Définition :

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{p} \frac{n_i}{x_i}} \qquad (x_i > 0)$$

Propriété : $\bar{x}_H \leq \bar{x}_G \leq \bar{x}$

d) **La moyenne quadratique**

Définition :

$$\bar{x}_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2} \qquad (x_i \geq 0)$$

Propriété : $\bar{x}_Q \geq \bar{x} \geq \bar{x}_G \geq \bar{x}_H$

e) **La médiane**

Définition : $\widetilde{x}$ est un nombre qui laisse autant d'observations à sa gauche qu'à sa droite; c'est donc un nombre tel que $F^*(\widetilde{x}) = \frac{1}{2}$, où F^* est la fonction de distribution de la variable observée; si F^* présente un saut de discontinuité passant d'une valeur de $< \frac{1}{2}$ à une valeur $> \frac{1}{2}$, on convient de prendre l'abscisse correspondant à ce saut de discontinuité; si F^* présente un palier à l'ordonnée $\frac{1}{2}$, on convient de prendre le milieu du palier.

Propriété : la médiane est égale à la moyenne (arithmétique) lorsque la distribution de fréquences est symétrique (symétrie du polygone des fréquences par rapport à un axe vertical).

f) **Le Mode**

Un mode est une abscisse dont la fréquence est un maximum local. Une distribution peut posséder plusieurs modes; dans le cas des distributions groupées, on parle de classe modale (classe de fréquence localement maximale).

1.2.2 Utilisation des différents paramètres de position

Le paramètre de position le plus utilisé est la moyenne arithmétique : c'est le plus naturel et aussi celui qui possède les propriétés les plus intéressantes (existe toujours, toujours unique, dépend de toutes les valeurs observées, possède des propriétés algébriques facilitant les calculs,...). Il existe cependant des situations particulières où les autres paramètres peuvent être utiles.

1ère situation

Un capital de 1000 F a été placé pendant 4 ans; le taux d'intérêt a été de 8% pendant les 2 premières années et de 10% pendant les 2 années suivantes. Quel a été le taux d'intérêt moyen ?

Le taux d'intérêt moyen est le taux d'intérêt constant auquel il faut placer 1000 F pendant 4 ans pour obtenir à la fin le même capital que dans la situation décrite. Il faut donc rechercher la valeur de r telle que

$$1000 \times (1+r)^4 = 1000 \times (1+0{,}08)^2(1+0{,}1)^2,$$

ce qui donne

$$1+r = \sqrt[4]{(1+0{,}08)^2(1+0{,}1)^2} \neq 1{,}09.$$

Le taux d'intérêt moyen s'obtient donc en calculant une *moyenne géométrique* et non une moyenne arithmétique.

2ème situation

Une voiture parcourt un premier kilomètre à une vitesse de 60 km/h et un second kilomètre à 100 km/h. Quelle est sa vitesse moyenne ?
La vitesse moyenne s'obtient en divisant l'espace total parcouru par le temps mis pour le parcourir, ce qui donne

$$\frac{2}{\frac{1}{60}+\frac{1}{100}} = 75 \text{ km/h}.$$

Il s'agit donc ici d'une *moyenne harmonique*.

3ème situation

2 disques ont respectivement comme rayons r_1 et r_2 cm; quel est le rayon du disque d'aire moyenne ?

Le rayon r cherché doit satisfaire l'égalité

$$\pi r^2 = \frac{\pi r_1^2 + \pi r_2^2}{2}$$

d'où

$$r = \sqrt{\frac{1}{2}(r_1^2 + r_2^2)},$$

qui est une *moyenne quadratique*.

4ème situation

Dans une entreprise de 5 personnes, les salaires mensuels sont respectivement

de 900, 1000, 800, 850, et 3000 euros. Prétendre que le salaire moyen, dans cette entreprise, est de 1310 euros (moyenne arithmétique) ne traduit évidemment pas correctement la réalité. Dans ce cas, la médiane (900 euros) est beaucoup plus satisfaisante. Ce sera le cas chaque fois que les valeurs observées contiennent quelques valeurs "aberrantes" (c'est-à-dire beaucoup plus grandes ou beaucoup plus petites que la plupart des autres valeurs). La médiane présente l'avantage, par rapport à la moyenne, de ne pas dépendre des valeurs extrêmes qui, dans beaucoup de problèmes, peuvent être considérées comme anormales ou douteuses (erreurs de mesure). Le mode présente également cet avantage, mais il ne suffit certainement pas à caractériser la position d'une distribution : ses principaux inconvénients sont de ne pas être toujours unique et d'être sensible à de petites modifications des données.

1.2.3 Les paramètres de dispersion

a) **La variance**

Définition :

$$s^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2$$

où les x_i désignent toujours les valeurs observées pour les distributions non groupées et les milieux de classes pour les distributions groupées.

Propriétés :

1) la variance est nulle ssi toutes les valeurs observées sont égales entre elles;
2) si on multiplie les données par k, leur variance est multipliée par k^2 : $x'_i = kx_i, \forall i \Rightarrow s'^2 = k^2 s^2$;
3) une translation des données n'affecte pas leur variance : $x'_i = x_i + t, \forall i \Rightarrow s'^2 = s^2$;

b) **L'écart type**

Définition : $s = \sqrt{s^2}$ (racine carrée positive de la variance)

Propriétés : résultent de celles de la variance.

c) **Le coefficient de variation**

Définition : $v = \dfrac{s}{\bar{x}}$ (utilisé uniquement lorsque $\bar{x} > 0$)

Propriétés :

1) le coefficient de variation est nul ssi toutes les valeurs observées sont égales entre elles;
2) c'est un paramètre indépendant des unités de mesure utilisées : $x'_i = kx_i$, $\forall i \Rightarrow v' = v$.

d) **L'amplitude**

Définition : c'est la différence entre la plus grande et la plus petite valeurs observées.

Propriétés :

1) L'amplitude est nulle ssi toutes les valeurs observées sont égales entre elles;
2) $x'_i = kx_i, \quad \forall i \Rightarrow a' = ka$;
3) $x'_i = x_i + t, \quad \forall i \Rightarrow a' = a$.

e) **L'écart moyen absolu**

Définition :

$$e = \frac{1}{n} \sum_{i=1}^{p} n_i |x_i - \bar{x}|.$$

Propriétés :

1) $e = 0$ ssi $x_i = x_j, \forall i,j$;
2) $x'_i = kx_i \Rightarrow e' = ke$;
3) $x'_i = x_i + t \Rightarrow e' = e$.

f) **L'intervalle interquartile**

Définition : C'est l'intervalle entre le nombre qui laisse $\dfrac{1}{4}$ d'observations à sa

gauche (1e quartile) et le nombre qui laisse $\frac{3}{4}$ d'observations à sa gauche (3$^{\text{ème}}$ quartile).

Les 3 quartiles sont donc les nombres où la fonction de distribution F^* vaut respectivement $\frac{1}{4}$,$\frac{1}{2}$, et $\frac{3}{4}$ (le 2$^{\text{ème}}$ quartile n'est rien d'autre que la médiane). Lorsque F^* présente des points de discontinuité ou des paliers, on procède comme pour la médiane.

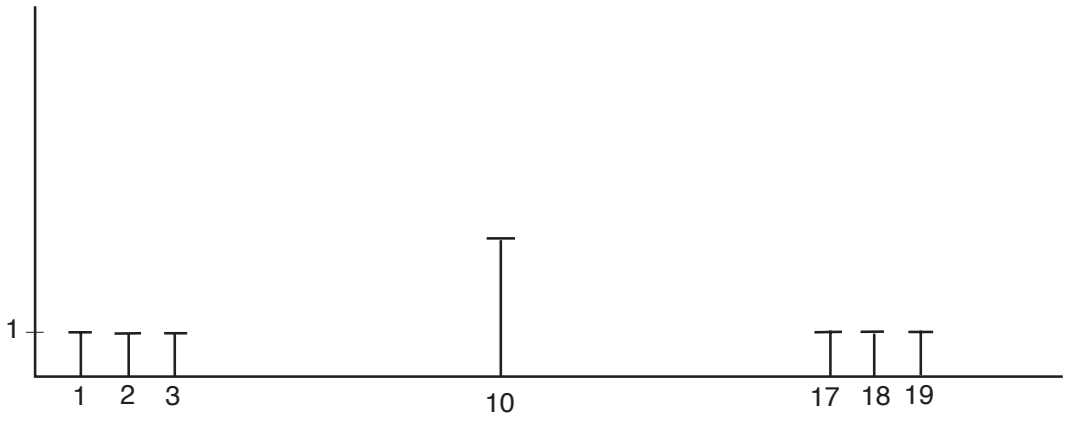
1.2.4 Utilisation des différents paramètres de dispersion

Le premier paramètre de dispersion qui se présente à l'esprit est l'écart moyen : $\frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})$ mais ce paramètre étant toujours nul, il ne présente aucun intérêt. C'est la raison pour laquelle les écarts $x_i - \bar{x}$ sont considérés soit en valeur absolue (écart moyen absolu), soit au carré (variance). La valeur absolue n'étant pas un objet mathématique facilement maniable, on utilise très peu l'écart moyen absolu. La variance, l'écart-type et le coefficient de variation sont donc les paramètres les plus utilisés. L'écart-type présente l'avantage de s'exprimer dans les mêmes unités que les observations. Le coefficient de variation est utile pour comparer les dispersions de deux distributions et pour relativiser l'écart-type (un écart-type de 1000 n'a pas la même signification selon que les valeurs observées sont de l'ordre de 10.000 ou de 100.000).

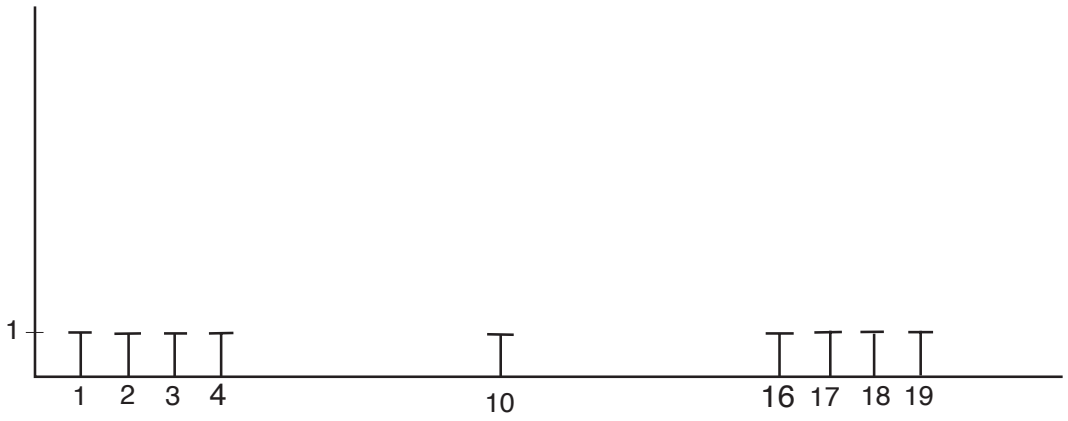
Les paramètres précédents dépendent de toutes les observations et en particulier des éventuelles observations aberrantes (comme la moyenne arithmétique), ce qui n'est pas le cas de l'intervalle interquartile (comme la médiane). Celui-ci a cependant le désavantage de ne pas être très sensible à des modifications des valeurs observées. L'amplitude apporte une information intéressante mais est tout à fait insuffisante pour caractériser la dispersion d'une distribution car elle ne dépend que des valeurs extrêmes.

Exemple

Valeurs observées	1	2	3	4	10	16	17	18	19
Fréquences absolues									
1 $^{\text{ère}}$ distribution	1	1	1	0	3	0	1	1	1
2 $^{\text{ère}}$ distribution	1	1	1	1	1	1	1	1	1



1ère distribution



2ème distribution

	1ère distribution	2ème distribution
Moyenne	10	10
Amplitude	18	18
Intervalle interquart.	17-3=14	17-3=14
Ecart-type	6,55	7,15

Dans ce cas, seul l'écart-type permet de voir que les dispersions des 2 distributions sont différentes.

1.2.5 Le paramètre de dissymétrie

$$m_3 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^3$$

sera nul si la distribution est symétrique, différent de 0 si la distribution présente une dissymétrie et d'autant plus grand en valeur absolue que la dissymétrie sera forte.

Ce paramètre ne varie pas avec une translation des données mais dépend des unités choisies. Pour avoir un nombre sans dimension, on utilise généralement le *paramètre de Fisher*

$$g_1 = \frac{m_3}{s^3}$$

Interprétation

Lorsqu'une distribution de probabilités est symétrique par rapport à sa moyenne, la somme des écarts $(x_i - \bar{x})^3$ positifs compense la somme des écarts négatifs, de sorte que $m_3 = g_1 = 0$.

Lorsque les $(x_i - \bar{x})$ négatifs sont petits en valeur absolue et fréquents et que les $(x_i - \bar{x})$ positifs sont grands en valeur absolue et peu fréquents, on dit que l'on a une dissymétrie gauche. Comme

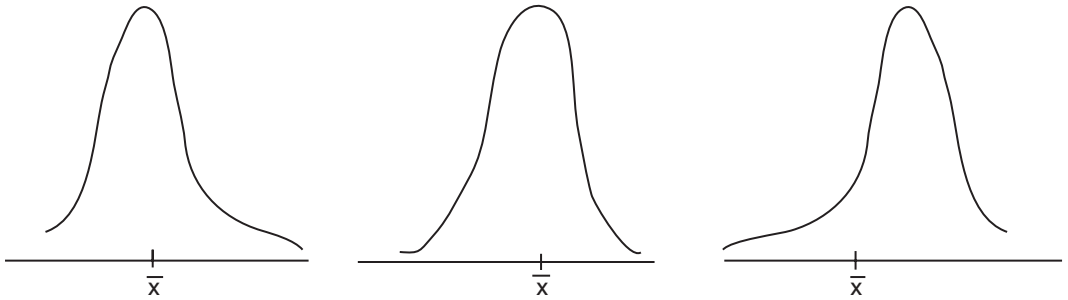
$$\frac{1}{n}\sum_{i} n_i(x_i - \bar{x}) = 0,$$

on a dans ce cas :

$$\sum_{i:x_i-\bar{x}>0} n_i(x_i - \bar{x})^3 > -\sum_{i:x_i-\bar{x}<0} n_i(x_i - \bar{x})^3$$

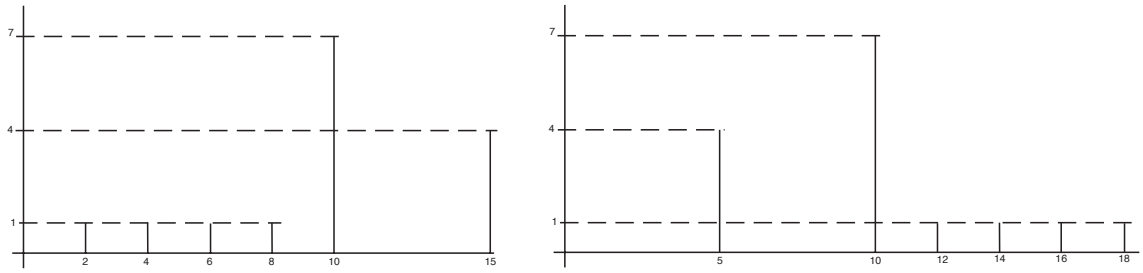
de sorte que $m_3 > 0$ et $g_1 > 0$.

Lorsque la dissymmétrie est droite, on a $m_3 < 0$ et $g_1 < 0$.



Dissymétrie gauche	Symétrie	Dissymétrie droite
$g_1 > 0$	$g_1 = 0$	$g_1 < 0$

Exemple.



Ces deux distributions ont même moyenne ($\bar{x} = 10$) et même variance ($s^2 = 14{,}66$). Celle de gauche présente une "dissymétrie droite" et celle de droite une "dissymétrie gauche" : c'est le paramètre g_1 qui permet de faire la différence ($g_1 = -0{,}35$ à gauche et $+0{,}35$ à droite).

1.2.6 Le paramètre d'aplatissement

On utilise habituellement

$$m_4 = \frac{1}{n} \sum_{i=1}^{p} n_i (x_i - \bar{x})^4$$

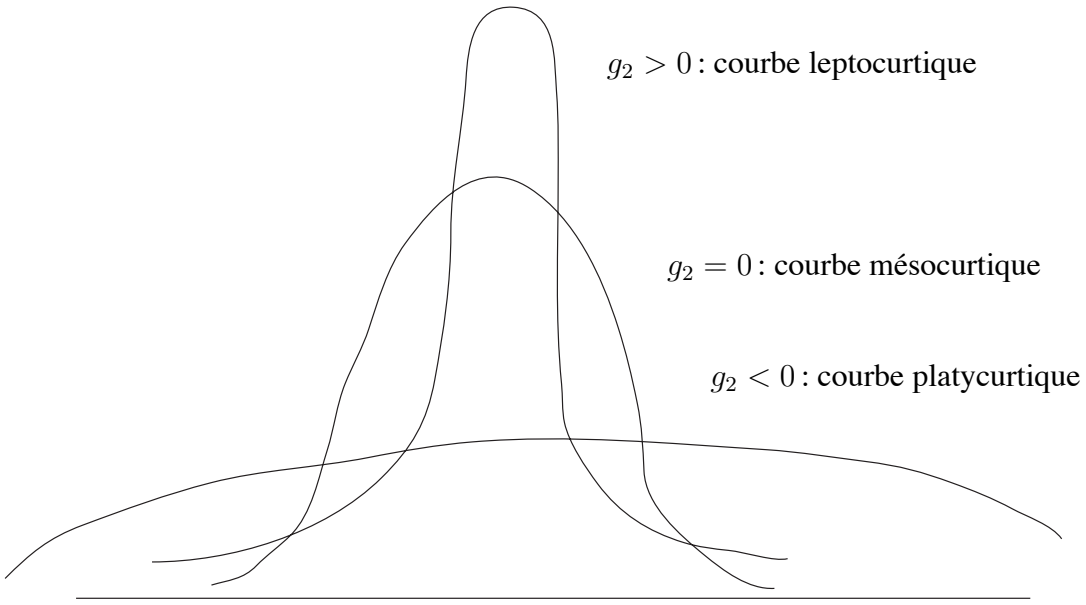
ou le *deuxième paramètre de Fisher*

$$g_2 = \frac{m_4}{s^4} - 3$$

Interprétation

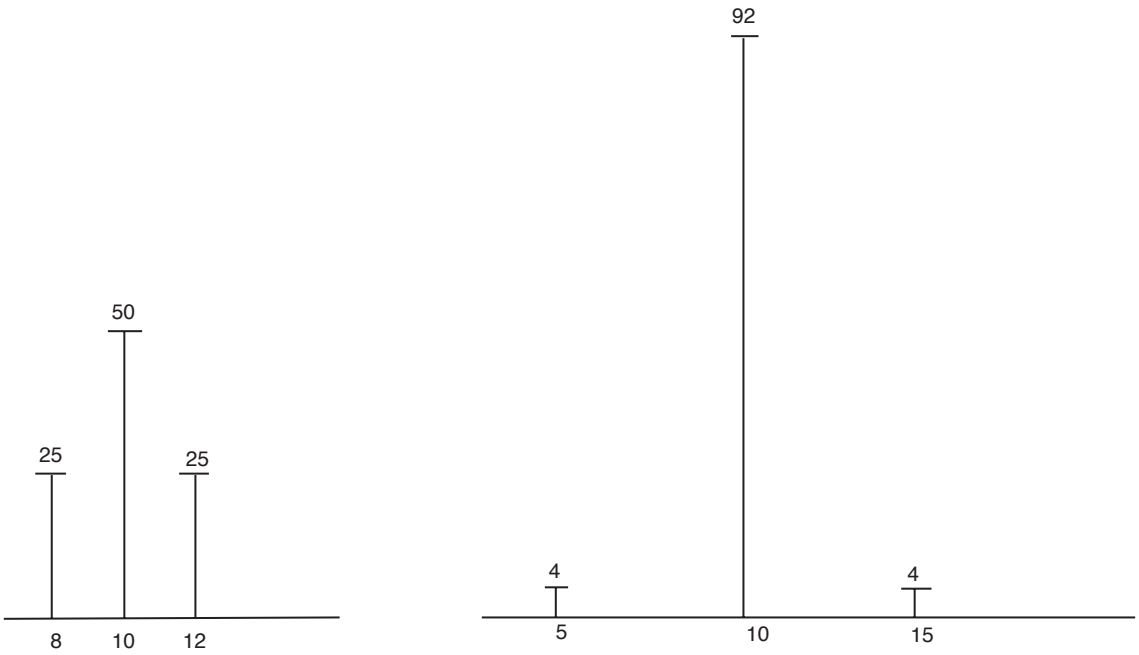
On verra plus loin que, pour une distribution particulière souvent utilisée en statistique, et appelée distribution normale, le paramètre g_2 est nul. Par conséquent, une distribution symétrique pour laquelle $g_2 > 0$ est plus pointue que la normale de même variance, en ce sens que les petits et les grands $(x_i - \bar{x})$ sont plus fréquents et que les $|x_i - \bar{x}|$ intermédiaires sont moins fréquents (cela est dû à la définition de g_2 qui est liée aux 4èmes puissances des $|x_i - \bar{x}|$).

Une distribution pour laquelle $g_2 < 0$ est plus aplatie que la normale de même variance.



Le coefficient g_2 permet donc de comparer une distribution symétrique à la normale de même variance.

Exemple



Ces deux distributions ont même moyenne $(\bar{x} = 10)$, même variance $(s^2 = 2)$ et même paramètre de dissymétrie $(m_3 = 0$: distributions symétriques) mais leurs formes sont très différentes. C'est le paramètre g_2 qui permet de faire la différence : $g_2 = -1$ à gauche et $+9{,}5$ à droite.

1.2.7 Les moments

Le moment d'ordre k par rapport au point a est la quantité

$$\frac{1}{n}\sum_{i=1}^{p} n_i (x_i - a)^k.$$

En pratique, on utilise surtout les moments par rapport à l'origine :

$$a_k = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^k,$$

et les moments par rapport à la moyenne

$$m_k = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^k.$$

Propriétés.

- $a_1 = \bar{x}$
- $m_1 = 0$
- $m_2 = s^2 = a_2 - \bar{x}^2$
- m_2 est le plus petit moment d'ordre 2 : en effet, $\forall a$:

$$\frac{1}{n}\sum_{i=1}^{p} n_i(x_i - a)^2 = m_2 + (\bar{x} - a)^2.$$

- $x'_i = bx_i, \forall i \Rightarrow \begin{cases} a'_k = b^k a_k \\ m'_k = b^k a_k \end{cases}$
- $x'_i = x_i + t, \quad \forall i \Rightarrow m'_k = m_k$

1.2.8 Remarques sur le calcul des valeurs typiques

a) Les formules de définition de la variance et des moments m_3 et m_4, notamment, se prêtent mal au calcul. Il faut en effet calculer tous les écarts $(x_i - \bar{x})$, les élever à la puissance 2, 3 ou 4 et en faire la somme. Si les écarts $x_i - \bar{x}$ ne sont pas extrêmement précis, les erreurs commises peuvent devenir énormes. C'est la raison pour laquelle on préfère calculer ces valeurs typiques à partir des formules suivantes (à vérifier à titre d'exercice) :

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2 \\ m_3 &= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^3 - \frac{3}{n}\bar{x} \cdot \left(\sum_{i=1}^{p} n_i x_i^2\right) + 2\bar{x}^3 \\ m_4 &= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^4 - \frac{4}{n}\bar{x} \cdot \left(\sum_{i=1}^{p} n_i x_i^3\right) + \frac{6}{n}\bar{x}^2 \cdot \left(\sum_{i=1}^{p} n_i x_i^2\right) - 3\bar{x}^4 \end{aligned}$$

En effet, l'utilisation de ces formules nécessitent uniquement le calcul de $\sum_{i=1}^{p} n_i x_i^k$ pour $k = 1,2,3$ et 4.

Il suffira donc de dresser le tableau suivant :

x_i	n_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$	$n_i x_i^4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$S_1 = n\bar{x}$	S_2	S_3	S_4

Ce tableau se remplit très facilement ligne par ligne puisque dans chaque ligne on passe d'un élément au suivant en multipliant par x_i. D'autre part, les sommes des colonnes de ce tableau fournissent les termes nécessaires au calcul de s^2, m_3 et m_4.

b) Afin d'éviter d'introduire des nombres gigantesques dans les calculs, on procède souvent à une translation des données

$$x'_i = x_i - t, \forall i,$$

où t est choisi à peu près au milieu des valeurs observées de telle sorte que les x'_i soient petits. On calcule alors $\bar{x}'$, s'^2, m'_3 et m'_4 comme indiqué dans la remarque précédente, en utilisant les x'_i au lieu des x_i. On retrouve ensuite les valeurs typiques grâce aux relations :

$$\begin{cases} \bar{x} = \bar{x}' + t \\ s^2 = s'^2 \\ m_3 = m'_3 \\ m_4 = m'_4. \end{cases}$$

Si l'on a une distribution groupée avec des classes de même longueur ℓ, on peut aussi poser

$$x'_i = \frac{x_i - t}{\ell}, \forall i,$$

effectuer les calculs à partir des x'_i et en déduire les valeurs typiques grâce aux relations

$$\bar{x} = \ell\bar{x}' + t \qquad m_3 = \ell^3 m'_3$$
$$s^2 = \ell^2 s'^2 \qquad m_4 = \ell^4 m'_4$$

Chapitre 2

Statistique descriptive à deux dimensions

2.1 Les distributions de fréquences

On dispose cette fois de n couples de valeurs observées; la 1ère valeur de chaque couple se rapporte à une variable numérique discrète x dont les valeurs possibles, rangées dans l'ordre croissant, sont x_1,x_2, . . . ,x_p; la 2ème valeur de chaque couple se rapporte à une variable numérique discrète y dont les valeurs possibles, rangées dans l'ordre croissant, sont y_1,y_2, . . . ,y_q.

2.1.1 Définitions

– La *fréquence absolue* d'un couple (x_i,y_j) est le nombre n_{ij} de couples observés égaux à (x_i,y_j);

– La *fréquence relative* du couple (x_i,y_j) est $\frac{n_{ij}}{n}$;

– La *distribution de fréquences* (absolues ou relatives) *à 2 dimensions* d'un couple de variables est un tableau à double entrée dont les lignes correspondent aux valeurs d'une variable, rangées dans l'ordre croissant, dont les colonnes correspondent aux valeurs de l'autre variable, rangées dans l'ordre croissant, et dont les cases contiennent les fréquences (absolues ou relatives) correspondantes.

– La *fréquence marginale absolue* de la valeur x_i (resp. y_j) est la somme des fréquences absolues des couples contenant x_i (resp. y_j), c'est-à-dire la quan-

tité $\sum_{j=1}^{q} n_{ij}$, notée $n_{i.}$ (resp. la quantité $\sum_{i=1}^{p} n_{ij}$, notée $n_{.j}$).

– La *fréquence marginale relative* de la valeur x_i (resp. y_j) est la somme des fréquences relatives des couples contenant x_i (resp. y_j), c'est-à-dire la quantité $\sum_{j=1}^{q} \frac{n_{ij}}{n} = \frac{n_{i.}}{n}$ (resp. la quantité $\sum_{i=1}^{p} \frac{n_{ij}}{n} = \frac{n_{.j}}{n}$).

– La *distribution de fréquences marginales* (absolues ou relatives) de la 1ère (resp. 2ème) variable est la distribution à 1 dimension où à chaque valeur x_i (resp. y_j) est associée sa fréquence marginale (absolue ou relative).

– La *fréquence conditionnelle absolue* de la valeur x_i (resp. y_j) sous la condition y_k (resp. x_ℓ) est égale à n_{ik} (resp. $n_{\ell j}$).

– La *fréquence conditionnelle relative* de la valeur x_i (resp. y_j) sous la condition y_k (resp. x_ℓ) est la quantité $\frac{n_{ik}}{n_{.k}}$, notée n_i/y_k (resp. la quantité $\frac{n_{\ell j}}{n_{\ell .}}$, notée n_j/x_ℓ).

– La *distribution de fréquences conditionnelles* (absolues ou relatives) de la 1ère (resp. 2ème) variable sous la condition y_k (resp. x_ℓ) est la distribution à 1 dimension où à chaque valeur x_i (resp. y_j) est associée sa fréquence conditionnelle (absolue ou relative) sous la condition y_k (resp. x_ℓ).

Remarque : on a les relations suivantes :

$$\begin{array}{ll} \sum_{i=1}^{p} n_{i.} = n & \sum_{i=1}^{p} \frac{n_{i.}}{n} = 1 \\ \sum_{j=1}^{q} n_{.j} = n & \sum_{j=1}^{q} \frac{n_{.j}}{n} = 1 \\ \forall k, \ \sum_{i=1}^{p} n_i/y_k = 1 & \sum_{j=1}^{q} n_j/x_\ell = 1, \forall \ell \end{array}$$

2.1.2 Groupement des données

On peut grouper les valeurs d'une ou des 2 variables en classes; les fréquences définies précédemment se rapportent alors aux classes plutôt qu'aux valeurs, comme pour les distributions à 1 dimension.

2.1.3 Représentations graphiques

Un ensemble de n couples observés peut être représenté par un *nuage de points* dans le plan : il suffit de porter les valeurs de la 1ère variable en abscisse et les valeurs de la 2ème variable en ordonnée (on donne éventuellement aux points des grosseurs proportionnelles à leurs fréquences ou on indique ces fréquences à côté des points).

On peut aussi représenter une distribution à 2 dimensions au moyen de graphiques à 3 dimensions : *diagrammes en bâtons* pour les distributions non groupées et *stéréogrammes* pour les distributions groupées (voir exemple qui suit).

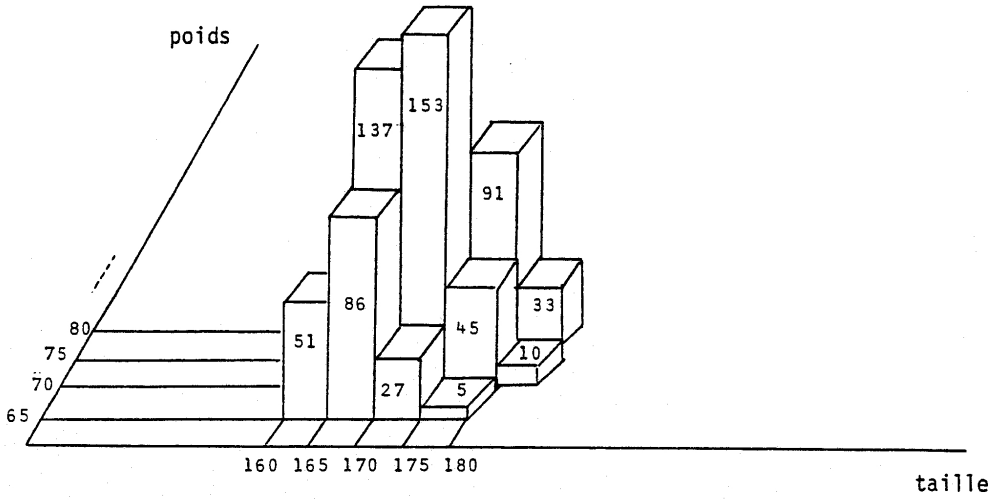
2.1.4 Exemple

On a mesuré le poids et la taille de 1000 individus : le tableau ci-dessous donne la distribution de fréquences absolues (après groupement des données).

Poids / Taille	[65-70)	[70-75)	[75-80)	[80-85)	[85-90)	[90-95)	[95-100)	
[160-165)	51	46	5	4	3			109
[165-170)	86	137	46	11	2			282
[170-175)	27	153	89	25	7			301
[175-180)	5	45	91	40	6			187
[180-185)		10	33	21	16	1	1	82
[185-190)		1	4	11	10	3		29
[190-195)				1	2	4	1	8
[195-200)					1	1		2
	169	392	268	113	47	9	2	1000

La dernière colonne donne la distribution des fréquences marginales absolues des tailles; la dernière ligne donne la distribution des fréquences marginales absolues des poids; en divisant ces chiffres par 1000, on obtient les distributions de fréquences marginales relatives.

La distribution conditionnelle relative des poids des individus dont la taille est comprise entre 170 et 175 cm est donnée par la 3ème ligne du tableau, après division par 301 (somme des éléments de cette ligne).



Partie du stéréogramme représentant la distribution précédente

2.2 Valeurs typiques des distributions à 2 dimensions

2.2.1 Valeurs typiques relatives aux distributions unidimensionnelles

Les distributions de fréquences marginales et conditionnelles sont des distributions à 1 dimension; on peut donc leur associer les valeurs typiques vues dans le chapitre

précédent :

– moyennes marginales :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{P} n_{i.}x_i \ \cdot$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j \cdot$$

– variances marginales :

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(x_i - \bar{x})^2 \ \cdot$$

$$s_2^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(y_j - \bar{y})^2 \cdot$$

– moyenne conditionnelle de la 1ère variable sous la condition y_k :

$$\bar{x}_k = \frac{1}{n_{.k}}\sum_{i=1}^{p} n_{ik}x_i \cdot$$

On obtient, pour l'exemple précédent :

Moyennes marginales : tailles : 172,4 cm
poids : 75 kg.

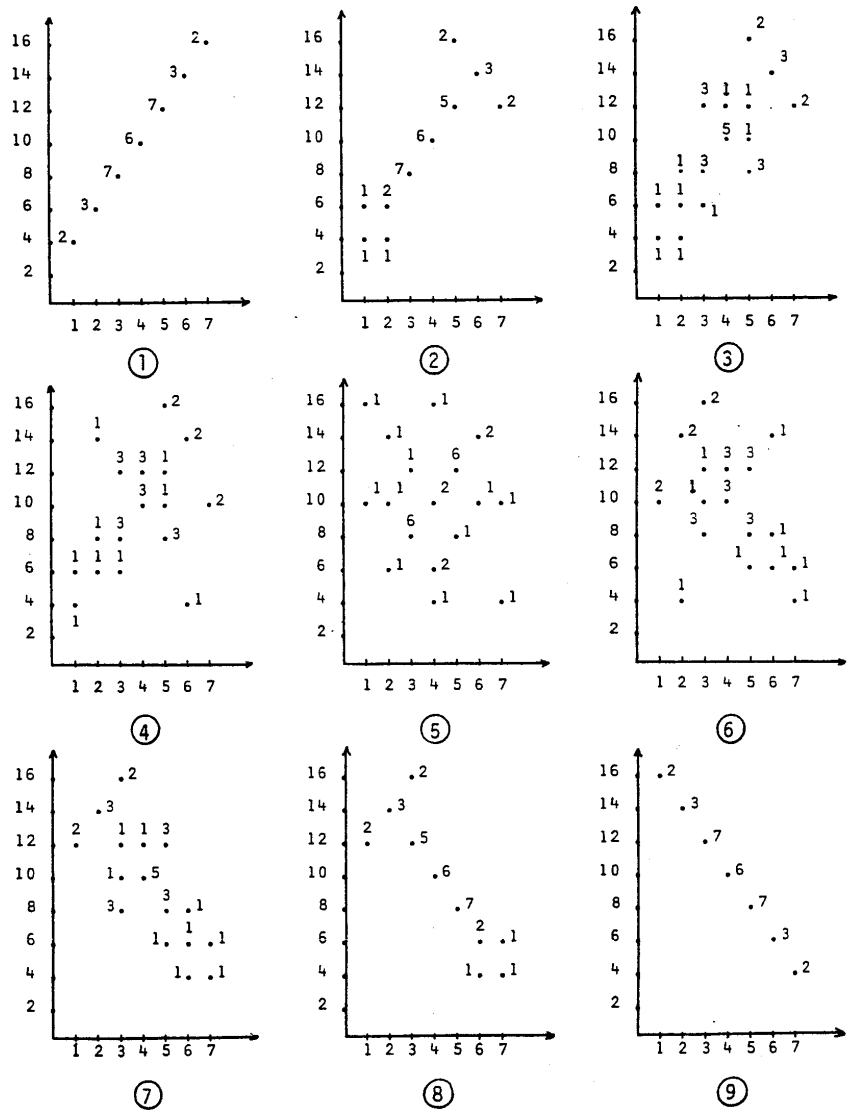
Ecarts-types marginaux : taille : 6,46 cm
poids : 5,58 kg.

Poids moyen des individus mesurant entre 170 et 175 cm (moyenne conditionnelle) : 74,7 kg.

2.2.2 Etude de la dépendance linéaire entre les 2 variables

a) **Introduction**

La connaissance des 2 distributions marginales n'est en général pas suffisante pour connaître entièrement une distribution à 2 dimensions. Les 9 distributions ci-dessous, par exemple, ont les mêmes distributions marginales mais sont complètement différentes (on a indiqué à côté de chaque point, sa fréquence absolue).



Dans chaque cas, la distribution marginale de la 1ère variable (valeurs en abscisse) est donnée par

x_i	1	2	3	4	5	6	7
$n_{i.}$	2	3	7	6	7	3	2

et celle de la 2e variable en ordonnées) par

y_j	4	6	8	10	12	14	16
$n_{.j}$	2	3	7	6	7	3	2

Les moyennes et variances marginales sont données par

$$\bar{x} = 4,\ \bar{y} = 10,\ s_1 = 1{,}57,\ s_2 = 3{,}14$$

(exercice : expliquer pourquoi, dans cet exemple, $s_2 = 2s_1$).

Dans les 3 premiers graphiques, on constate une liaison linéaire positive assez nette : les points sont essentiellement placés au-dessus à droite et en-dessous à gauche du centre de masse de la distribution, qui est ici le point $(4{,}10)$; cela signifie que les points $(x_i{,}y_j)$ sont, pour la plupart, tels que $(x_i - \bar{x})$ et $(y_j - \bar{y})$ ont le même signe, ou encore tels que $(x_i - \bar{x}) \cdot (y_j - \bar{y}) > 0$.

Dans les 3 derniers graphiques, on constate une liaison linéaire négative assez nette : les points sont essentiellement placés au-dessus à gauche et en-dessous à droite du centre de masse de la distribution; cela signifie que les points $(x_i{,}y_j)$ sont, pour la plupart, tels que $(x_i - \bar{x})$ et $(y_j - \bar{y})$ ont des signes contraires, ou encore tels que $(x_i - \bar{x}) \cdot (y_j - \bar{y}) < 0$.

Dans les 3 graphiques du milieu, les points sont éparpillés de sorte que les $(x_i - \bar{x}) \cdot (y_j - \bar{y}) < 0$ compensent plus ou moins les $(x_i - \bar{x}) \cdot (y_j - \bar{y}) > 0$.

Pour mesurer le lien linéaire entre les 2 variables, on va donc se baser sur la quantité $\sum_i \sum_j n_{ij}(x_i - \bar{x})(y_j - \bar{y})$, qui sera fortement positive dans les 3 premiers cas, fortement négative dans les 3 derniers cas et proche de 0 dans les 3 cas intermédiaires.

b) **La covariance**

Définition :

$$m_{11} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$$

Propriétés :

1) $\left.\begin{array}{l} x'_i = ax_i + b, \forall i \\ y'_j = cy_j + d, \forall j \end{array}\right\} \Rightarrow m'_{11} = acm_{11};$

2) $|m_{11}| \leq s_1 s_2$

Démonstration de 2)

$\forall u \in \mathbb{R}$, on a

$$\frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} \left[u(x_i - \bar{x}) + (y_j - \bar{y})\right]^2 \geq 0,$$

ce qui peut encore s'écrire

$$u^2 s_1^2 + 2um_{11} + s_2^2 \geq 0, \forall u \in \mathbb{R}.$$

Cette relation implique que

$$m_{11}^2 - s_1^2 s_2^2 \leq 0,$$

d'où l'inégalité annoncée. (C.Q.F.D.)

3) $|m_{11}| = s_1 s_2$ *si et seulement si les points observés se trouvent sur une droite non parallèle aux axes. (On suppose* s_1 *et* $s_2 \neq 0$*).*

Démonstration.

Si $|m_{11}| = s_1 s_2$, alors il existe une valeur $u_0 \neq 0$ telle que

$$u_0^2 s_1^2 + 2u_0 m_{11} + s_2^2 = 0,$$

ce qui implique (cf. démonstration précédente)

$$u_0(x_i - \bar{x}) + (y_j - \bar{y}) = 0 \quad , \qquad \forall i,j \text{ tels que } n_{ij} \neq 0 :$$

cela signifie que tous les couples (x_i,y_j) observés au moins une fois sont sur une droite. Inversément, si tous les points observés sont sur la droite

$$ax + by + c = 0 \quad , \quad (a \text{ et } b \neq 0)$$

alors on a

$$ax_i + by_j + c = 0, \quad \forall i,j \text{ tels que } n_{ij} \neq 0.$$

En multipliant par $\frac{n_{ij}}{n}$ et en sommant sur i et j, on obtient

$$a\bar{x} + b\bar{y} + c = 0,$$

ce qui signifie que le centre de masse de la distribution est aussi sur la droite.

On a donc

$$a(x_i - \bar{x}) + b(y_j - \bar{y}) = 0, \quad \forall i,j \text{ tels que } n_{ij} \neq 0,$$

ou encore, en posant $u_0 = \frac{a}{b}(\neq 0)$,

$$u_0(x_i - \bar{x}) + (y_j - \bar{y}) = 0, \forall i,j \text{ tels que } n_{ij} \neq 0.$$

En se rapportant à la démonstration de la propriété 2), on en tire

$$u_0^2 + s_1^2 + 2u_0 m_{11} + s_2^2 = 0,$$

d'où

$$|m_{11}| = s_1 s_2.$$

(C.Q.F.D.)

c) **Le coefficient de corrélation**

Définition :

$$r = \frac{m_{11}}{s_1 \cdot s_2} \qquad (s_1 \text{ et } s_2 \text{ supposés } \neq 0)$$

Propriétés :

1) r est un nombre sans dimension;

2) $\left.\begin{array}{ll} x'_i = ax_i + b, & \forall i \\ y'_j = cy_j + d, & \forall j \end{array}\right\} \Rightarrow r' = r;$

3) $-1 \leq r \leq 1$

4) $|r| = 1$ *si et seulement si les points observés se trouvent sur une droite non parallèle aux axes*. (Ces propriétés résultent immédiatement de celles de la covariance).

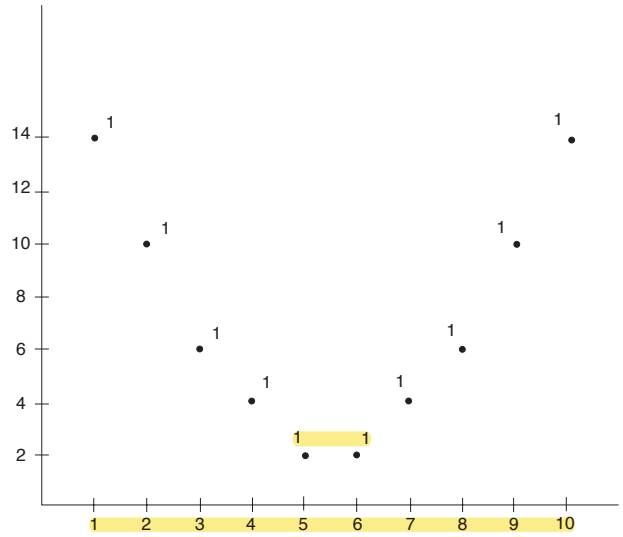
d) **Remarques**

1) En pratique, on utilisera plutôt le coefficient de corrélation qui a l'avantage d'être un nombre sans dimension, indépendant des unités utilisées pour mesurer les variables et qui est toujours compris entre -1 et $+1$.
Si r est proche de 1, c'est qu'il existe une dépendance linéaire positive entre les 2 variables; s'il est proche de -1, il existe une dépendance linéaire négative; s'il est proche de 0, il n'y a pas de lien linéaire entre les 2 variables.
Dans les exemples précédents, on a respectivement

	1	2	3	4	5	6	7	8	9
r=	1	0,88	0,69	0,34	0	-0,34	-0,69	-0,88	-1

2) Nous supposons ici que la distribution à 2 dimensions n'est pas dégénérée, c'est-à-dire que les points ne sont pas tous situés sur une parallèle à un axe. Cela signifierait en effet qu'une des variables a la même valeur pour toutes les observations. Dans ce cas, la variance de cette variable et la covariance des 2 variables sont nulles, le coefficient de corrélation n'étant pas défini.

3) Le coefficient de corrélation ne concerne que l'éventuel lien *linéaire* entre 2 variables. Dans l'exemple ci-dessous, le coefficient de corrélation est nul (à vérifier) alors qu'il existe manifestement un lien (non linéaire) entre les 2 variables.



Moyennant une transformation (logarithme, racine carrée, fonction trigonométrique,...) d'une variable ou des 2 variables, on pourra se servir du coefficient de corrélation pour établir un lien non linéaire (exponentiel, quadratique, trigonométrique,...) entre les 2 variables. On pourra aussi utiliser le "rapport de corrélation" (cf. plus loin).

4) Le fait qu'il existe un lien (linéaire ou non) entre 2 variables n'implique pas nécessairement qu'il y ait relation de cause à effet entre elles. Elles peuvent simplement dépendre d'une même 3ème variable (il n'existe par de lien de cause à effet entre l'augmentation du prix du mazout de chauffage et l'augmentation du prix de l'essence pour voiture : les deux résultent de l'augmentation du prix du pétrole), ou même être totalement indépendantes (taille d'un enfant et taille d'un arbuste au cours du temps). L'interprétation du coefficient de corrélation (et des autres valeurs typiques aussi d'ailleurs) doit toujours tenir compte du problème concret considéré et reposer sur un minimum d'intuition et de bon sens.

5) A partir de quelle valeur estime-t-on que r est proche de 1, de -1 ou de 0 ? Pour répondre à cette question, il faudrait connaître le "comportement habituel de r". S'il est "rarement" supérieur à 0,8 (par exemple), on dira que 0,8 est une valeur significativement grande. C'est le calcul des probabilités

et l'inférence statistique qui nous permettront d'étudier cette question de façon plus précise.

6) Comme pour les valeurs typiques unidimensionnelles, on simplifie le calcul de la covariance en utilisant la formule (à vérifier à titre d'exercice)

$$m_{11} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \bar{x}\cdot\bar{y}$$

Si l'on effectue des translations

$$\begin{cases} x'_i = x_i - t_1, \forall i \\ y'_j = y_j - t_2, \forall j, \end{cases}$$

on aura

$$m_{11} = m'_{11}.$$

Si l'on pose

$$\begin{cases} x'_i = \dfrac{x_i - t_1}{\ell_1}, & \forall i, \\ y'_j = \dfrac{y_j - t_2}{\ell_2}, & \forall j, \end{cases}$$

on aura

$$m_{ll} = \ell_1\ell_2 m'_{11}.$$

e) **Les droites de régression**

1) *Introduction*

Lorsqu'on s'attend à ce que la variable y dépende plus ou moins linéairement de la variable x, on peut rechercher ce lien linéaire en déterminant la droite qui "ajuste" au mieux le nuage des points observés, les écarts des points à la droite étant considérés parallèlement à l'axe y : on obtient ainsi la droite de régression de y en x. Si c'est x qui dépend de y, on s'intéresse à la droite de régression de x en y, en considérant les écarts parallèlement à l'axe des x. Si aucune des variables n'a un rôle privilégié, on peut calculer la droite qui minimise la somme des carrés des distances des points observés à cette droite, les distances étant prises perpendiculairement à la

droite. Cette droite est en général peu utilisée car les calculs pour l'obtenir sont assez longs.
D'autre part, la connaissance des 2 droites de régression fournit une bonne information sur la disposition du nuage de points.

2) *Droites de régression de y en x et de x en y*

La *droite de régression de y* en x est la droite qui minimise la somme des carrés des écarts des points observés à cette droite, les écarts étant pris parallèlement à l'axe y.
C'est donc la droite d'équation $y = ax + b$ qui minimise la quantité

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}[y_j - ax_i - b]^2.$$

En annulant les dérivées par rapport à a et b, on obtient le système d'équations

$$\begin{cases} a\sum_{i=1}^{p} n_{i.}x_i^2 + b\sum_{i=1}^{p} n_{i.}x_i = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_iy_j \\ an\bar{x} + bn = n\bar{y} \end{cases}$$

En multipliant la 2ème équation par $\bar{x}$ et en la soustrayant de la 1ère, on obtient

$$a = \frac{m_{11}}{s_1^2} \text{ et } b = \bar{y} - \frac{m_{11}}{s_1^2}\bar{x}.$$

L'équation de la droite de régression de y en x est donc

$$y = \frac{m_{11}}{s_1^2}(x - \bar{x}) + \bar{y};$$

$\frac{m_{11}}{s_1^2}\left(= r\frac{s_2}{s_1}\right)$ est appelé *coefficient de régression de y en x* et est noté b_{21}.

La droite de régression de x en y est la droite qui minimise la somme des carrés des écarts des points observés à cette droite, les écarts étant pris parallèlement à l'axe x. Son équation est

$$x = \frac{m_{11}}{s_2^2}(y - \bar{y}) + \bar{x};$$

$\frac{m_{11}}{s_2^2}\left(= r\frac{s_1}{s_2}\right)$ est appelé *coefficient de régression de x en y* et est noté b_{12}.

Propriétés des droites de régression

- Les deux droites de régression passent par le point $(\bar{x},\bar{y})$.
- Les droites de régression de y en x et de x en y ont comme coefficient angulaire respectivement $r\frac{s_2}{s_1}$ et $\frac{1}{r}\frac{s_2}{s_1}$.

On en déduit qu'elles sont toujours inclinées dans le même sens (coefficients angulaires de même signe), que l'une est toujours "plus verticale" que l'autre (car $|r| \leq \frac{1}{|r|}$), qu'elles sont confondues ssi tous les points observés sont sur une droite non parallèle aux axes (ssi $r = \pm 1$) et qu'elles sont perpendiculaires ssi $r = 0$.

3) *Variances résiduelles*

La *variance résiduelle de y en x* est la quantité

$$s_{21}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left[y_j - ax_i - b\right]^2,$$

où l'on remplace a et b par les valeurs obtenues plus haut.
On obtient :

$$\begin{aligned} s_{21}^2 &= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left[(y_j - \bar{y}) - \frac{m_{11}}{s_1^2}(x_i - \bar{x})\right]^2 \\ &= s_2^2 - 2\frac{m_{11}^2}{s_1^2} + \frac{m_{11}^2}{s_1^2} = s_2^2(1 - r^2). \end{aligned}$$

De même, la *variance résiduelle de x en y* vaut

$$s_{12}^2 = s_1^2(1 - r^2).$$

Propriétés et interprétation

- $s_{21}^2 = 0$ ssi $r = \pm 1$ ssi tous les points observés sont sur une droite.

- $s_{21}^2 = s_2^2$ ssi $r = 0$ ssi les droites de régression sont parallèles aux axes.
- $s_2^2 r^2$ représente une certaine proportion de s_2^2, d'autant plus grande que la dépendance linéaire entre x et y est forte : on peut donc la considérer comme la part de la variance de y qui est expliquée par le lien linéaire entre x et y, tandis que la variance résiduelle s_{21}^2 est la part de la variance de y qui n'est pas expliquée par ce lien linéaire (d'où son nom).

4) *Droite de régression orthogonale et droite des moindres rectangles*

La *droite de régression orthogonale* est celle qui minimise la somme des carrés des écarts des points observés à cette droite, les écarts étant pris perpendiculairement à la droite. On peut démontrer qu'elle a pour équation :

$$y = c(x - \bar{x}) + \bar{y}$$

où

$$c = \frac{2m_{11}}{s_1^2 - s_2^2 + \sqrt{(s_1^2 - s_2^2)^2 + 4m_{11}^2}}\,.$$

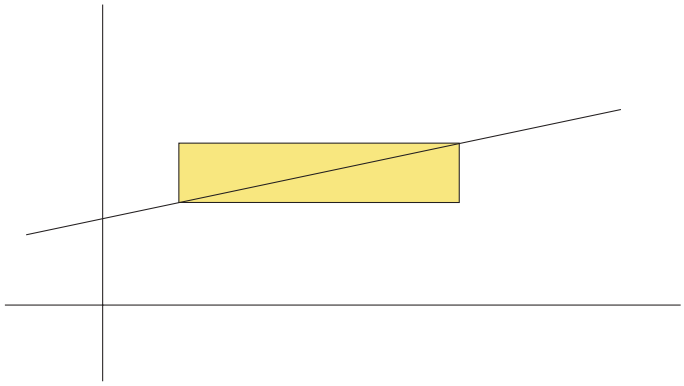
On voit que l'inclinaison de cette droite dépend très fort de l'écart $s_1^2 - s_2^2$. En particulier, si $s_1^2 = s_2^2$, on obtient $c = \pm 1$. Pour éviter cet inconvénient, on peut travailler à l'aide des variables réduites en remplaçant x_i et y_j par

$$\begin{aligned} x_i' &= \frac{x_i - \bar{x}}{s_1}, \\ y_j' &= \frac{y_j - \bar{y}}{s_2}. \end{aligned}$$

On obtient alors la droite d'équation

$$y = \pm\frac{s_2}{s_1}(x - \bar{x}) + \bar{y},$$

le signe à adopter étant celui de la covariance.
Cette dernière est appelée *droite des moindres rectangles* car c'est celle que l'on obtient en minimisant la somme des aires des rectangles définis par les points observés et cette droite, comme illustré dans la figure ci-dessous.



Cette droite est intermédiaire entre les droites de régression de y en x et de x en y : son coefficient angulaire est, au signe près, la moyenne géométrique des coefficients angulaires des 2 autres droites.

2.2.3 Les moments

Le moment d'ordre (k,ℓ) par rapport au point (c,d) est la quantité

$$\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - c)^k (y_j - d)^\ell.$$

En choisissant $c = \bar{x}$ et $d = \bar{y}$, on obtient les moments centrés

$$m_{k\ell} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \bar{x})^k (y_j - \bar{y})^\ell.$$

En particulier : $m_{10} = m_{01} = 0$,

m_{11} est la covariance

$m_{20} = s_1^2$; $m_{02} = s_2^2$

2.2.4 Etude de la dépendance non linéaire entre les 2 variables

a) **Définitions**

La moyenne conditionnelle de la 2ème variable sous la condition x_ℓ est, par définition de la moyenne d'une distribution unidimensionnelle :

$$\bar{y}_\ell = \sum_{j=l}^{q} \frac{n_{\ell j}}{n_{\ell .}} y_j$$

Le *diagramme de régression de y en x* est l'ensemble des points $(x_\ell,\bar{y}_\ell)$. Le *rapport de corrélation de y en x* est la quantité r_{yx} telle que

$$r_{yx}^2 = \frac{\sum_{\ell=1}^{p} n_{\ell.}(\bar{y}_\ell - \bar{y})^2}{n\, s_2^2}$$

On définit de façon analogue le diagramme de régression de x en y et le rapport de corrélation de x en y.

b) **Propriétés**

1) $r_{yx} = 0$ ssi le diagramme de régression est une droite parallèle à l'axe des x (démonstration immédiate). Dans ce cas, toutes les moyennes conditionnelles $\bar{y}_\ell$ sont égales entre elles, c'est-à-dire ne dépendent pas de x_ℓ.

2) $r_{yx}^2 \leq 1$.
En effet,

$$\begin{aligned}
n s_2^2 &= \sum_{j=1}^{q} n_{.j}(y_j - \bar{y})^2 \\
&= \sum_{\ell=1}^{p}\sum_{j=1}^{q} n_{\ell j}\left[(y_j - \bar{y}_\ell) + (\bar{y}_\ell - \bar{y})\right]^2 \\
&= \sum_{\ell=1}^{p}\sum_{j=1}^{q} n_{\ell j}(y_j - \bar{y}_\ell)^2 + \sum_{\ell=1}^{p} n_{\ell.}(\bar{y}_\ell - \bar{y})^2 \qquad \text{(justifier)} \\
&\geq \sum_{\ell=1}^{p} n_{\ell.}(\bar{y}_\ell - \bar{y})^2
\end{aligned}$$

3) $r_{yx}^2 = 1$ ssi $\forall \ell,j : n_{\ell j} \neq 0 \Rightarrow y_j = \bar{y}_\ell$
(résulte de la démonstration précédente).
Dans ce cas, pour chaque x_ℓ, les valeurs observées y_j sont confondues avec $\bar{y}_\ell$; autrement dit, tous les points observés sont sur le diagramme de régression : la 2ème variable dépend totalement de la 1ère.

Le rapport de corrélation de y en x mesure donc le degré de dépendance (non nécessairement linéaire) de y en x.

c) **Régression curvilinéaire.**

Lorsqu'une dépendance non linéaire existe entre x et y, on peut être intéressé par sa forme analytique : c'est l'objet de la régression curvilinéaire ou "ajustement d'une courbe à un nuage de points".
Lorsqu'une courbe doit être ajustée, deux problèmes se posent : le choix de l'équation de la courbe et la détermination des paramètres intervenant dans cette équation.

- Le choix de l'équation de la courbe se fera si possible sur base de considérations théoriques (d'après la connaissance qu'on a du phénomène étudié); sinon, suivant l'allure du graphique et après plusieurs essais éventuels.
- La détermination des paramètres se fait en général par la méthode des moindres carrés.

Par exemple, l'ajustement au moyen d'un polynôme

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_p x^p$$

se fera en recherchant les a_k qui minimisent

$$\frac{1}{n} \sum_i \sum_j n_{ij} (y_j - a_0 - a_1\, x_i - \ldots - a_p x_i^p)^2.$$

Certains ajustements peuvent se ramener à un ajustement linéaire ou polynomial par changements de variables, en travaillant à l'aide de graphiques logarithmiques ou semi-logarithmiques.

D'autres ajustements, par contre, se font par des techniques particulières liées à la forme analytique de la courbe choisie.

Exemples de courbes souvent utilisées dans les applications :

Courbe exponentielle : $y = ae^{bx}$ (ex. : développement d'une population dont le taux de croissance est constant).

Courbe logistique : $y = \dfrac{a}{1 + bc^x}$ (ex. : développement d'une population dont le taux de croissance est proportionnel à l'écart entre la population existante et un maximum que cette population ne peut dépasser).

Courbe de Gompertz : $y = e^{a+bc^x}$ (utilisée en matière d'assurances et dans le domaine économique).

Deuxième partie

La théorie des probabilités

Introduction

La statistique descriptive permet, nous l'avons vu, de mettre de l'ordre dans des ensembles de données afin de mettre en évidence des propriétés particulières des phénomènes étudiés. Les propriétés ainsi trouvées ne concernent évidemment que les données observées. En déduire des lois générales, des décisions ou des prévisions présente un danger d'erreur certain : cela ne peut être fait que de façon qualitative, imprécise et avec un maximum de précautions. Si l'on veut quantifier les propriétés intéressantes et connaître les risques d'erreurs que l'on commet, il faut faire appel à des modèles mathématiques. C'est la théorie des probabilités qui sert de base aux modèles utilisés en statistique. Les variables aléatoires et leurs distributions de probabilité seront les modèles théoriques représentant les variables observées et leurs distributions de fréquences.

Il faut noter que la théorie des probabilités est utilisée dans des domaines autres que la statistique comme par exemple la mécanique quantique, la génétique ou l'actuariat. Il est important de savoir également que les phénomènes d'imprécision, d'incertitude ou d'ignorance (nécessairement présents dans toutes les activités humaines) ne sont pas tous modélisables au moyen de la théorie des probabilités. D'autres théories ont été récemment développées pour combler cette lacune (théorie des possibilités, des fonctions de croyance, des ensembles flous,...).

Chapitre 1

La notion de probabilité et ses propriétés

1.1 Expérience et événement aléatoires

La définition de la probabilité est liée aux notions d'expérience et d'événement aléatoires.
Une *expérience aléatoire* est une expérience dont on ne peut prévoir exactement le *résultat* (ex. : jet d'un dé).
Un *événement* est aléatoire si l'on ne peut prévoir s'il se réalisera ou non au cours de l'expérience aléatoire (ex. : obtenir un chiffre pair). Un événement donné se réalise au cours d'une expérience si le résultat de cette expérience est conforme à cet événement (ex. : l'événement "obtenir un chiffre pair" se réalise si le résultat de l'expérience "jet du dé" est 2 ou 4 ou 6). Dorénavant, on assimilera tout événement à l'ensemble des résultats pour lesquels il se réalise (ex. : "obtenir un chiffre pair" $= \{2,4,6\}$) et on définira un *événement* comme un *sous-ensemble de l'ensemble E de tous les résultats possibles de l'expérience* et on dira que l'événement "se réalise" au cours de l'expérience si le résultat de l'expérience appartient à l'événement.

L'"événement impossible" est celui qui ne se réalise jamais (ex. : obtenir 7 lors du jet d'un dé) : il sera assimilé à l'ensemble vide $\emptyset$ (qui est bien un sous-ensemble de E).

L'"événement certain" est celui qui se réalise toujours (ex. : obtenir un chiffre inférieur à 7 lors du jet d'un dé) : il sera assimilé à l'ensemble E (qui est bien un sous-ensemble de E).

"La réunion de 2 événements" sera l'événement assimilé à la réunion des 2 sous-ensembles correspondants. (ex. : obtenir un chiffre strictement inférieur à 3 *ou* obtenir

un chiffre strictement supérieur à $4 = \{1,2\} \cup \{5,6\} = \{1,2,5,6\}$).

"L'intersection de 2 événements" sera l'événement assimilé à l'intersection des 2 sous-ensembles correspondants. (ex. : obtenir un chiffre pair *et* obtenir un chiffre strictement supérieur à $3 = \{2,4,6\} \cap \{4,5,6\} = \{4,6\}$).

1.2 Probabilité d'un événement aléatoire

Supposons que l'on répète n fois une expérience aléatoire et qu'un événement A se réalise f_A fois. On dira que f_A est la *fréquence absolue* de A et que $\frac{f_A}{n}$ est sa *fréquence relative* .

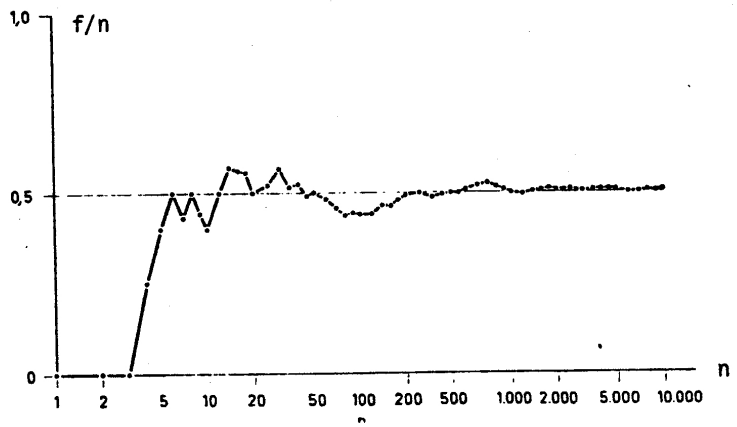
L'introduction du concept de probabilité résulte de l'observation expérimentale suivante : si l'on répète un très grand nombre de fois une expérience, la fréquence relative d'un événement tend à se stabiliser, à converger vers une valeur; c'est le phénomène de *régularité statistique*, qui est à la base du *postulat* suivant : *à tout événement on peut associer un nombre vers lequel converge sa fréquence relative; ce nombre est appelé probabilité mathématique de l'événement considéré*. Nous noterons $P(A)$ la probabilité de l'événement A.

On peut se poser la question suivante : étant donné une expérience et un événement qui lui est associé, comment peut-on déterminer la probabilité de cet événement ?
Dans certains problèmes, on rencontrera des probabilités qui peuvent être déterminées a priori, sans faire appel aux fréquences observées. Tel est le cas par exemple des jeux de hasard (non truqués) où les résultats ont tous les mêmes chances de se produire (loteries, dés, cartes, jeux de casino,...) : la probabilité d'un événement est alors simplement le nombre de résultats pour lesquels il se réalise, divisé par le nombre total de résultats possibles. Dans d'autres problèmes, il faudra estimer la probabilité a posteriori à partir de l'étude des fréquences de l'événement (on verra plus loin comment se fait cette estimation).
Même lorsqu'une probabilité a priori peut être calculée, la méthode a posteriori est supérieure car la plus proche de la réalité (un dé n'est jamais parfaitement symétrique), mais elle nécessite la réalisation d'une série d'expériences.

Comme illustration du phénomène de régularité statistique, le tableau suivant, qui

résulte d'une expérience effectivement réalisée, donne les fréquences absolues et relatives de l'apparition du côté "face" au cours de dix mille jets d'une pièce de monnaie (résultats publiés par KERRICH en 1946).

n	f	f/n	n	f	f/n	n	f	f/n
1	0	0,000	60	29	0,483	1.200	596	0,497
2	0	0,000	70	32	0,457	1.400	704	0,503
3	0	0,000	80	35	0,438	1.600	810	0,506
4	1	0,250	90	40	0,444	1.800	918	0,510
5	2	0,400	100	44	0,440	2.000	1.013	0,506
6	3	0,500						
7	3	0,429	120	53	0,442	2.500	1.272	0,509
8	4	0,500	140	65	0,464	3.000	1.510	0,503
9	4	0,444	160	74	0,462	3.500	1.772	0,506
10	4	0,400	180	86	0,478	4.000	2.029	0,507
			200	98	0,490	4.500	2.293	0,510
12	6	0,500				5.000	2.533	0,507
14	8	0,571	250	125	0,500			
16	9	0,562	300	146	0,487	6.000	3.009	0,502
18	10	0,556	350	173	0,494	7.000	3.516	0,502
20	10	0,500	400	199	0,498	8.000	4.034	0,504
			450	226	0,502	9.000	4.538	0,504
25	13	0,520	500	255	0,510	10.000	5.067	0,507
30	17	0,567				⋮	⋮	⋮
35	18	0,514	600	312	0,520			
40	21	0,525	700	368	0,526			
45	22	0,489	800	413	0,516			
50	25	0,500	900	458	0,509			
			1000	502	0,502			



1.3 Axiomes de la théorie des probabilités

Après avoir postulé l'existence d'une probabilité associée à chaque événement, on introduit un certain nombre d'axiomes qui vont permettre de travailler avec ce nouveau concept. Ces axiomes ne sont pas choisis au hasard : puisque la probabilité est une fréquence relative idéalisée, on lui donne les mêmes propriétés que les fréquences relatives.

Soit E l'ensemble des résultats liés à une expérience, n le nombre de répétitions de cette expérience, A,B,C des événements associés à l'expérience (des sous-ensembles de E) et f_A,f_B,f_C les fréquences absolues de ces événements. Il est clair que (propriétés aisément démontrables):

$$0 \leq \frac{f_A}{n} \leq 1, \quad \forall A \subset E,$$

$$\frac{f_A}{n} = \frac{n}{n} = 1 \text{ dès que } A = E,$$

$$f_A = 0 \text{ lorsque } A = \emptyset,$$

$$f_A \leq f_B \text{ lorsque } A \subset B,$$

$$f_A + f_{A*} = n, \text{ où } A* = E \setminus A$$

$$f_{A \cup B} = f_A + f_B - f_{A \cap B}$$

$$\ldots$$

Les axiomes des probabilités se déduisent de ces propriétés en remplaçant $\frac{f_A}{n}$ par $P(A)$, ce qui donne notamment :

$0 \leq P(A) \leq 1, \forall A \subset E;$
$P(E) = 1,$
$P(\emptyset) = 0,$
$A \subset B \Rightarrow P(A) \leq P(B),$
$P(A) + P(A*) = 1, (A* = E \setminus A),$
$P(A \cup B) = P(A) + P(B) - P(A \cap B),$
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C)$
$+P(A \cap B \cap C),$
$A_i \cap A_j = \emptyset \ \forall i,j \Rightarrow P(A_1 \cup \ldots \cup A_m) = \sum_{i=1}^{m} P(A_i),$
$\ldots$

Toutes ces propriétés peuvent en fait se déduire des 3 axiomes suivants, que nous appellerons *axiomes de la théorie des probabilités* :

$$\begin{cases} P(A) \geq 0, \forall A \subset E, \\ P(E) = 1, \\ A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B). \end{cases}$$

Lorsque $A \cap B = \emptyset$, on dit que A et B sont *incompatibles* (ils ne se réalisent jamais simultanément).

Remarques

1. Si $E = \{e_1, e_2, \ldots, e_m\}$ est l'ensemble des résultats possibles d'une expérience et si tous ces résultats ont les mêmes chances de se présenter (jeu de hasard "idéal"), alors il résulte des deuxième et troisième axiomes que

$$P(\{e_i\}) = \frac{1}{m}, \quad \forall i$$

et si $A = \{e_{i_1}, e_{i_2}, \ldots, e_{i_k}\}$, alors le troisième axiome donne

$$P(A) = \frac{k}{m}.$$

Dans ce cas, on obtient la probabilité de A en divisant le nombre de résultats appartenant à A par le nombre total de résultats possibles.

Exemple : jet d'un dé $\rightarrow E = \{1,2,3,4,5,6\}$

$A =$ "obtenir un chiffre pair" $= \{2,4,6\}$

$\rightarrow P(A) = \dfrac{3}{6} = \dfrac{1}{2}$.

2. Il peut arriver qu'un événement, tout en étant possible, ait une fréquence relative qui tend vers 0 (ex. : "retomber sur la tranche" pour une pièce de monnaie jetée en l'air). Cet événement a donc une probabilité nulle, bien qu'il ne soit pas impossible : on parle alors d'événement *stochastiquement impossible* . De même un événement non certain mais de probabilité 1 est dit *stochastiquement certain* .

1.4 Probabilité conditionnelle et indépendance

Soit E l'ensemble des résultats liés à une expérience, n le nombre de répétitions de cette expérience, A et B deux événements associés à l'expérience et f_A,f_B les fréquences absolues de ces événements.

La fréquence conditionnelle relative de A sous la condition B s'obtient en ne considérant que les répétitions de l'expérience où B s'est réalisé et en comptant, *parmi celles-là*, le nombre de cas où A s'est réalisé aussi :

$$f_{A/B} = \frac{f_{A\cap B}}{f_B}\,.$$

Par analogie, on définira la *probabilité conditionnelle de A sous la condition B* par la quantité

$$P(A/B) = \frac{P(A \cap B)}{P(B)},$$

$P(B)$ étant supposée $\neq 0$. Cette quantité est bien comprise entre 0 et 1 (à vérifier).

La probabilité conditionnelle de B sous la condition A sera

$$P(B/A) = \frac{P(A \cap B)}{P(A)},$$

$P(A)$ étant supposée $\neq 0$.

On dira que *A est indépendant de B* ssi

$$P(A/B) = P(A).$$

On constate que A est indépendant de B ssi

$$P(A \cap B) = P(A) \cdot P(B)$$

ssi

$$P(B/A) = P(B)$$

ssi B *est indépendant de* A.

En conclusion, on dira que A et B *sont indépendants* ssi

$$P(A/B) = P(A)$$

ssi

$$P(B/A) = P(B)$$

ssi

$$P(A \cap B) = P(A) \cdot P(B).$$

Plus généralement, m événements A_1,A_2, . . . ,A_m sont indépendants si, pour tout ensemble formé de k de ces événements :

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \ldots P(A_{i_k}).$$

Remarques

1. L'indépendance 2 à 2 de m événements ne suffit pas pour avoir l'indépendance globale de ces événements (voir exercices).
2. Il ne faut pas confondre la probabilité conditionnelle de A sous la condition B $(P(A/B))$ avec la probabilité que A et B se réalisent simultanément $P(A \cap B))$. Dans le 1$^{\text{ier}}$ cas, on se place dans l'hypothèse que B *s'est réalisé*; dans le 2$^{\text{ème}}$ cas, on ne fait aucune hypothèse.

Exemple : jet d'un dé

$$\begin{aligned}
A &= \text{“obtenir un chiffre strictement supérieur à 3”}\\
B &= \text{“obtenir un chiffre pair”}\\
A/B &= \text{“obtenir un chiffre strictement supérieur à 3 sous la condition qu'on a un chiffre pair”}\\
&= \text{“obtenir un chiffre strictement supérieur à 3 } \textit{sachant} \text{ qu'on a obtenu un chiffre pair”}\\
A \cap B &= \text{“obtenir un chiffre strictement supérieur à 3 } \textit{et} \text{ obtenir un chiffre pair”}\\
&= \text{“obtenir 4 ou 6”}\\
P(A) &= 1/2\\
P(B) &= 1/2\\
P(A/B) &= 2/3\\
P(A \cap B) &= 1/3
\end{aligned}$$

On a bien

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

3. Il ne faut pas confondre *événements indépendants* et *événements incompatibles*. (Démontrer à titre d'exercice, que si deux événements de probabilités non nulles sont incompatibles, alors ils ne peuvent pas être indépendants).

1.5 La formule de Bayes

Soit A_1,A_2,...,A_m et B des événements liés à une expérience tels que

$$\begin{cases} B \subset A_1 \cup A_2 \cup \ldots \cup A_m, \\ A_i \cap A_j = \emptyset, \forall i,j, \\ P(B) \neq 0. \end{cases}$$

Il en résulte que

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_m \cap B)$$

et que

$$(A_i \cap B) \cap (A_j \cap B) = \emptyset, \forall i,j (i \neq j).$$

Par conséquent,

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) + \ldots + P(A_m \cap B). \\ &= P(B/A_1) \cdot P(A_1) + P(B/A_2) \cdot P(A_2) + \ldots P(B/A_m) \cdot P(A_m) \end{aligned}$$

et donc, pour k fixé,

$$P(A_k/B) = \frac{P(B/A_k) \cdot P(A_k)}{\sum_{j=1}^{m} P(B/A_j) \cdot P(A_j)}.$$

C'est la formule de Bayes, fondamentale notamment en théorie de la décision. Un exemple classique de situation où cette formule s'applique est le suivant. Supposons que l'événement B ne puisse se réaliser que si l'un des événements A_1,A_2, $\ldots$,A_m se produit (c'est ce qu'exprime l'hypothèse $B \subset A_1 \cup A_2 \cup \ldots \cup A_m$) et que les A_i soient incompatibles 2 à 2 ($A_i \cap A_j = \emptyset, \forall i,j,i \neq j$).

On connaît la probabilité que B se réalise sous la condition $A_i(P(B/A_i))$, $\forall i$ et on connaît $P(A_i)$, $\forall i$ (probabilités "a priori" des A_i). Au cours d'une expérience, B se réalise. Cette information supplémentaire permet de revoir la propbabilité de chaque A_i et de déterminer sa probabilité "a posteriori" : $P(A_i/B)$.

Exemples

A) Trois machines produisent respectivement 50, 30 et 20% du nombre total de pièces fabriquées dans une usine. Les probabilités que ces machines fabriquent une pièce mauvaise sont respectivement $0{,}03$,$0{,}04$ et $0{,}05$. En prélevant une pièce au hasard, on observe qu'elle est défectueuse. Quelle est la probabilité qu'elle ait été fabriquée par la 1ère machine ?

$$\begin{aligned} A_i &= \text{"la pièce a été fabriquée par la i}^{\text{ème}}\text{ machine"} \quad i = 1{,}2{,}3. \\ B &= \text{"la pièce est défectueuse"} \\ P(A_1) &= 0{,}5 \qquad P(A_2) = 0{,}3 \qquad P(A_3) = 0{,}2 \\ P(B/A_1) &= 0{,}03 \qquad P(B/A_2) = 0{,}04 \qquad P(B/A_3) = 0{,}05. \\ P(A_1/B) &= \frac{0{,}03 \times 0{,}5}{0{,}03 \times 0{,}5 + 0{,}04 \times 0{,}3 + 0{,}05 \times 0{,}2} = 0{,}405 \end{aligned}$$

B) Vous avez une chance sur 1000 d'avoir la maladie M. Vous passez un test de dépistage qui s'avère positif. On sait que le test est positif dans 99% des cas pour les personnes malades et dans 2% des cas pour les personnes non malades. Quelle est la probabilité que vous ayez la maladie ?
(Réponse : moins de 5% ! Courage !)

Chapitre 2

Variables aléatoires

2.1 Variables aléatoires et fonctions associées

Soit E l'ensemble des résultats d'une expérience aléatoire.

Définitions :

- une *variable aléatoire* V est une application de E dans $\mathbb{R}$;
- la *fonction de répartition* de V est la fonction F de $\mathbb{R}$ dans $[0,1]$: $x \to F(x) = P\{e \in E : V(e) \leq x\}$.
- une variable aléatoire est *discrète* si l'ensemble de ses valeurs possibles est fini ou dénombrable; elle est *continue* si l'ensemble de ses valeurs possibles est un intervalle de la droite réelle ou la droite réelle complète.
- la *distribution de probabilité d'une variable discrète* est un tableau contenant les valeurs possibles x_i de cette variable et les probabilités p_i d'obtenir chacune de ces valeurs $(i = 1,2,\ldots) : p_i = P\{e \in E : V(e) = x_i\}$.

Remarque : la distribution de probabilité d'une variable aléatoire discrète joue exactement le même rôle que la distribution de fréquences d'une variable observée (cf. statistique descriptive) et peut également être représentée au moyen d'un diagramme en bâtons; de même, la fonction de répartition d'une variable aléatoire discrète joue le rôle de la fonction de distribution d'une variable observée; la seule différence est que les fréquences relatives observées ont été remplacées par des probabilités théoriques.

Lorsque la variable est continue, la probabilité qu'elle prenne une valeur précise est nulle (événement possible mais stochastiquement impossible : sa fréquence relative

tend vers 0) : la notion de distribution de probabilité n'a plus de sens et est remplacée par la densité de probabilité.

Définition : la *densité de probabilité f d'une variable continue* est la dérivée (si elle existe) de sa fonction de répartition.
(Dans la suite du cours, on ne considérera que des variables continues qui ont une densité de probabilité) : $f = \dfrac{\partial F}{\partial x}$

Propriétés (à démontrer)

- F est une fonction croissante
- $\lim\limits_{x\to\infty} F(x) = 1$
- $\lim\limits_{x\to-\infty} F(x) = 0$

V **discrète**	V **continue**
$p_i \geq 0, \forall i$	$f(x) \geq 0, \forall x$
$\sum\limits_i p_i = 1$	$\int_{-\infty}^{\infty} f(\xi)\, d\xi = 1$
$F(x) = \sum\limits_{x_i \leq x} p_i$	$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi$

Notation

$$F(x) = P\{e \in E : V(e) \leq x\} = P[V \leq x]$$

Lorsqu'on considère plusieurs variables aléatoires, on placera les noms de ces variables en indices des fonctions associées : fonction de répartition de $V = F_V$.

2.2 Commentaires et exemples

Comme l'indique la définition, une variable aléatoire permet d'associer un nombre réel à chaque résultat d'une expérience. Dans beaucoup de cas d'ailleurs, le résultat de l'expérience est lui-même un nombre réel, de sorte que V sera souvent l'application identique.

Exemples :

1. Au jet d'un dé, on peut associer la variable aléatoire dont les valeurs possibles sont 1,2,3,4,5,6. Par ailleurs, si à chaque résultat du jet du dé est associé un gain, on peut considérer la variable aléatoire dont les valeurs possibles sont les différents gains que l'on peut réaliser.
2. A l'expérience qui consiste à prélever un individu dans une population, on peut associer la variable aléatoire dont les valeurs possibles sont les tailles que peuvent avoir les individus. On considère en général que ce type de variable aléatoire est continue, même si la précision limitée des instruments de mesure fait que l'ensemble des valeurs possibles est fini.

Lorsque les résultats d'une expérience ne sont pas numériques (ex. : tirage d'une boule dans une urne contenant des boules blanches et des boules noires), on peut toujours conventionnellement leur associer des nombres (1 si la boule est blanche, 0 si elle est noire).

A toute variable aléatoire V, on peut faire correspondre une fonction de répartition : cette fonction est définie sur la droite réelle et associe à chaque x réel, la probabilité de l'ensemble des résultats (c'est donc par définition un événement) qui sont appliqués par V sur des nombres $\leq x$.

Exemple

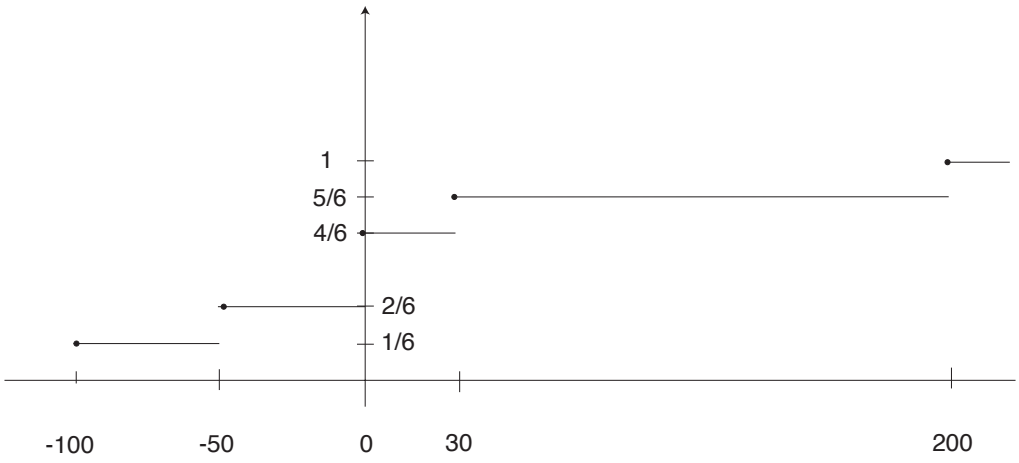
Supposons qu'aux résultats 1,2,3,4,5 et 6 du jet du dé soient associés respectivement les gains -100F, -50F, 0F, 0F, 30F et 200F. On a donc une variable aléatoire V telle que

$$V(1) = -100 \quad V(2) = -50 \quad V(3) = 0 \quad V(4) = 0 \quad V(5) = 30 \quad V(6) = 200.$$

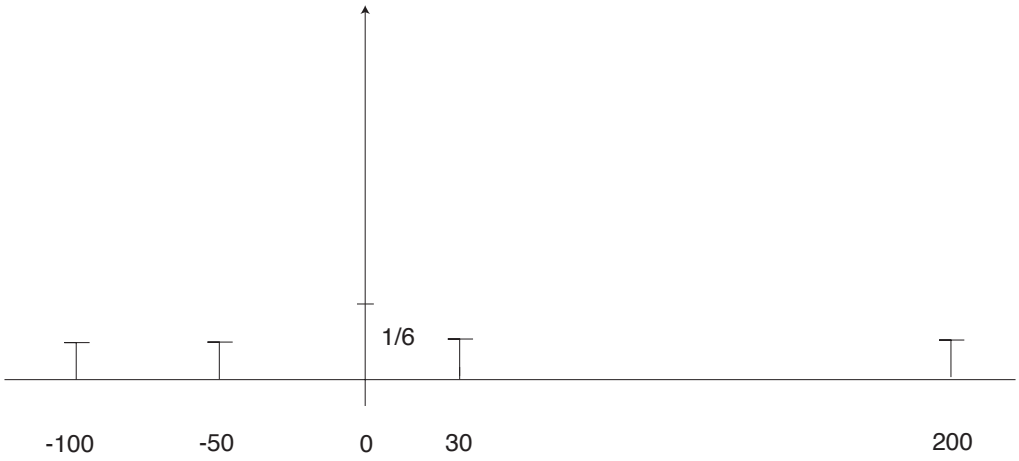
La valeur de la fonction de répartition au point $x = 5$ sera donc

$$F(5) = P\{1,2,3,4\} = 2/3$$

Dans cet exemple, le graphique de la fonction F se présente comme suit :



La variable aléatoire considérée ici est discrète : on peut donc définir sa distribution de probabilité, dont le graphique se présente comme suit :



Pour toute variable aléatoire V (discrète ou continue), on peut s'intéresser à la probabilité de l'ensemble des résultats qui sont appliqués par V dans l'intervalle $(x,x+\Delta x]$, c'est-à-dire

$$P\{e \in E : x < V(e) \leq x + \Delta x\},$$

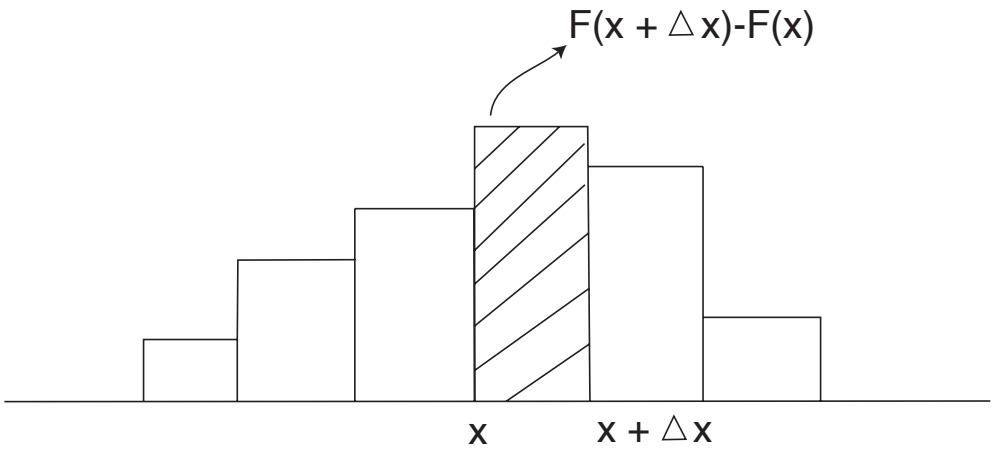
que nous noterons, en abrégé,

$$P[x < V \leq x + \Delta x].$$

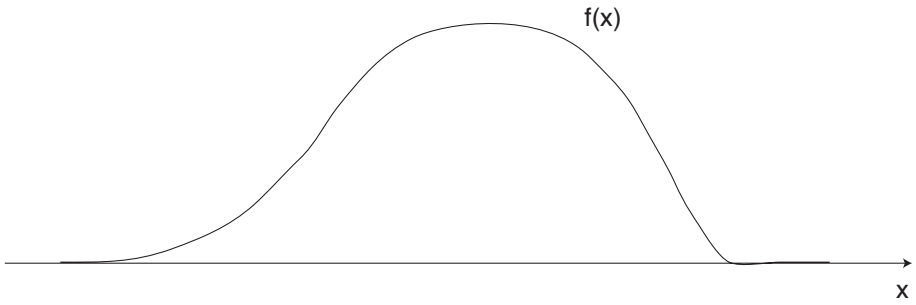
Il résulte de la définition de la fonction de répartition et des axiomes de la théorie des probabilités que cette quantité est égale à

$$F(x + \Delta x) - F(x) \qquad \text{(justifier)}.$$

Lorsque V est continue, on peut songer à lui associer un histogramme, comme on l'a fait pour les distributions de fréquences dans le cas des données groupées, constitué de rectangles de bases $= \Delta x$ et d'aires $= P[x < V \leq x + \Delta x]$, et donc de hauteurs = $\dfrac{F(x + \Delta x) - F(x)}{\Delta x}$.



Néanmoins, cela revient à supposer qu'à l'intérieur de chaque intervalle, toutes les valeurs sont également possibles. Pour se rapprocher de la réalité on a donc intérêt à considérer un Δx aussi petit que possible. A la limite, lorsque $\Delta x \to 0$, les rectangles se réduisent à des bâtons de hauteur $F'(x)$ (dérivée de $F(x)$) : la fonction ainsi obtenue s'appelle la densité de probabilité de la variable et est notée f.



C'est une fonction non-négative telle que

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \qquad\qquad (justifier).$$

Toute fonction ayant ces 2 propriétés pourra être considérée comme une densité de probabilité.

Exemple :

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \\ 0 & \text{ailleurs} \end{cases}$$

est une densité de probabilité.

Lorsque la variable aléatoire prend ses valeurs dans un intervalle $[a,b]$, on peut toujours poser $f(x) = 0$ à l'extérieur de cet intervalle.

Remarques

1. Les variables aléatoires discrètes et continues définies dans ce cours ne couvrent évidemment pas tous les types de variables aléatoires que l'on peut rencontrer, mais elles suffisent pour la plupart des applications.

2. Il résulte de la définition de la densité de probabilité que la probabilité associée à un intervalle n'est rien d'autre que l'aire située en-dessous de $f(x)$ et limitée à cet intervalle.

2.3 Valeurs typiques d'une variable aléatoire

A toute variable aléatoire, on peut donc associer une distribution de probabilité dans le cas discret, une densité de probabilité dans le cas continu et une fonction de répartition dans les 2 cas. Comme pour les distributions de fréquences en statistique descriptive, ces fonctions peuvent être représentées graphiquement. Elles peuvent aussi être caractérisées au moyen de quelques paramètres bien choisis. Ces paramètres sont les mêmes que ceux vus en statistique descriptive, les fréquences relatives étant remplacées par des probabilités.

		V **discrète**	V **continue**
Moyenne	$\mu =$	$\sum_i p_i x_i$	$\int_{-\infty}^{\infty} x f(x)\, dx$
Variance	$\sigma^2 =$	$\sum_i p_i (x_i - \mu)^2$	$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$
Moments d'ordre k **par rapport à** μ	$\mu_k =$	$\sum_i p_i (x_i - \mu)^k$	$\int_{-\infty}^{\infty} (x-\mu)^k f(x)\, dx$
Moments d'ordre k **par rapport à** 0	$\alpha_k =$	$\sum_i p_i x_i^k$	$\int_{-\infty}^{\infty} x^k f(x)\, dx$

Remarques :

1. La moyenne est aussi appelée *espérance mathématique*.

2. Dans le cas continu, des valeurs typiques peuvent ne pas exister puisqu'elles sont définies à partir d'intégrales généralisées; il en est de même dans le cas discret lorsque la variable peut prendre une infinité dénombrable de valeurs (on a alors des séries).

L'interprétation des valeurs typiques est analogue à celle vue pour les valeurs typiques des distributions observées (cf. statistique descriptive) : μ est un paramètre de position, σ^2 est un paramètre de dispersion, μ_3 est un paramètre de dissymétrie et μ_4 un paramètre d'aplatissement (on aura l'occasion d'y revenir lors de l'étude des variables aléatoires particulières).

La *médiane* est l'abscisse $\widetilde{\mu}$ telle que $F(\widetilde{\mu}) = 1/2$; dans le cas d'un saut de discontinuité ou d'un palier, on procède comme pour les distributions de fréquences.

Un *mode* est une abscisse où la distribution de probabilité ou la densité de probabilité présente un maximum local.

Paramètre de dissymétrie de Fisher : $\gamma_1 = \dfrac{\mu_3}{\sigma^3}$

Paramètre d'aplatissement de Fisher : $\gamma_2 = \dfrac{\mu_4}{\sigma^4} - 3$

Variable réduite.
Si V est une variable aléatoire de moyenne μ et d'écart-type σ, la *variable réduite* correspondante est $\dfrac{V-\mu}{\sigma}$: il s'agit d'une variable de moyenne 0 et d'écart-type 1 (conséquences des propriétés des valeurs typiques : voir plus loin).

Exemples numériques

1ier exemple (cf. exemple du dé du §précédent)

$$\begin{aligned}
\text{Moyenne des gains} &= \frac{1}{6}(-100) + \frac{1}{6}(-50) + \frac{2}{6}(0) + \frac{1}{6}(30) + \frac{1}{6}(200) = 13{,}33 \\
\text{Variance des gains} &= \frac{1}{6}(-100-13{,}33)^2 + \frac{1}{6}(-50-13{,}33)^2 \\
&\quad + \frac{2}{6}(0-13{,}33)^2 + \frac{1}{6}(30-13{,}33)^2 + \frac{1}{6}(200-13{,}33)^2
\end{aligned}$$

2ème exemple (cf. exemple de densité de probabilité du §précédent).

$$\begin{aligned}
\text{Moyenne} &= \int_0^1 x^2\,dx + \int_1^2 x(2-x)\,dx = 1 \\
\text{Variance} &= \int_0^1 (x-1)^2 x\,dx + \int_1^2 (x-1)^2(2-x)\,dx
\end{aligned}$$

Médiane : 1ier exemple : 0, 2ème exemple : 1

Mode : 1ier exemple : 0, 2 ème exemple : 1

2.4 Vecteurs aléatoires à deux dimensions

2.4.1 Définitions

– Un *vecteur aléatoire* Z est un couple (V,W) de variables aléatoires : il définit donc une application de E (ensemble des résultats de l'expérience) dans $\mathbb{R}^2$.

– La *fonction de répartition* de Z est la fonction F de $\mathbb{R}^2$ dans $[0,1]$:

$$(x,y) \to F(x,y) = P\{e \in E : V(e) \leq x \text{ et } W(e) \leq y\}$$

– Un vecteur aléatoire Z est *discret* si l'image de E par Z est un ensemble fini ou dénombrable de points du plan; il est *continu* si cette image est un domaine connexe du plan (on ne considérera pas ici les autres types de vecteurs aléatoires).

– La *distribution de probabilité* d'un vecteur aléatoire *discret* $Z = (V,W)$ est l'ensemble des couples (x_i,y_j) de valeurs possibles de (V,W) et l'ensemble p_{ij} des probabilités correspondantes :

$$p_{ij} = P\{e \in E : V(e) = x_i \text{ et } W(e) = y_j\}.$$

– La *densité de probabilité* d'un vecteur aléatoire *continu* est la dérivée seconde mixte de sa fonction de répartition, à condition qu'elle existe : $f = \dfrac{\partial F}{\partial y\, \partial x}$. On suppose dans la suite du cours que les dérivées secondes mixtes existent et sont égales entre elles.

– La *fonction de répartition marginale* de V est la fonction F_1 de $\mathbb{R}$ dans $[0,1]$:

$$x \to F_1(x) = P\{e \in E : V(e) \leq x\}.$$

– La *distribution de probabilité marginale* de V est l'ensemble des valeurs possibles x_i de la variable V auxquelles sont associées les probabilités :

$$p_{i.} = \sum_j p_{ij}. \qquad \text{(cas discret)}$$

– La *densité de probabilité marginale* de V est la fonction

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,n)\, dn \qquad \text{(cas continu)}$$

On définit de façon analogue la fonction de répartition marginale de W (notée $F_2(y)$), la distribution de probabilité marginale de $W(p_{.j})$ et la densité de probabilité marginale de $W(f_2(y))$.

– La *distribution de probabilité conditionnelle* de V (discrète) sous la condition $W = y_j$ est l'ensemble des valeurs possibles x_i de la variable V auxquelles sont associées les probabilités

$$p_{i/j} = \frac{p_{ij}}{p_{.j}}$$

– La *densité de probabilité conditionnelle* de V (continue) sous la condition $W = y_0$ est la fonction

$$f(x/y_0) = \frac{f(x,y_0)}{f_2(y_0)}$$

2.4.2 Propriétés

Elles résultent des définitions précédentes et des propriétés des probabilités.

- F est une fonction croissante en x et en y;
- $\lim\limits_{\substack{x\to+\infty \\ y\to+\infty}} F(x,y) = 1 \quad ; \quad \lim\limits_{\substack{x\to-\infty \\ y\to-\infty}} F(x,y) = 0$;
- $F_1(x) = \lim\limits_{y\to+\infty} F(x,y)$; $F_2(y) = \lim\limits_{x\to+\infty} F(x,y)$;

Z discret	**Z continu**
$p_{ij} \geq 0, \forall i,j$	$f(x,y) \geq 0, \forall x,y$
$\sum_i \sum_j p_{ij} = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi,\eta)\, d\xi\, d\eta = 1$
$F(x,y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij}$	$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi,\eta) d\xi\, d\eta$
$F_1(x) = \sum_{x_i \leq x} p_{i.}$	$F_1(x) = \int_{-\infty}^{x} f_1(\xi) d\xi$ $f_1(x) = F_1'(x)$
$F_2(y) = \sum_{y_j \leq y} p_{.j}$	$F_2(y) = \int_{-\infty}^{y} f_2(\eta) d\eta$ $f_2(y) = F_2'(x)$
$\sum_i p_{i.} = 1$	$\int_{-\infty}^{\infty} f_1(\xi)\, d\xi = 1$
$\sum_j p_{.j} = 1$	$\int_{-\infty}^{\infty} f_2(\eta) d\,\eta = 1$

2.4.3 Valeurs typiques

Aux distributions à 1 dimension (marginales et conditionnelles), on associe les valeurs typiques habituelles.

	Discret	**Continu**
Moyennes	$\mu_1 = \sum_i p_{i.} x_i$	$\mu_1 = \int_{-\infty}^{\infty} x f_1(x)\, dx$
marginales	$\mu_2 = \sum_j p_{.j} y_j$	$\mu_2 = \int_{-\infty}^{\infty} y f_2(y)\, dy$
Variances	$\sigma_1^2 = \sum_i p_{i.}(x_i - \mu_1)^2$	$\sigma_1^2 = \int_{-\infty}^{\infty} (x-\mu_1)^2 f_1(x)\, dx$
marginales	$\sigma_2^2 = \sum_{.j} p_{.j}(y_j - \mu_2)^2$	$\sigma_2^2 = \int_{-\infty}^{\infty} (y-\mu_2)^2 f_2(y)\, dy$
Moyennes	$\mu_{1/j} = \sum_i p_{i/j} x_i$	$\mu_{1/y_0} = \int_{-\infty}^{\infty} x f(x/y_0)\, dx$
conditionnelles	$\mu_{2/j} = \sum_j p_{j/i} y_j$	$\mu_{2/x_0} = \int_{-\infty}^{\infty} y f(y/x_0)\, dy$
Variances	$\sigma_{1/j}^2 = \sum_i p_{i/j}(x_i - \mu_{1/j})^2$	$\sigma_{1/y_0}^2 = \int_{-\infty}^{\infty} (x-\mu_{1/y_0})^2 f(x/y_0)\, dx$
conditionnelles	$\sigma_{2/i}^2 = \sum_j p_{j/i}(y_j - \mu_{2/i})^2$	$\sigma_{2/x_0}^2 = \int_{-\infty}^{\infty} (x-\mu_{2/x_0})^2 f(y/x_0)\, dy$

2.4.4 Etude de la dépendance linéaire entre les 2 variables

a) **La covariance**

Définition :

$$\mu_{11} = \sum_i \sum_j p_{ij}(x_i - \mu_1)(y_j - \mu_2) \qquad \text{(discret)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_1)(y-\mu_2) f(x,y) dx\, dy \qquad \text{(continu)}$$

Interprétation : identique à celle vue pour la covariance de 2 variables observées.

Propriétés :

1) $|\mu_{11}| \leq \sigma_1 \sigma_2$

2) $|\mu_{11}| = \sigma_1\sigma_2$ *si et seulement si toute la masse de probabilité est concentrée sur une droite non parallèle aux axes*.

Les démonstrations sont identiques à celles vues pour les distributions observées : il suffit de remplacer les fréquences relatives par des probabilités. Dans le cas discret, on se base sur la relation :

$$\sum_i \sum_j p_{ij}[u(x_i - \mu_1) + (y_j - \mu_2)]^2 \geq 0, \quad \forall u;$$

dans le cas continu, on se base sur la relation

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [u(x-\mu_1) + (y-\mu_2)]^2 \, f(x,y)\, dx\, dy \geq 0, \forall u.$$

b) **Le coefficient de corrélation**

Définition :

$$\rho = \frac{\mu_{11}}{\sigma_1\sigma_2} \qquad (\sigma_1 \text{ et } \sigma_2 \text{ supposés } \neq 0)$$

Interprétation : identique à celle vue pour le coefficient de corrélation de 2 variables observées.

Propriétés :

1) $-1 \leq \rho \leq 1$

2) $|\rho| = 1$ *si et seulement si toute la masse de probabilité est concentrée sur une droite non parallèle aux axes*.
(ces propriétés résultent immédiatement de celles de la covariance).

c) **Les droites de régression**

– La droite de régression de W en V a pour équation

$$y = \frac{\mu_{11}}{\sigma_1^2}(x - \mu_1) + \mu_2;$$

– La droite de régression de V en W a pour équation

$$x = \frac{\mu_{11}}{\sigma_2^2}(y - \mu_2) + \mu_1;$$

– $\dfrac{\mu_{11}}{\sigma_1^2} = \rho\dfrac{\sigma_2}{\sigma_1}$ est appelé coefficient de régression de W en V et est noté β_{21};

– $\dfrac{\mu_{11}}{\sigma_2^2} = \rho\dfrac{\sigma_1}{\sigma_2}$ est appelé coefficient de régression de V en W et est noté β_{12};

– la variance résiduelle de W en V est la quantité

$$\sigma_{21}^2 = \sigma_2^2(1 - \rho^2);$$

– la variance résiduelle de V en W est la quantité

$$\sigma_{12}^2 = \sigma_1^2(1 - \rho^2).$$

Tout ce qui précède s'obtient par des raisonnements analogues à ceux vus en statistique descriptive, les fréquences relatives étant simplement remplacées par des probabilités.

2.4.5 Les moments

Le moment d'ordre (k,ℓ) par rapport au point (c,d) est la quantité

$$\sum_i \sum_j p_{ij}(x_i - c)^k (y_j - d)^\ell \qquad \text{(cas discret)}$$

ou

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - c)^k (y - d)^\ell f(x,y)dx\,dy \qquad \text{(cas continu)}$$

On note $\mu_{k\ell}$ le moment d'ordre (k,ℓ) par rapport au point (μ_1,μ_2) et $\alpha_{k\ell}$ le moment d'ordre (k,ℓ) par rapport au point $(0,0)$.

On a

$$\alpha_{00} = 1$$

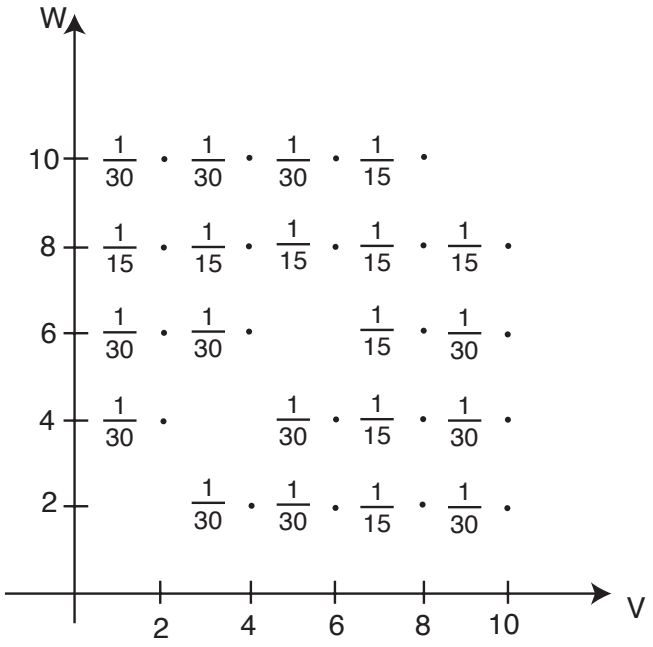
$$\begin{array}{ll} \alpha_{10} = \mu_1 & \alpha_{01} = \mu_2 \\ \mu_{10} = 0 & \mu_{01} = 0 \\ \mu_{20} = \sigma_1^2 & \mu_{02} = \sigma_2^2 \end{array}$$

$$\mu_{11} = \text{ covariance.}$$

2.4.6 Commentaires et exemples

Un vecteur aléatoire à 2 dimensions permet d'associer un point du plan à chaque résultat d'une expérience. Dans certains cas d'ailleurs, le résultat de l'expérience est lui-même un couple de nombres réels (ex. : à l'expérience qui consiste à prélever un individu dans une population, on peut associer le vecteur aléatoire dont les valeurs possibles sont les couples (taille, poids) des individus).

Exemple 1 : on prélève, sans remplacement, 2 cartes d'un paquet de 6 cartes contenant un 2, un 4, un 6, deux 8 et un 10. Le résultat de l'expérience est un des couples $(2,4),(2,6),(2,8),(2,10),(4,2),(4,6),\ldots,(8,10)$. La distribution de probabilité se présente comme suit (on a indiqué à côté de chaque point la probabilité correspondante).



V = valeur de la 1^{ere} carte

W = valeur de la 2^{eme} carte

La fonction de répartition F associe à chaque point du plan la probabilité de se trouver "en-dessous et à gauche" du point; la fonction de répartition marginale F_1 associe à chaque abscisse la probabilité de se trouver à gauche de cette abscisse.
On a donc par exemple $F(x,y) = 0, \forall x < 2$ et $\forall y < 2, F(2,2) = 0, F(3,3) = 0, F(4,3) = \frac{1}{30}, F(4,4) = \frac{1}{15}, F(6,5) = \frac{2}{15}, F(20,3) = \frac{1}{6}, F(4,8) = \frac{4}{15}, F(10,10) = 1, F(15,15) = 1, F_1(3) = \frac{1}{6}, F_1(8) = \frac{5}{6}, F_1(1) = 0, F_1(10) = 1.$

Les distributions de probabilité marginales s'obtiennent "par projection" sur les axes; dans cet exemple, elles sont identiques pour les 2 variables :

Valeur	2	4	6	8	10
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$

La distribution de probabilité conditionnelle de V sous la condition $W = 6$ se présente comme suit :

Valeur	2	4	6	8	10
Probabilité	$\frac{1}{5}$	$\frac{1}{5}$	0	$\frac{2}{5}$	$\frac{1}{5}$

Moyenne marginale de $V = \frac{1}{6} \times 2 + \frac{1}{6} \times 4 + \frac{1}{6} \times 6 + \frac{1}{3} \times 8 + \frac{1}{6} \times 10 = \frac{19}{3} =$ moyenne marginale de W.

Variance marginale de $V = \frac{1}{6}(2 - \frac{19}{3})^2 + \frac{1}{6}(4 - \frac{19}{3})^2 + \frac{1}{6}(6 - \frac{19}{3})^2 + \frac{1}{3}(8 - \frac{19}{3})^2 + \frac{1}{6}(10 - \frac{19}{3})^2$

N.B. : pour le calcul des valeurs typiques (notamment la variance), on dispose des mêmes formules simplifiées qu'en statistique descriptive.

Moyenne conditionnelle de V sous la condition $W = 6$

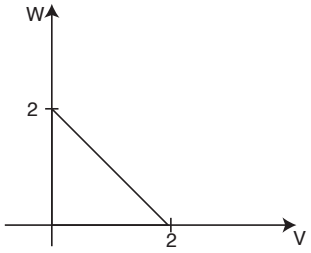
$$= \frac{1}{5} \times 2 + \frac{1}{5} \times 4 + \frac{2}{5} \times 8 + \frac{1}{5} \times 10$$

$$\text{Covariance} = \frac{1}{30}(4-\frac{19}{3})(2-\frac{19}{3}) + \frac{1}{30}(6-\frac{19}{3})(2-\frac{19}{3}) + \ldots + \frac{1}{15}(8-\frac{19}{3})(10-\frac{19}{3}).$$

Exercice : achever les calculs, déterminer le coefficient de corrélation et les droites de régression.

Exemple 2 : la fonction $f(x,y) = \begin{cases} 1/2 & 0 \leq x+y \leq 2,\ x \geq 0, y \geq 0, \\ 0 & \text{ailleurs} \end{cases}$ est une densité de probabilité dans le plan (fonction non négative dont l'intégrale double, étendue au plan, vaut 1).

Fonction de répartition : $F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi,\eta)\, d\xi\, d\eta.$



Ainsi :

$$\begin{aligned}
F(x,y) &= 0 \qquad \forall x \leq 0, \forall y \leq 0, \\
F(1,1) &= \int_0^1 \int_0^1 \frac{1}{2} d\xi\, d\eta = 1/2 \\
F(2,1) &= \int_0^1 \int_0^1 \frac{1}{2} d\xi\, d\eta + \int_1^2 \left[\int_0^{2-\xi} \frac{1}{2} d\eta \right] d\xi = 3/4 \\
F(2,2) &= 1 \\
F(1,\frac{1}{2}) &= \int_0^1 \int_0^{1/2} \frac{1}{2}\, d\xi\, d\eta = \frac{1}{4}
\end{aligned}$$

Densité de probabilité marginale de V :

$$f_1(x) = \begin{cases} 0 & x \leq 0 \text{ ou } x \geq 2 \\ \displaystyle\int_0^{2-x} \frac{1}{2}\, d\eta = \frac{1}{2}(2-x) & 0 < x < 2 \end{cases}$$

$$F_1(1) = \int_0^1 \frac{1}{2}(2-x)\, dx$$

Densité de probabilité conditionnelle de W pour $V = 1$

$$\begin{aligned} f(y/1) &= \frac{f(1,y)}{f_1(1)} \\ &= \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & \text{ailleurs} \end{cases} \end{aligned}$$

Moyenne marginale de $V = \int_0^2 x\frac{1}{2}(2-x)\,dx = \frac{2}{3}$

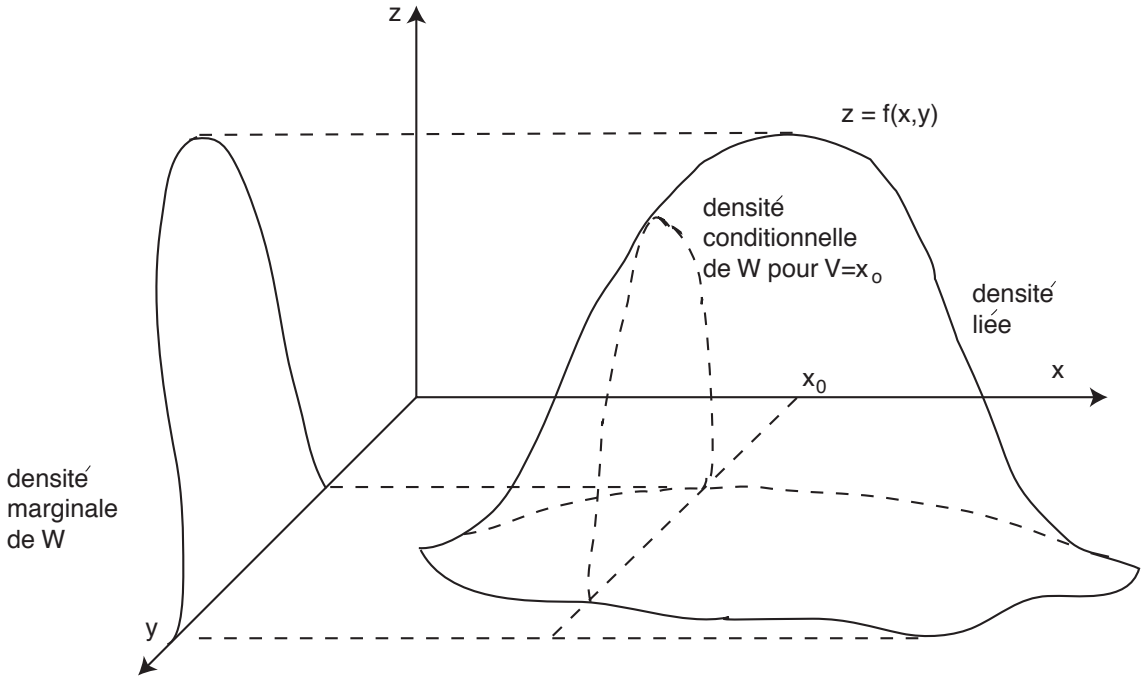
Variance marginale de $V = \int_0^2 (x-\frac{2}{3})^2\frac{1}{2}(2-x)\,dx$

Moyenne conditionnelle de W pour $V = 1$

$$= \int_0^1 y\,dy = \frac{1}{2}$$

Exercice : calculer densité de probabilité et moyenne marginales de W, covariance, coefficient de corrélation et droites de régression.

Interprétation géométrique des densités marginales et conditionnelles.



2.4.7 Indépendance des 2 variables

Définition : V et W sont indépendantes ssi, $\forall x_1,x_2,y_1,y_2$:

$$P\left[x_1 \leq V \leq x_2, y_1 \leq W \leq y_2\right] = P\left[x_1 \leq V \leq x_2\right] \cdot P\left[y_1 \leq W \leq y_2\right],$$

ssi $\forall x,y$:

$$F(x,y) = F_1(x) \cdot F_2(y) \qquad \text{(à démontrer)}$$

Propriétés : si V et W sont indépendantes, alors (à démontrer)

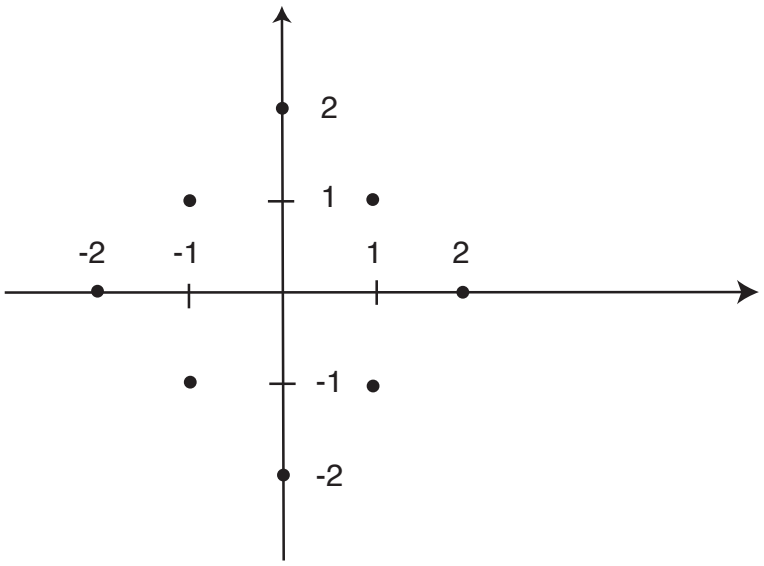
- $p_{ij} = p_{i.} \cdot p_{.j}$ (cas discret)
- $f(x,y) = f_1(x) \cdot f_2(y)$ (cas continu)
- $p_{i/j} = p_{i.}$ et $p_{j/i} = p_{.j}, \quad \forall i,j$ (cas discret)
- $f(x/y_0) = f_1(x)$ et $f(y/x_0) = f_2(y), \forall x,y,x_0,y_0$ (cas continu)

- $$\begin{aligned} u_{11} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)f_1(x)\cdot f_2(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}(x-\mu_1)f_1(x)\,dx \cdot \int_{-\infty}^{\infty}(y-\mu_2)f_2(y)\,dy \\ &= (\mu_1-\mu_1)\cdot(\mu_2-\mu_2) \\ &= 0 \end{aligned}$$
- les droites de régression sont parallèles aux axes (puisque $\mu_{11} = 0$)
- la connaissance des distributions marginales suffit pour connaître la distribution liée.

Lorsque $\mu_{11} = 0$, les variables sont dites *non-corrélées*. Le raisonnement ci-dessus montre que *des variables indépendantes sont non-corrélées. La réciproque n'est pas vraie*, comme le montre l'exemple ci-dessous.

Exemple de variables non corrélées mais dépendantes.

Soit la distribution de probabilité représentée ci-dessous, où à chaque point est associée une probabilité $\frac{1}{8}$.



$$\left.\begin{array}{l} F(-1,-1) = 1/8 \\ F_1(-1) = 3/8 \text{ et } F_2(-1) = 3/8 \end{array}\right\} \Rightarrow \text{ pas indépendance}$$

Un rapide calcul montre que $\mu_{11} = 0$, et donc $\rho = 0$.

2.5 Opérations sur les variables aléatoires

Toute fonction d'une ou de plusieurs variables aléatoires est elle-même une variable aléatoire. Il est important de pouvoir en déterminer la distribution, à partir de celles des variables initiales.

2.5.1 Distribution d'une fonction monotone d'une variable aléatoire

Soit V une variable aléatoire et $W = G(V)$ où G est dérivable et croissante. La fonction de répartition de W est, par définition,

$$\begin{aligned} F_W(x) &= P[W \leq x] \\ &= \mathbb{P}[G(V) \leq x] \\ &= \mathbb{P}[V \leq G^{-1}(x)] \\ &= F_V(G^{-1}(x)); \end{aligned}$$

on peut donc la calculer à partir de la fonction de répartition de V. Dans le cas continu, on obtient

$$\begin{aligned} f_W(x) &= F'_W(x) \\ &= F'_V(G^{-1}(x)) \cdot \frac{1}{G'(G^{-1}(x))} \end{aligned}$$

$$\Rightarrow \boxed{f_W(x) = f_V(G^{-1}(x)) \cdot \frac{1}{G'(G^{-1}(x))}}$$

Si G est décroissante, on obtient (à démontrer)

$$\boxed{f_W(x) = -f_V(G^{-1}(x)) \cdot \frac{1}{G'(G^{-1}(x))}}$$

2.5.2 Distribution de la somme de 2 variables aléatoires

Soit V et W deux variables aléatoires; la fonction de répartition de $Z = V + W$ est donnée par

$$\begin{aligned} F_Z(x) &= P[Z \leq x] \\ &= P[V + W \leq x]. \end{aligned}$$

Dans le cas discret, on obtient

$$F_Z(x) = \sum_i \sum_j p_{ij} \text{ pour les couples } (i,j) \text{ tels que } v_i + w_j \leq x,$$

et où

$$p_{ij} = P[V = v_i, W = w_j].$$

Dans le cas continu,

$$\begin{aligned} F_Z(x) &= \iint_{\xi+\eta \le x} f_{(V,W)}(\xi,\eta)\, d\xi\, d\eta \\ &= \iint_{v \le x} f_{(V,W)}(u, v-u) du\, dv \qquad \begin{cases} \xi = u \\ \eta = v - u \\ J = 1 \end{cases} \\ &= \int_{-\infty}^{x} dv \int_{-\infty}^{\infty} f_{(V,W)}(u, v-u)\, du \end{aligned}$$

d'où

$$\boxed{f_{V+W}(x) = \int_{-\infty}^{\infty} f_{(V,W)}(u, x-u)\, du}$$

Si V et W sont indépendantes, on obtient

$$f_{V+W}(x) = \int_{-\infty}^{\infty} f_V(u) \cdot f_W(x-u)\, du.$$

2.5.3 Distribution du produit de 2 variables aléatoires

Soit V et W deux variables aléatoires; la fonction de répartition de $Z = V \cdot W$ est donnée par

$$F_Z(x) = \mathbb{P}[V \cdot W \le x].$$

Dans le cas continu, on obtient

$$\begin{aligned} F_Z(x) &= \iint_{\xi \cdot \eta \le x} f_{(V,W)}(\xi,\eta)\, d\xi\, d\eta \\ &= \iint_{v \le x} f_{(V,W)}(u, \frac{v}{u}) |\frac{1}{u}| du\, dv \qquad \begin{cases} \xi = u \\ \eta = \dfrac{v}{u} \\ J = \dfrac{1}{u} \end{cases} \\ &= \int_{-\infty}^{x} dv \int_{-\infty}^{\infty} f_{(V,W)}(u, \frac{v}{u}) |\frac{1}{u}|\, du \end{aligned}$$

d'où

$$\boxed{f_{V \cdot W}(x) = \int_{-\infty}^{\infty} f_{(V,W)}(u, \frac{x}{u}) \cdot |\frac{1}{u}|\, du}$$

Si V et W sont indépendantes, on obtient

$$f_{V\cdot W}(x) = \int_{-\infty}^{\infty} f_V(u) \cdot f_W(\frac{x}{u}) \cdot |\frac{1}{u}|\, du.$$

2.5.4 Exemples

a) Soit V une variable aléatoire et $W = 3V + 5$. On a

$$\begin{aligned} F_W(x) &= P[W \leq x] \\ &= P[3V + 5 \leq x] \\ &= P[V \leq \frac{x-5}{3}] = F_V\left(\frac{x-5}{3}\right) \end{aligned}$$

Dans le cas discret, on obtient

$$F_W(x) = \sum_{v_i \leq \frac{x-5}{3}} P[V = v_i];$$

dans le cas continu,

$$f_W(x) = \frac{1}{3} f_V\left(\frac{x-5}{3}\right).$$

b) Soit V une variable aléatoire et $W = V^2$.

$$\begin{aligned} F_W(x) &= P\left[W \leq x\right] \\ &= \begin{cases} 0 & \text{si } x < 0 \\ P[-\sqrt{x} \leq V \leq \sqrt{x}] & \text{si } x \geq 0 \end{cases} \\ &= \begin{cases} 0 & \text{si } x < 0 \\ F_V(\sqrt{x}) - F_V(-\sqrt{x} - 0) & \text{si } x \geq 0 \end{cases} \end{aligned}$$

Dans le cas continu, on obtient

$$f_W(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{1}{2\sqrt{x}} f_V(\sqrt{x}) + \frac{1}{2\sqrt{x}} f_V(-\sqrt{x}) & \text{si } x > 0 \end{cases}$$

2.6 Propriétés de la moyenne et de la variance

2.6.1 La moyenne ou espérance mathématique

La moyenne μ d'une variable aléatoire V,

$$\mu = \sum_i p_i x_i \text{ ou } \int_{-\infty}^{\infty} x f_V(x)\, dx$$

est encore appelée espérance mathématique de V et notée $E(V)$.

1ère propriété : si $W = G(V)$ (où G est dérivable), on a

$$\begin{aligned} E(W) &= \int_{-\infty}^{\infty} y f_W(y)\, dy \\ &= \int_{G\nearrow} y f_V(G^{-1}(y)) \frac{dy}{G'(G^{-1}(y))} - \int_{G\searrow} y f_V(G^{-1}(y)) \frac{dy}{G'(G^{-1}(y))} \end{aligned}$$

d'où, en posant $x = G^{-1}(y)$, (dans la 2ème intégrale, le signe - compense la permutation des bornes)

$$\boxed{E(G(V)) = \int_{-\infty}^{\infty} G(x) f_V(x)\, dx = \int_{-\infty}^{\infty} y f_{G(V)}(y)\, dy}$$

En particulier, l'espérance mathématique d'une constante est cette constante, d'autre part,

$$\begin{aligned} E(aV+b) &= \int_{-\infty}^{\infty} (ax+b) f_V(x)\, dx \\ &= a \int_{-\infty}^{\infty} x f_V(x)\, dx + b \int_{-\infty}^{\infty} f_V(x)\, dx \\ &= aE(V) + b \end{aligned}$$

Si on effectue une transformation linéaire d'une variable aléatoire, sa moyenne subit la même transformation linéaire.

N.B. : les démonstrations présentées ici concernent le cas continu; elles se transposent sans difficulté au cas discret.

La propriété précédente se généralise au cas d'une fonction de plusieurs variables :

$$\begin{aligned} E(G(V,W)) &= \int_{-\infty}^{\infty} z f_{G(V,W)}(z)\, dz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y) f_{(V,W)}(x,y)\, dx\, dy. \end{aligned}$$

2ème propriété : si $Z = V + W$, on a

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} x f_{V+W}(x)\, dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{(V,W)}(u, x-u)\, du\, dx \qquad \left\{ \begin{array}{l} u = \xi \\ x = \xi + \eta \\ J = 1 \end{array} \right. \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\xi+\eta) \cdot f_{(V,W)}(\xi,\eta) d\xi\, d\eta \\ &= \int_{-\infty}^{\infty} \xi f_V(\xi)\, d\xi + \int_{-\infty}^{\infty} \eta f_W(\eta)\, d\eta, \end{aligned}$$

f_V et f_W étant les densités de probabilité marginales de V et de W,

$$= E(V) + E(W).$$

La moyenne d'une somme de variables aléatoires est toujours égale à la somme des moyennes.

3ème propriété : si $Z = V \cdot W$ et si V et W sont indépendantes, on a

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} x \cdot f_{V\cdot W}(x)\, dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_V(u) \cdot f_W\left(\frac{x}{u}\right) \cdot \left|\frac{1}{u}\right| du\, dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta \cdot f_V(\xi) \cdot f_W(\eta) d\xi\, d\eta \qquad \left\{ \begin{array}{l} u = \xi \\ x = \xi\eta \\ J = \xi \end{array} \right. \\ &= \int_{-\infty}^{\infty} \xi \cdot f_V(\xi)\, d\xi \cdot \int_{-\infty}^{\infty} \eta \cdot f_W(\eta)\, d\eta, \\ &= E(V) \cdot E(W). \end{aligned}$$

La moyenne du produit de 2 variables aléatoires **indépendantes** *est égale au produit des moyennes.*

2.6.2 La variance

En vertu de ce qui précède, la variance de V

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_V(x)\, dx$$

n'est rien d'autre que $E[(V-\mu)^2]$; la variance de V est parfois notée $D^2(V)$.

1ère propriété : Si $W = aV + b$, on a

$$\begin{aligned} E(W) &= aE(V) + b \\ &= a\mu + b \end{aligned}$$

et donc

$$\begin{aligned} D^2(W) &= E\left[(W - a\mu - b)^2\right] \\ &= E\left[(aV + b - a\mu - b)^2\right] \\ &= E\left[a^2(V-\mu)^2\right] \\ &= a^2 \cdot D^2(V). \qquad \text{(justifier)} \end{aligned}$$

La variance de $aV + b$ est égale à la variance de V multipliée par a^2.

2ème propriété : Si $Z = V + W$, on a (μ,μ_1,μ_2 désignant les moyennes de Z,V et W):

$$\begin{aligned} D^2(Z) &= E\left[(Z-\mu)^2\right] \\ &= E\left[(V + W - \mu_1 - \mu_2)^2\right] && \text{(justifier)} \\ &= E\left[(V-\mu_1)^2\right] + E\left[(W-\mu_2)^2\right] + 2E\left[(V-\mu_1)(W-\mu_2)\right] && \text{(justifier)} \\ &= D^2(V) + D^2(W) + 2\mu_{11} && \text{(justifier)} \end{aligned}$$

Il en résulte que la variance d'une somme de variables aléatoires **indépendantes** *est égale à la somme des variances.*

3ème propriété : la variance est le plus petit moment d'ordre 2.

En effet, $\forall c$:

$$\begin{aligned} E\left[(V-c)^2\right] &= E\left[(V-\mu+\mu-c)^2\right] \\ &= E\left[(V-\mu)^2\right]+(\mu-c)^2, \qquad \text{(justifier)} \end{aligned}$$

qui est minimum pour $c=\mu$.

2.6.3 Moyenne et variance d'une variable réduite

Il résulte des propriétés précédentes que si μ et σ sont la moyenne et l'écart-type de V, alors

$$E\left(\frac{V-\mu}{\sigma}\right)=0 \qquad \text{(justifier)}$$

et

$$D^2\left(\frac{V-\mu}{\sigma}\right)=1 \qquad \text{(justifier)}$$

2.7 Fonctions génératrices et caractéristiques

Définition : la fonction génératrice des moments d'une variable V est la fonction $\psi(t)$ définie par

$$\psi(t)=E(e^{tV}).$$

En vertu des propriétés de l'espérance mathématique, elle est donc égale à

$$\begin{aligned} \psi(t) &= \int_{-\infty}^{\infty} e^{tx} f_V(x)\,dx \qquad \text{(cas continu)}, \\ &= \sum_i e^{tx_i} p_i \qquad \text{(cas discret)}. \end{aligned}$$

En remplaçant l'exponentielle par son développement en série

$$e^{tx}=1+tx+\frac{t^2x^2}{2!}+\ldots+\frac{t^nx^n}{n!}+\ldots \quad ,$$

on obtient (moyennant des conditions assez fortes de convergence pour que les opérations soient permises) :

$$\psi(t) = \sum_{k=0}^{\infty} \alpha_k \frac{t^k}{k!},$$

où α_k sont les moments d'ordre k par rapport à l'origine de la variable V; on en déduit

$$\alpha_k = \left(\frac{\partial^k \psi}{\partial t^k} \right)_{t=0};$$

c'est la raison pour laquelle $\psi(t)$ est appelée la fonction génératrice des moments.
Les difficultés éventuelles concernant la convergence peuvent être évitées en introduisant la fonctions complexe

$$\varphi(t) = \psi(it) \qquad (i = \sqrt{-1}),$$

appelée *fonction caractéristique* de V, qui existe toujours.

On peut démontrer qu'il existe une correspondance biunivoque entre l'ensemble des fonctions de répartition et l'ensemble des fonctions caractéristiques; la distribution de probabilité ou la densité de probabilité d'une variable aléatoire est donc entièrement déterminée par la fonction $\varphi(t)$ (d'où son nom). On a aussi

$$i^k \alpha_k = \left(\frac{\partial^k \varphi}{\partial t^k} \right)_{t=0}.$$

Propriétés

1. Si $W = aV + b$ et si $\psi(t) = E(e^{tV})$, alors

$$\begin{aligned} E(e^{tW}) &= E(e^{atV+tb}) \\ &= E(e^{atV} \cdot e^{tb}) \\ &= e^{tb} \cdot \psi(at). \end{aligned}$$

2. *La fonction génératrice d'une somme de variables aléatoires indépendantes est égale au produit des fonctions génératrices de ces variables*.
   En effet,

$$\begin{aligned} E(e^{t(V+W)}) &= E(e^{tV} \cdot e^{tW}) \\ &= E(e^{tV}) \cdot E(e^{tW}) \qquad \text{(justifier)} \end{aligned}$$

Chapitre 3

Variables aléatoires particulières

3.1 Variable indicatrice

voir exercices

3.2 Variable uniforme

voir exercices

3.3 Variable triangulaire

voir exercices

3.4 Variable binomiale

On considère n répétitions indépendantes d'une expérience aléatoire; à chaque répétition, le phénomène A peut se produire avec une probabilité p. On s'intéresse à la variable aléatoire "nombre de réalisations de A au cours des n expériences". C'est une variable discrète dont les valeurs possibles sont $0,1,2,\ldots,n$ et dont la distribution de probabilité est caractérisée par

$$P[\text{variable } = k] = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0,\ldots,n.$$

En effet, il existe plusieurs manières de réaliser k fois A au cours de n répétitions de l'expérience : le nombre de manières est donné par le nombre de combinaisons de k objets parmi n :

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

La probabilité de chaque manière étant égale à $p^k \cdot (1-p)^{n-k}$ (justifier) et toutes ces manières étant incompatibles (justifier), on en déduit que la probabilité demandée est égale à

$$\binom{n}{k} p^k (1-p)^{n-k} \qquad \textbf{(formule du binôme)}$$

Il en résulte également que la probabilité que A se réalise au plus k fois au cours de n expériences est donnée par

$$\sum_{r=0}^{k} \binom{n}{r} p^r (1-p)^{n-r}$$

(pour $k = n$, cette probabilité vaut 1 : justifier).

Cette variable aléatoire est appelée *variable binomiale de paramètre p et d'exposant n et est notée $B(n,p)$*. Il y a autant de variables binomiales qu'il y a de couples (n,p) où n est un entier positif et $p \in [0,1]$. Pour $n = 1$, on retrouve les variables indicatrices.

Calcul de la moyenne

$$\begin{aligned}
\mu &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \underbrace{\sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^\ell (1-p)^{n-1-\ell}}_{1} \qquad \ell = k-1 \\
&= np.
\end{aligned}$$

Calcul de la variance

$$\begin{aligned}
\sigma^2 &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - \mu^2 \\
&= \sum_{k=2}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} - \mu^2
\end{aligned}$$

$\ell = k - 2$

$$\begin{aligned}
&= n(n-1)\, p^2 \underbrace{\sum_{\ell=0}^{n-2} \binom{n-2}{\ell} p^\ell (1-p)^{n-2-\ell}}_{1} + \mu - \mu^2 \\
&= n^2 p^2 - np^2 + np - n^2 p^2 \\
&= np(1-p)
\end{aligned}$$

Calcul de la fonction génératrice

$$\begin{aligned}
\psi(t) &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k q^{n-k} \qquad\qquad (q = 1-p) \\
&= (p\, e^t + q)^n
\end{aligned}$$

(Exercice : retrouver la moyenne et la variance à partir de la fonction génératrice).

Stabilité : la somme de 2 variables binomiales indépendantes de paramètre p est encore une variable binomiale de paramètre p.
On dit que la famille des variable binomiales de paramètre p est stable.

Démonstration.

Soit $B(n_1,p)$ et $B(n_2,p)$ deux variables binomiales indépendantes de paramètre p et d'exposants n_1 et n_2; soit $\psi_1(t)$ et $\psi_2(t)$ leurs fonctions génératrices. La fonction

génératrice de leur somme vaut

$$\begin{aligned}\psi(t) &= \psi_1(t)\cdot\psi_2(t) \qquad \text{(justifier)}\\ &= (pe^t+q)^{n_1}\cdot(pe^t+q)^{n_2}\\ &= (pe^t+q)^{n_1+n_2}.\end{aligned}$$

C'est donc la fonction génératrice d'une binomiale de paramètre p et d'exposant $n_1 + n_2$. Comme les fonctions génératrices caractérisent entièrement les variables aléatoires, la somme de $B(n_1,p)$ et $B(n_2,p)$ est égale à $B(n_1 + n_2,p)$.

Allure de la distribution de probabilité.

En faisant le rapport de 2 probabilités successives, on obtient

$$\frac{P[B(n,p)=k]}{P[B(n,p)=k-1]} = \frac{n-k+1}{k}\cdot\frac{p}{1-p}.$$

Ce rapport est > 1 pour tout $k < np + p$
$= 1$ pour $k = np + p$
< 1 pour tout $k > np + p$.

Il en résulte que $B(n,p)$ a une distribution en i si $np + 1 \leq 1$, en j si $np + p \geq n$ et en cloche dans les autres cas. Si $np + p$ est entier, on a 2 modes consécutifs pour $k = np + p - 1$ et $k = np + p$. Si $np + p$ n'est pas entier, on a 1 mode en la valeur entière située entre $np + p - 1$ et $np + p$.

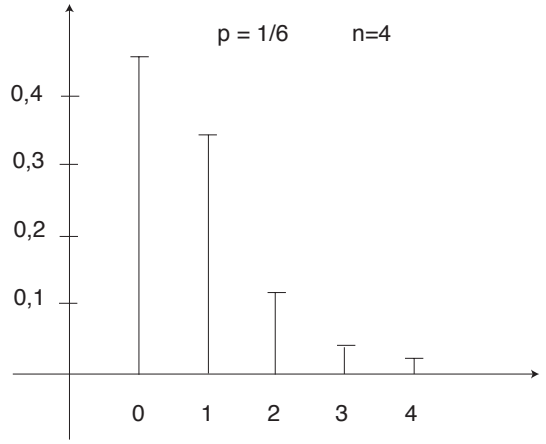
Remarquons également que le rapport précédent fournit une formule de récurrence pour calculer de proche en proche les probabilités successives de la distribution.

On peut démontrer que, pour une $B(n,p)$:

$$\mu_3 = n\,p\,q(q-p) \qquad \mu_4 = n\,p\,q(1-6pq+3npq)$$
$$\gamma_1 = \frac{q-p}{\sqrt{npq}} \qquad \gamma_2 = \frac{(1-6pq)}{npq}$$

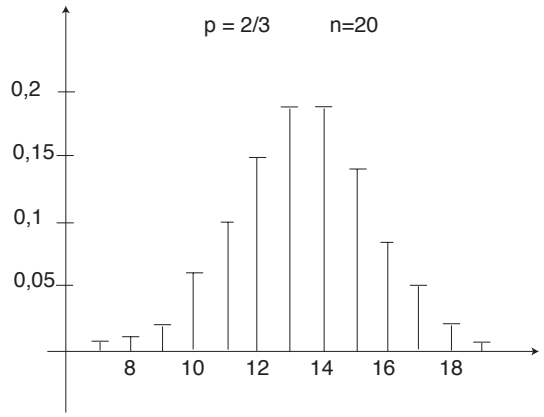
Il en résulte que la distribution de probabilité est symétrique lorsque $p = 1/2$ $(\mu_3 = 0)$ et qu'elle présente une dissymétrie gauche $(\mu_3 > 0)$ ou droite $(\mu_3 < 0)$ suivant que $p < 1/2$ ou $p > 1/2$.

Exemples numériques



Distribution en $i(np + p < 1)$
Mode unique à l'extrémité gauche

$$\mu = 2/3 \qquad \sigma^2 = 5/9$$
$$\mu_3 = 10/27 \quad \mu_4 = 55/54$$
$$\gamma_1 = 2/\sqrt{5} \quad \gamma_2 = 3/10$$



Distribution en cloche avec 2 modes consécutifs égaux ($np + p$ entier, > 1 et $< n$)

$$\mu = \frac{40}{3} \qquad \sigma^2 = \frac{40}{9}$$
$$\mu_3 = -\frac{40}{27} \qquad \mu_4 = \frac{520}{9}$$
$$\gamma_1 = -\frac{1}{\sqrt{40}} \quad \gamma_2 = -\frac{3}{40}$$

Remarque : une variable binomiale de paramètre p et d'exposant n peut être vue comme la somme de n variables indicatrices indépendantes de paramètre p (justifier). En se basant sur cette constatation, on peut retrouver aisément la moyenne, la variance, la fonction génératrice et la stabilité (à faire comme exercice).

Lecture des tables : voir exercices.

3.5 Variable hypergéométrique

Une urne contient N_1 boules rouges et N_2 boules blanches ($N = N_1 + N_2$); on prélève n boules dans l'urne ($n \leq N$) *sans les replacer* et on s'intéresse à la variable aléatoire "nombre de boules rouges prélevées".

C'est une variable discrète dont les valeurs possibles sont $0,1,2,\ldots,n$ et dont la distribution de probabilité est caractérisée par

$$\mathbb{P}[\text{ variable } = k] = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}} \qquad \text{(justifier)}$$

Cette variable aléatoire est appelée *variable hypergéométrique*. Ses propriétés sont semblables à celles des variables binomiales.

On démontre d'autre part que si $N_1 \to \infty$ et $N_2 \to \infty$ avec $\dfrac{N_1}{N} \to p$ et $\dfrac{N_2}{N} \to 1-p$, la formule précédente coïncide avec la formule du binôme, ce qui est intuitivement évident puisque cette dernière correspond à l'expérience où le prélèvement des boules se fait *avec remplacement*.

Lorsque N est grand, le calcul des probabilités relatives aux distributions hypergéométriques est long, alors que l'écart par rapport aux distributions binomiales est faible. Celles-ci sont donc en général utilisées comme approximation, dès que $N > 10n$.

C'est en particulier le cas dans les *sondages*, bien que ceux-ci soient le plus souvent des tirages sans remplacement.

3.6 Variable de Poisson

Si l'on considère une binomiale $B(n,p)$ dont le paramètre p tend vers 0, dont l'exposant n tend vers ∞ et dont le produit np tend vers un nombre λ, on trouve

$$P[\text{ variable } = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,2,\ldots$$

En effet, il suffit de remarquer que $\binom{n}{k}p^k(1-p)^{n-k}$ peut encore s'écrire

$$\frac{(np)^k}{k!}\left(1-\frac{np}{n}\right)^n \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)}{(1-p)^k}.$$

On obtient donc la distribution de probabilité d'une variable discrète dont les valeurs possibles sont tous les entiers non négatifs : c'est la *variable de Poisson de paramètre λ; nous la noterons P_λ*. Il y a autant de variables de Poisson qu'il y a de valeurs réelles non négatives pour λ.

Calcul de la moyenne

$$\begin{aligned}
\mu &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda}\underbrace{\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}}_{e^\lambda} \\
&= \lambda
\end{aligned}$$

Calcul de la variance

$$\begin{aligned}
\sigma^2 &= \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} - \mu^2 \\
&= \lambda^2 e^{-\lambda}\underbrace{\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}}_{e^\lambda} + \mu - \mu^2 \\
&= \lambda
\end{aligned}$$

Calcul de la fonction génératrice

$$\begin{aligned}\psi(t) &= \sum_{k=0}^{\infty} e^{kt} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda} \cdot e^{\lambda e^t} \\ &= e^{\lambda(e^t-1)}.\end{aligned}$$

Stabilité : la famille des variables de Poisson est stable.

Soit P_{λ_1} et P_{λ_2} deux variables de Poisson indépendantes et $\psi_1(t)$ et $\psi_2(t)$ leurs fonctions génératrices. La fonction génératrice de leur somme vaut

$$\begin{aligned}\psi(t) &= \psi_1(t) \cdot \psi_2(t) \\ &= e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} \\ &= e^{(\lambda_1+\lambda_2)(e^t-1)}.\end{aligned}$$

C'est donc la fonction génératrice d'une Poisson de paramètre $\lambda_1 + \lambda_2$. Comme les fonctions génératrices caractérisent entièrement les variables aléatoires, la somme de P_{λ_1} et P_{λ_2} est égale à $P_{\lambda_1+\lambda_2}$.

Allure de la distribution de probabilité

En faisant le rapport de 2 probabilités successives, on obtient

$$\frac{P\left[P_\lambda = k\right]}{P\left[P_\lambda = k-1\right]} = \frac{\lambda}{k}.$$

Il en résulte que P_λ a une distribution en i si $\lambda \leq 1$ et une distribution en cloche si $\lambda > 1$. Si λ est entier, on a 2 modes consécutifs en $k = \lambda - 1$ et $k = \lambda$. Si λ n'est pas

entier, on a 1 mode en la valeur entière située entre $\lambda - 1$ et λ. Le rapport précédent fournit également une formule de récurrence pour calculer de proche en proche les probabilités successives de la distribution.

On peut démontrer que, pour une P_λ :

$$\mu_3 = \lambda \qquad \mu_4 = 3\lambda^2 + \lambda$$
$$\gamma_1 = \frac{1}{\sqrt{\lambda}} \quad \gamma_2 = \frac{1}{\lambda}$$

Il en résulte que les distributions de probabilité de toutes les variables de Poisson présentent une dissymétrie gauche d'autant plus marquée que λ est faible.

Le tableau suivant illustre le passage des distributions binomiales aux distributions de Poisson (pour $\lambda = 1$). En pratique, lorsque n est suffisamment grand et p suffisamment petit, $B(n,p)$ peut être remplacée sans erreur importante par P_λ (où $\lambda = np$), qui est plus facile à manier (voir exercices).

k	Distributions binomiales						Distribution
	$n = 5$ $p = 1/5$	$n = 10$ $p = 1/10$	$n = 20$ $p = 1/20$	$n = 40$ $p = 1/40$	$n = 80$ $p = 1/80$	...	de Poisson : $\lambda = 1$
0	0,3277	0,3487	0,3585	0,3632	0,3656	...	0,3679
1	0,4096	0,3874	0,3774	0,3726	0,3702	...	0,3679
2	0,2048	0,1937	0,1887	0,1863	0,1851	...	0,1839
3	0,0512	0,0574	0,0596	0,0605	0,0609	...	0,0613
4	0,0064	0,0112	0,0133	0,0143	0,0148	...	0,0153
5	0,0003	0,0015	0,0022	0,0026	0,0029	...	0,0031
6	–	0,0001	0,0003	0,0004	0,0004	...	0,0005
7	–	0,0000	0,0000	0,0001	0,0001	...	0,0001
8	–	0,0000	0,0000	0,0000	0,0000	...	0,0000
⋮	–	⋮	⋮	⋮	⋮		⋮
Totaux	1,0000	1,0000	1,0000	1,0000	1,0000	...	1,0000

Lecture des tables : voir exercices

Applications

Considérons un évènement aléatoire pouvant se réaliser au cours de chaque intervalle de temps Δt avec une probabilité $a\Delta t$. Le nombre de réalisations de cet événement au cours de n intervalles de temps consécutifs est donc une variable binomiale de paramètre $a\Delta t$ et d'exposant n à condition que l'événement se réalise au plus une fois dans chaque intervalle de temps. Si l'événement peut se produire à tout instant, on doit considérer des intervalles Δt infiniment petits et un nombre n infiniment grand, le produit $n\Delta t$ restant égal à l'intervalle de temps global considéré. Dans ce cas, le nombre de réalisations de l'événement devient une variable de Poisson. Cette variable est utilisée dans de nombreux problèmes; elle peut représenter le nombre d'arrivées de clients à un guichet, de bateaux dans un port, de voitures à un feu de circulation, de programmes à traiter par l'unité centrale d'un ordinateur, le nombre de pièces prises dans un stock, ...

Le raisonnement précédent peut aussi être appliqué au cas où l'on considère le nombre de réalisations d'un certain événement sur une surface donnée (que l'on divise en surfaces infiniment petites) ou dans un volume donné (que l'on divise en volumes infiniment petits). (**Exemple :** nombre de défauts sur une surface d'acier ou de carton,...)

N.B. : il peut arriver que la probabilité de réalisation d'un événement soit plus élevée à certains moments qu'à d'autres; dans ce cas, le modèle de la distribution de Poisson ne peut plus s'appliquer, mais d'autres distributions ont été définies (binomiales négatives, de Poisson-Pascal,...).

3.7 Variable exponentielle négative

Considérons à nouveau un événement aléatoire pouvant se réaliser au cours de chaque intervalle de temps Δt avec une probabilité $a\ \Delta t$ et soit $F(t)$ la probabilité que l'attente avant la prochaine réalisation soit inférieure ou égale à t.
On a

$$F(t+\Delta t) - F(t) = (1 - F(t)) \cdot a\Delta t \qquad \text{(justifier)}$$

d'où

$$F'(t) = (1 - F(t)) \cdot a \qquad (t > 0).$$

On résout cette équation différentielle en tenant compte du fait que $F(0) = 0$ et on trouve

$$F(t) = \begin{cases} 1 - e^{-at} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Il s'agit donc de la fonction de répartition de la variable "temps d'attente avant la prochaine réalisation". La densité de probabilité correspondante est

$$f(t) = F'(t) = \begin{cases} ae^{-at} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Les variables exponentielles négatives et les variables de Poisson sont étroitement liées. Dans notre exemple, les unes concernent les intervalles de temps entre 2 réalisations successives d'un événement et les autres les nombres de réalisations de l'événement au cours d'un intervalle de temps donné.

Définition

La variable exponentielle négative de paramètre $a > 0$ est une variable continue, notée Exp_a, dont la densité de probabilité est

$$f(x) = \begin{cases} ae^{-ax} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Calcul de la moyenne

$$\begin{aligned} \mu &= \int_0^\infty xae^{-ax}\,dx \\ &= \left[-xe^{-ax}\right]_0^\infty + \int_0^\infty e^{-ax}\,dx \\ &= \left[-\frac{e^{-ax}}{a}\right]_0^\infty = \frac{1}{a} \end{aligned}$$

Calcul de la variance

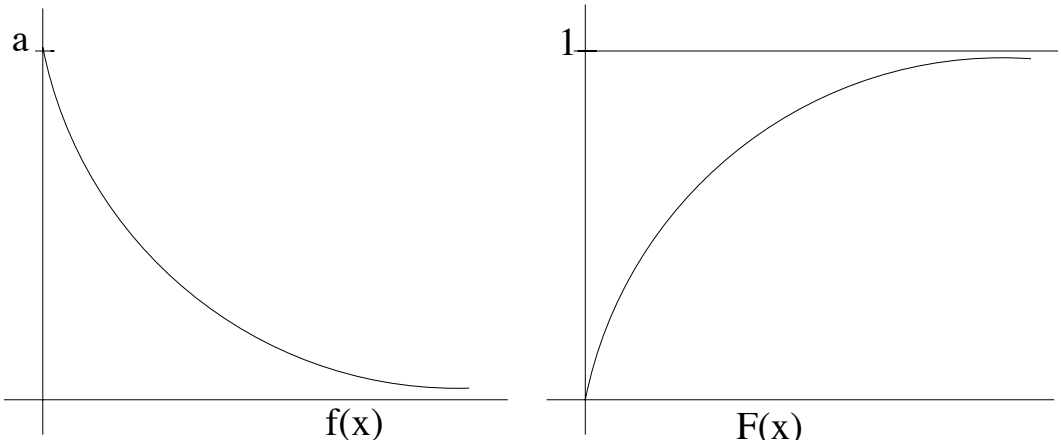
$$\begin{aligned}
\sigma^2 &= \int_0^\infty x^2 a e^{-ax}\,dx - \mu^2 \\
&= \left[-x^2 e^{-ax}\right]_0^\infty + 2\int_0^\infty x e^{-ax}\,dx - \mu^2 \\
&= 2\frac{\mu}{a} - \mu^2 \\
&= \frac{1}{a^2}
\end{aligned}$$

Calcul de la fonction génératrice

$$\begin{aligned}
\psi(t) &= \int_0^\infty e^{xt} a e^{-ax}\,dx \\
&= \frac{a}{t-a}\left[e^{(t-a)x}\right]_0^\infty \qquad \text{(pour } t < a\text{)} \\
&= \frac{a}{a-t}
\end{aligned}$$

Stabilité : la somme de 2 exponentielles négatives indépendantes n'est pas une exponentielle négative car le produit de $\frac{a}{a-t}$ par $\frac{b}{b-t}$ n'est pas du type $\frac{c}{c-t}$.

Allures de la densité de probabilité et de la fonction de répartition



Propriété : soit V une exponentielle négative de paramètre a. Il vient

$$\begin{aligned}\mathbb{P}\left[V \leq t+\Delta t / V > t\right] &= \frac{\mathbb{P}\left[t < V \leq t+\Delta t\right]}{\mathbb{P}\left[V > t\right]} \\ &= \frac{1-e^{-a(t+\Delta t)}-1+e^{-at}}{e^{-at}} = 1-e^{-a\Delta t} = \mathbb{P}\left[V \leq \Delta t\right].\end{aligned}$$

La probabilité que l'événement survienne dans l'intervalle Δt à venir ne dépend pas du temps t que l'on a déjà attendu. Il s'agit de la "propriété d'oubli", caractéristique de cette variable.

3.8 Variable normale

La variable normale est de loin la plus importante et la plus utilisée en statistique. Cela résulte du fait que de très nombreux phénomènes aléatoires peuvent être modélisés par cette variable. En effet, tout variable mesurant un phénomène qui dépend d'un grand nombre de causes aléatoires aura approximativement un comportement de variable normale : c'est l'objet du théorème central-limite qui sera vu dans le chapitre suivant. On verra également que la variable normale permet, sous certaines conditions, de faciliter les calculs relatifs aux variables binomiales et de Poisson.

Définition

La variable normale de paramètres a (quelconque) et b (> 0) est une variable continue, notée $N(a,b)$, dont la densité de probabilité est

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2b^2}}, \qquad x \in \mathbb{R}$$

Calcul de la moyenne

$$\begin{aligned}
\mu &= \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-a)^2}{2b^2}}\, dx \qquad y = \frac{x-a}{b\sqrt{2}} \\
&= \frac{b\sqrt{2}}{\sqrt{\pi}} \underbrace{\int_{-\infty}^{\infty} y e^{-y^2}\, dy}_{0} + \frac{a}{\sqrt{\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-y^2}\, dy}_{\sqrt{\pi}} \\
&= a
\end{aligned}$$

Calcul de la variance

$$\begin{aligned}
\sigma^2 &= \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-a)^2 e^{-\frac{(x-a)^2}{2b^2}}\, dx \\
&= 0 + \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} b^2 e^{-\frac{(x-a)^2}{2b^2}}\, dx && \text{(par parties)} \\
&= \frac{b^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\, dy = b^2 && \text{(où } y = \frac{x-a}{b\sqrt{2}}\text{)}
\end{aligned}$$

Calcul de la fonction génératrice

$$\begin{aligned}
\psi(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \frac{e^{\mu t} \cdot e^{\frac{\sigma^2 t^2}{2}}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}} - \frac{\sigma t}{\sqrt{2}}\right)^2}\, dx \\
&= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\, dy = e^{\mu t + \frac{\sigma^2 t^2}{2}} && \text{(où } y = \frac{x-\mu}{\sigma\sqrt{2}} - \frac{\sigma t}{\sqrt{2}}\text{)}
\end{aligned}$$

Stabilité : la famille des variables normales est stable.

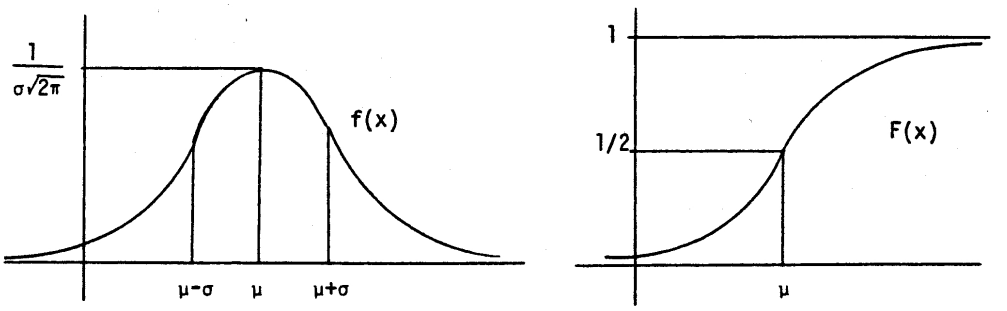
Soit $N(\mu_1,\sigma_1)$ et $N(\mu_2,\sigma_2)$ deux variables normales indépendantes et $\psi_1(t)$ et $\psi_2(t)$ leurs fonctions génératrices.
La fonction génératrice de leur somme vaut

$$\begin{aligned}\psi(t) &= \psi_1(t)\cdot\psi_2(t)\\ &= e^{\mu_1 t+\sigma_1^2 t^2/2}\cdot e^{\mu_2 t+\sigma_2^2 t^2/2}\\ &= e^{(\mu_1+\mu_2)t+\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}\end{aligned}$$

C'est donc la fonction génératrice d'une normale de paramètres $(\mu_1+\mu_2)$ et $\sqrt{\sigma_1^2+\sigma_2^2}$. Comme les fonctions génératrices caractérisent entièrement les variables aléatoires, la somme de $N(\mu_1,\sigma_1)$ et $N(\mu_2,\sigma_2)$ est égale à $N(\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2})$.

Allures de la densité de probabilité et de la fonction de répartition

L'étude de la fonction $y=f(x)$ et de ses dérivées montre que la densité de probabilité de $N(\mu,\sigma)$ est une courbe en cloche symétrique par rapport à la verticale $x=\mu$, possédant un maximum en $x=\mu$ (qui vaut $\frac{1}{\sigma\sqrt{2\pi}}$) et deux points d'inflexion en $x=\mu+\sigma$ et $x=\mu-\sigma$; elle est d'autre part asymptotique à l'axe des x.
La fonction de répartition est croissante (comme toute fonction de répartition), possède un point d'inflexion en $x=\mu$ (où elle vaut $1/2$) et est asymptotique aux horizontales $y=0$ et $y=1$.



Les paramètres de dissymétrie et d'aplatissement de Fisher sont tous les deux nuls : $\gamma_1=\gamma_2=0$. C'est d'ailleurs la valeur obtenue pour la normale qui justifie l'introduction du terme -3 dans la définition de γ_2. A variances égales, les courbes symétries

telles que $\gamma_2 > 0$ sont plus pointues que la normale (courbes leptocurtiques) et celles pour lesquelles $\gamma_2 < 0$ sont plus aplaties (courbes platycurtiques).

Les résultats suivants sont intéressants à retenir :

$$P\left[\mu - \tfrac{2}{3}\sigma \leq N(\mu,\sigma) \leq \mu + \tfrac{2}{3}\sigma\right] = 0{,}50$$
$$P\left[\mu - \sigma \leq N(\mu,\sigma) \leq \mu + \sigma\right] = 0{,}69$$
$$P\left[\mu - 2\sigma \leq N(\mu,\sigma) \leq \mu + 2\sigma\right] = 0{,}95$$
$$P\left[\mu - 3\sigma \leq N(\mu,\sigma) \leq \mu + 3\sigma\right] = 0{,}99$$

Parmi la famille des variables normales, il en est une qui joue un rôle particulier : il s'agit de la variable normale réduite $(\mu = 0,\sigma = 1)$ encore appelée *variable de Gauss*. La connaissance du comportement de cette variable suffit pour étudier n'importe quelle autre variable, puisque toute variable ayant une distribution $N(\mu,\sigma)$ est égale à $\mu+\sigma V$, où V a la distribution $N(0,1)$ (justifier).

Lecture des tables voir exercices.

3.9 Variable Gamma

Définition

La variable gamma de paramètres $\alpha(> 0)$ et $\beta(> 0)$, notée $\Gamma(\alpha,\beta)$, est une variable continue dont la densité de probabilité est

$$f(x) = \begin{cases} \dfrac{\alpha^\beta x^{\beta-1} e^{-ax}}{\Gamma(\beta)} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

où

$$\Gamma(\beta) = \int_0^\infty u^{\beta-1} e^{-u}\, du \qquad \text{(intégrale eulérienne)}$$

L'intérêt de l'introduction de cette variable est essentiellement théorique; pour $\beta = 1$, on retrouve l'exponentielle négative.
Nous verrons aussi qu'elle intervient dans l'étude de la variable χ^2. Dans ce but, notons

que sa fonction génératrice est donnée par

$$\begin{aligned} \psi(t) &= \int_0^\infty \frac{e^{xt}\alpha^\beta x^{\beta-1}e^{-\alpha x}}{\Gamma(\beta)}\,dx \\ y = (\alpha - t)x & \qquad\qquad (t < \alpha) \\ &= \frac{\alpha^\beta}{\Gamma(\beta)}\int_0^\beta e^{-y}\frac{y^{\beta-1}}{(\alpha-t)^\beta}\,dy \\ &= (1-\frac{t}{\alpha})^{-\beta} \qquad\qquad (\text{ pour } t<\alpha) \end{aligned}$$

3.10 Variable χ^2

Définition

La variable χ^2 à n degrés de liberté est la somme des carrés de n variables normales réduites indépendantes. Elle se présente de façon naturelle dans certaines applications (voir inférence statistique); nous la noterons $\chi^2_{(n)}$.

Recherche de la densité de probabilité

La densité de probabilité du carré d'une $N(0,1)$ est donnée par

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}} & x > 0 \\ 0 & x \le 0 \end{cases} \qquad \text{(justifier)}$$

Elle est donc identique à la densité de probabilité d'une $\Gamma(\frac{1}{2},\frac{1}{2})$. Il en résulte que sa fonction génératrice est

$$(1-2t)^{-1/2}$$

et que la fonction génératrice de $\chi^2_{(n)}$ est

$$(1-2t)^{-\frac{n}{2}} \qquad\qquad (\text{justifier})$$

On constate donc que la $\chi^2_{(n)}$ n'est rien d'autre qu'une $\Gamma(\frac{1}{2},\frac{n}{2})$; par conséquent, sa densité de probabilité vaut

$$f(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{n/2} \cdot \Gamma(\frac{n}{2})} & x > 0 \\ 0 & x \leq 0. \end{cases}$$

Moyenne et variance

Elles peuvent se calculer directement; elles peuvent aussi se déduire du fait que la moyenne et la variance du carré d'une $N(0,1)$ valent respectivement 1 et 2 (justifier ces 2 valeurs et expliquer comment se fait la déduction).
On obtient

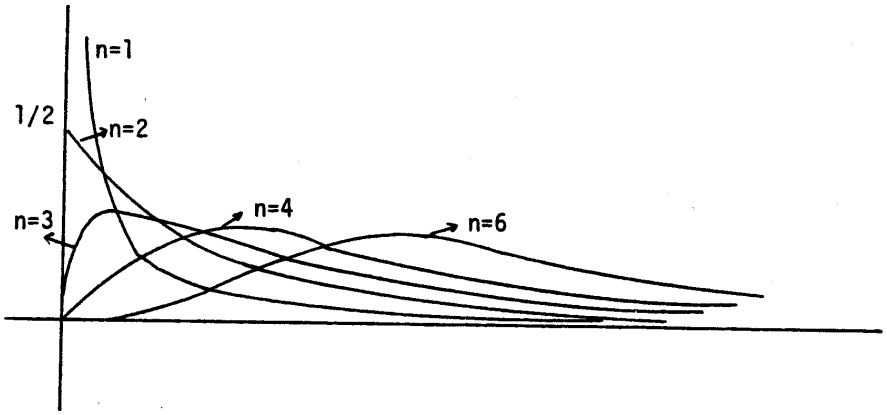
$$E(\chi^2_{(n)}) = n \qquad\qquad D^2(\chi^2_{(n)}) = 2n.$$

Stabilité

La famille des variables χ^2 est stable. Cela s'établit soit en considérant les fonctions génératrices, soit directement à partir de la définition (justifier).

Allure de la densité de probabilité

Les variables $\chi^2_{(n)}$ présentent toutes une dissymétrie gauche. Leur densité de probabilité est une courbe en i pour $n = 1$ et 2 et une courbe en cloche pour $n \geq 3$, le mode se situant en $n-2$. La courbe est tangente à l'origine à l'axe des ordonnées pour $n = 3$ et à l'axe des abscisses pour $n > 4$.
Le calcul donne $\gamma_1 = \dfrac{\sqrt{8}}{n}$ et $\gamma_2 = \dfrac{12}{n}$.



Lecture des tables : voir exercices.

3.11 Variable de Student

Définition

La variable de Student à n degrés de liberté, notée $t_{(n)}$ est le rapport $\dfrac{V}{\sqrt{\dfrac{1}{n}\sum_{i=1}^{n} V_i^2}}$,

où V,V_1,V_2, . . . ,V_n sont des variables $N(0,\sigma)$ indépendantes. Elle se présente de façon naturelle dans certaines applications (cf. inférence statistique).

Recherche de la densité de probabilité : voir exercices

On obtient

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

N.B. : elle est indépendante de σ.

Moyenne $= 0 \qquad (n > 1)$

Variance $= \dfrac{n}{n-2} \qquad (n > 2)$

Allure de la densité de probabilité

Courbe en cloche symétrique par rapport à l'axe des ordonnées. Sa variance est supérieure à 1 et son coefficient d'aplatissement $\gamma_2 = \dfrac{6}{n-4}$ $(n > 4)$ est > 0 : elle est donc plus dispersée et plus pointue que la $N(0,1)$ mais, lorsque n augmente, elle s'en rapproche. On verra d'ailleurs que les variables de Student sont "asymptotiquement normales".

Lecture des tables : voir exercices.

3.12 Variable de Snedecor

Définition

La variable de Snedecor à m et n degrés de liberté, notée $F_{m,n}$ est le rapport $\dfrac{\dfrac{1}{m}\displaystyle\sum_{i=1}^{m} V_i^2}{\dfrac{1}{n}\displaystyle\sum_{j=1}^{n} W_j^2}$, où $V_1, V_2, \ldots, V_m, W_1, W_2, \ldots, W_n$ sont des variables $N(0,\sigma)$ indépendantes.
Elle se présente de façon naturelle dans certaines applications (cf. inférence statistique).

Densité de probabilité

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right) \cdot \Gamma\left(\dfrac{n}{2}\right)} \left(\dfrac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \dfrac{m}{n}x\right)^{-\frac{m+n}{2}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Moyenne $= \dfrac{n}{n-2}$ $\qquad (n > 2)$

Variance $= \dfrac{2n^2(m+n-2)}{m(n-2)^2(m-4)}$ $\qquad (n > 4)$

Allure de la densité de probabilité

Dissymétrie gauche; courbe en i pour $m \leq 2$ et en cloche pour $m > 2$; tangente à l'origine à l'axe des ordonnées pour $m = 3$ et à l'axe des abscisses pour $m > 4$.

Lecture des tables : voir exercices

Chapitre 4

Vecteurs aléatoires particuliers

4.1 La normale à deux dimensions

La densité de probabilité d'une normale à deux dimensions est définie par

$$f(x,y) = \frac{1}{2\pi b_1 b_2 \sqrt{1-c}}\, e^{-\frac{Q}{2(1-c^2)}},$$

où

$$Q = \left(\frac{x-a_1}{b_1}\right)^2 - 2c\frac{x-a_1}{b_1}\frac{y-a_2}{b_2} + \left(\frac{y-a_2}{b_2}\right)^2,$$

et où

$$b_1 > 0, b_2 > 0, \text{ et } -1 < c < 1.$$

On peut démontrer les propriétés suivantes :

- a_1 et a_2 sont les deux moyennes marginales μ_1 et μ_2;
- b_1 et b_2 sont les deux écarts-types marginaux σ_1 et σ_2;
- c est le coefficient de corrélation ρ;
- $f(x,y)$ définit une surface en cloche asymptotique au plan (x,y), le sommet de la cloche correspondant au point moyen (μ_1,μ_2);
- les sections horizontales dans la surface sont des ellipses (des circonférences si $\rho = 0$ et $\sigma_1 = \sigma_2$);

- les distributions marginales et conditionnelles sont des normales;
- les variances conditionnelles sont constantes et se confondent avec les variances résiduelles;
- il suffit que $\rho = 0$ pour que les deux distributions marginales soient indépendantes (**particularité importante des normales**).

4.2 Variables binomiales et hypergéométriques généralisées

Une urne contient N_i boules de couleur i ($i = 1,2,\ldots,m$) et $N = \sum_{i=1}^{m} N_i$; on prélève n boules dans l'urne et on s'intéresse à la probabilité d'obtenir k_i boules de couleur i ($i = 1,2,\ldots,m$), avec $\sum_{i=1}^{m} k_i = n$.

Si le tirage se fait avec remplacement, cette probabilité est égale à

$$\frac{N!}{k_1!k_2!\ldots k_m!}\prod_{i=1}^{m}\left(\frac{N_i}{N}\right)^{k_i} \qquad \text{(justifier)}$$

Si le tirage se fait sans remplacement, on obtient

$$\frac{\prod_{i=1}^{m}\binom{N_i}{k_i}}{\binom{N}{n}} \qquad \text{(justifier)}$$

Les distributions de probabilité qui en découlent s'appellent respectivement *distribution polynomiale* (ou binomiale généralisée) et *distribution hypergéométrique généralisée*. Ce sont des distributions à $m-1$ dimensions qui, pour $m = 2$, coïncident respectivement avec la distribution binomiale et la distribution hypergéométrique.

Chapitre 5

Quelques théorèmes fondamentaux de la théorie des probabilités

5.1 L'inégalité de Bienayme-Tchebycheff

Enoncé *: pour toute variable aléatoire V de moyenne μ et d'écart-type σ et pour toute constante positive k, on a :*

$$P\left[|V-\mu| \geq k\sigma\right] \leq \frac{1}{k^2}.$$

Interprétation : cette propriété montre que, dans toute distribution, la proportion d'individus s'écartant de la moyenne de plus de k fois l'écart-type est toujours inférieure à $\frac{1}{k^2}$. On a donc, par exemple, pour toute variable V :

$$P\left[\mu - 2\sigma < V < \mu + 2\sigma\right] \geq 0{,}75$$

$$P\left[\mu - 3\sigma < V < \mu + 3\sigma\right] \geq 0{,}88$$

(voir en particulier les valeurs obtenues pour les normales).

Démonstration dans le cas continu (la démonstration est analogue dans le cas discret)

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\
&= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\,dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)\,dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 f(x)\,dx \\
&\geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)\,dx + \int_{\mu+k\sigma}^{+\infty} k^2\sigma^2 f(x)\,dx \\
&\geq k^2\sigma^2 P\left[|V-\mu| \geq k\sigma\right]
\end{aligned}$$

5.2 Le Théorème de Bernoulli (ou loi des grands nombres)

Enoncé *: soit n répétitions d'une expérience aléatoire au cours desquelles un événement A peut chaque fois se produire avec une probabilité p; la probabilité que la fréquence relative de A, au cours de ces n répétitions, diffère de p d'au moins une quantité arbitraire ϵ, tend vers 0 lorsque $n \to \infty$:*

$$P\left[|\frac{F}{n} - p| \geq \epsilon\right] n \underset{\to}{\longrightarrow} \infty 0$$

Interprétation : ce résultat n'est rien d'autre que la formalisation du phénomène de régularité statistique puisqu'il montre que la fréquence relative d'un événement converge vers sa probabilité.

Démonstration

F est une variable binomiale de paramètre p et d'exposant n (voir la définition de la binomiale); sa moyenne est np et sa variance npq. Par conséquent, la moyenne et la variance de $\frac{F}{n}$ sont respectivement égales à p et $\frac{pq}{n}$. En appliquant l'inégalité de Bienaymé-Tchebycheff avec $k = \epsilon\sqrt{\frac{n}{pq}}$, on obtient

$$P\left[|\frac{F}{n} - p| \geq \epsilon\right] \leq \frac{pq}{n\epsilon^2} \quad n \underset{\to}{\longrightarrow} \infty 0$$

5.3 Le théorème de de Moivre

Enoncé *: lorsque* $n \to \infty$*, la distribution de probabilité de la binomiale* $B(n,p)$ **réduite** *tend vers la densité de probabilité de la normale* $N(0,1)$*; on dit que la variable binomiale est* **asymptotiquement normale**.

$$\frac{B(n,p) - np}{\sqrt{npq}} \underset{n \to \infty}{\longrightarrow} N(0,1).$$

(ce théorème ne sera pas démontré ici, mais il est un cas particulier du Théorème Central-Limite présenté plus loin).

Remarque : le théorème de Bernoulli est un corollaire du théorème de de Moivre. En effet l'inégalité

$$|\frac{F}{n} - p| \geq \epsilon$$

est équivalente à

$$|\frac{F - np}{\sqrt{npq}}| \geq \epsilon \sqrt{\frac{n}{pq}}.$$

Pour tout nombre A positif, on peut choisir n tel que

$$\epsilon \sqrt{\frac{n}{pq}} > A,$$

d'où

$$P\left[|\frac{F}{n} - p| \geq \epsilon\right] \leq \mathbb{P}\left[|\frac{F - np}{\sqrt{npq}}| \geq A\right],$$

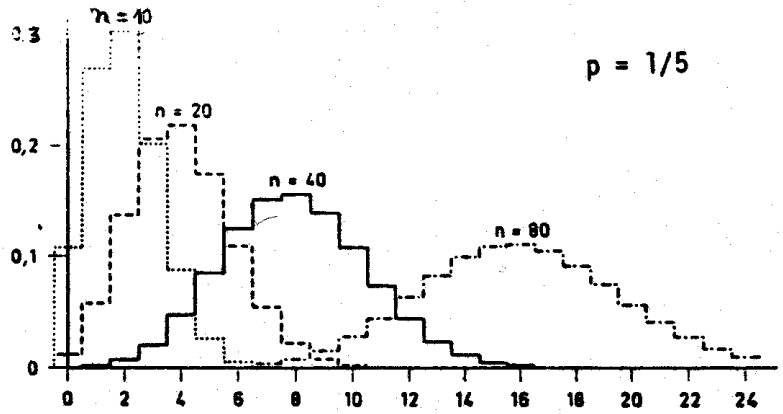
et donc, par le théorème de de Moivre :

$$\lim_{n \to \infty} P\left[|\frac{F}{n} - p| \geq \epsilon\right] \leq P\left[|N(0,1)| \geq A\right]$$

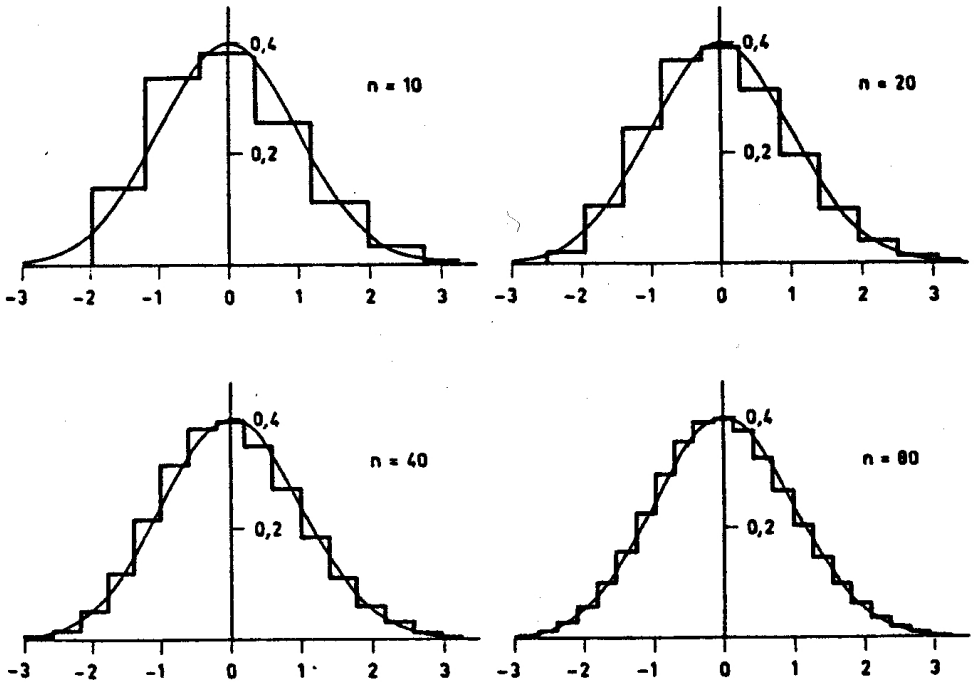
et on peut toujours choisir A de telle sorte que le $2^{\text{ème}}$ membre soit aussi petit que l'on veut.

Exemple : le 1[ier] graphique ci-dessous représente les distributions de probabilité de $B(10,\frac{1}{5})$, $B(20,\frac{1}{5})$, $B(40,\frac{1}{5})$ et $B(80,\frac{1}{5})$.

Les 4 graphiques suivants illustrent la convergence des distributions de probabilité des variables réduites correspondantes vers la densité de probabilité de la $N(0,1)$.



Variables réduites



5.4 Le Théorème Central-Limite

Enoncé *: sous certaines conditions très générales, la somme de n variables aléatoires indépendantes est une variable asymptotiquement normale :*

$$\left.\begin{array}{c} V_1,V_2,\ldots,V_n \text{ indépendantes} \\ V=\sum_{i=1}^{n} V_i \end{array}\right\} \Rightarrow \frac{V-\mu}{\sigma} \underset{n \to \infty}{\longrightarrow} N(0,1),$$

où μ et σ sont la moyenne et l'écart-type de V.

Interprétation : il est remarquable de constater que cette propriété est applicable quelles que soient les distributions des variables V_i. Ce théorème montre l'importance de la variable $N(0,1)$.

Il en résulte notamment que tout phénomène dépendant d'un grand nombre de causes aléatoires aura approximativement un comportement modélisable par une variable normale : c'est le cas par exemple des erreurs de mesure en physique ou en astronomie (qui peuvent être considérées comme résultant d'un grand nombre de petites erreurs additives et indépendantes), des tailles des organismes vivants (qui peuvent être considérées comme résultant de l'action d'un grand nombre de gènes), et de nombreuses autres variables relatives à des phénomènes biologiques, démographiques, économiques,...

Remarque : le théorème de de Moivre est un cas particulier du théorème central-limite puisqu'une variable binomiale peut toujours être considérée comme la somme de variables indicatrices indépendantes.

Démonstration : nous démontrons le théorème central-limite dans le cas particulier où les variables V_i ont la même distribution (c'est d'ailleurs dans ce cas-là que l'on va appliquer ce théorème en inférence statistique).

Soit

$$E(V_i)=\widehat{\mu}, D(V_i)=\widehat{\sigma}, \forall i,$$

d'où

$$E(V) = n\widehat{\mu}, D(V) = \sqrt{n}\widehat{\sigma},$$

et soit

$$W_i = V_i - \widehat{\mu}, \qquad W = \sum_{i=1}^{n} W_i, \qquad Z = \frac{W}{\sqrt{n}\widehat{\sigma}},$$

d'où

$$E(Z) = 0,\ D(Z) = 1.$$

On a

$$\begin{aligned} \psi_{W_i}(t) &= 1 + E(W_i) \cdot t + E(W_i^2)\frac{t^2}{2} + R(t^3) \\ &= 1 + 0 + \frac{\widehat{\sigma}^2 t^2}{2} + R(t^3), \end{aligned}$$

d'où

$$\psi_W(t) = \left[1 + \frac{\widehat{\sigma}^2 t^2}{2} + R(t^3)\right]^n,$$

ce qui implique

$$\begin{aligned} \psi_Z(t) &= E(e^{Zt}) = E\left(e^{\frac{Wt}{\sqrt{n}\widehat{\sigma}}}\right) = \psi_W\left(\frac{t}{\sqrt{n}\widehat{\sigma}}\right) \\ &= \left[1 + \frac{t^2}{2n} + R\left(\frac{t^3}{n^{3/2}}\right)\right]^n, \end{aligned}$$

et finalement

$$\lim_{n\to\infty} \psi_Z(t) = e^{t^2/2},$$

ce qui est bien la fonction génératrice d'une $N(0,1)$.

5.5 Quelques résultats asymptotiques sur les variables particulières

1. La binomiale est asymptotiquement normale (c'est le théorème de de Moivre)

$$\frac{B(n,p) - np}{\sqrt{npq}} \underset{n \to \infty}{\longrightarrow} N(0,1).$$

2. La variable de Poisson est asymptotiquement normale

$$\frac{P_\lambda - \lambda}{\sqrt{\lambda}} \underset{\lambda \to +\infty}{\longrightarrow} N(0,1).$$

3. La variable χ^2 est asymptotiquement normale mais, en général, on préfère utiliser le résultat suivant, où la convergence est beaucoup plus rapide

$$\sqrt{2\chi^2_{(n)}} - \sqrt{2n-1} \underset{n \to \infty}{\longrightarrow} N(0,1).$$

4. La variable de Student est asymptotiquement normale; en fait, lorsque le nombre de degrés de liberté est suffisamment grand, on peut directement assimiler la variable de Student à une $N(0,1)$:

$$t_{(n)} \underset{n \to \infty}{\longrightarrow} N(0,1).$$

5. Rappelons enfin que, par définition,

$$B(n,p) \underset{\substack{p \to 0 \\ np \to \lambda}}{\overset{n \to \infty}{\longrightarrow}} P_\lambda.$$

Règles pratiques habituellement suivies pour l'application de ces résultats

Le signe $\approx$ signifie ici "est assimilée à".

1. $\frac{B(n,p) - np}{\sqrt{npq}} \approx N(0,1)$ lorsque $np > 5$ et $nq > 5$.
2. $B(n,p) \approx P_{\lambda=np}$ lorsque $p < 0{,}1$ et $n > 30$.

3. Lorsque les deux approximations ci-dessus sont possibles, on préfère utiliser la 1ère.
4. $\dfrac{P_\lambda - \lambda}{\sqrt{\lambda}} \approx N(0,1)$ lorsque $\lambda > 15$
5. $\sqrt{2\chi^2_{(n)}} - \sqrt{2n-1} \approx N(0,1)$ lorsque $n > 30$
6. $t_{(n)} \approx N(0,1)$ lorsque $n > 60$.

(voir exemples aux exercices).

Troisième partie

L’inférence statistique

Introduction

L'inférence statistique permet d'obtenir de l'information sur un ensemble d'objets ou d'individus (la population) à partir d'informations recueillies sur une partie de cet ensemble (l'échantillon). L'information que l'on va considérer ici concernera

- l'estimation de paramètres (et la détermination de la précision de cette estimation);
- la vérification d'hypothèses (et la détermination du risque d'erreur associé).

Les outils et les modèles mathématiques qui sont à la base de l'inférence statistique sont fournis par la théorie des probabilités.

Population

La population est l'ensemble des individus, objets ou mesures auquel on s'intéresse. Elle peut être finie et bien définie (ensemble des habitants de ce pays, ensemble des tailles des habitants de cette ville, ensemble des voitures sortie de cette chaîne de fabrication,...). Elle peut être finie mais mal définie (ensemble des insectes de la forêt amazonienne). Elle peut être infinie (ensemble des prélèvements de 10 cl que l'on peut effectuer dans cet océan, ensemble des teneurs en sel dans ces prélèvements,...)

Echantillon

L'échantillon est la partie de la population qui est réellement observée. Pour pouvoir appliquer la théorie des probabilités, il est nécessaire que l'échantillon vérifie certaines conditions. On considérera ici des échantillons *aléatoires* et *simples*.

L'échantillon est *aléatoire* lorsque tous les individus de la population ont la même probabilité de faire partie de l'échantillon. Le procédé le plus couramment utilisé est l'emploi de tables de nombres aléatoires.

L'échantillon est *simple* lorsque les individus qui constituent l'échantillon sont prélevés dans une population indépendamment les uns des autres. Un échantillon prélevé sans remplacement dans une population finie n'est donc pas un échantillon simple. Néanmoins, en pratique, on peut considérer comme simple un échantillon prélevé dans une population dont l'effectif est nettement plus grand que celui de l'échantillon (à partir de 10 fois plus grand environ) (cf. distributions binomiales et hypergéométriques).

Distribution de la population

On considère une population d'individus (ensemble des habitants de ce pays) et l'expérience qui consiste à prélever au hasard un individu dans cette population. Il s'agit d'une expérience aléatoire. L'ensemble des résultats de cette expérience est la population de départ. Si, à chaque résultat, on associe un nombre (taille des habitants), on définit une variable aléatoire qui possède une distribution de probabilité discrète ou continue (suivant l'effectif de la population). Cette distribution de probabilité sera appelée la *distribution de la population*.

Convention

Sauf mention contraire, on supposera toujours qu'à tout individu d'une population est associé un nombre (dans certains cas, les individus seront eux-même des nombres). Le plus souvent d'ailleurs on travaillera directement sur ces nombres. Ainsi, lorsqu'on parlera d'individus x,y,x_1,x_2, . . . ,x_n, il faudra entendre "valeurs de ces individus".

Chapitre 1

Distributions échantillonnées

1.1 Introduction

On considère une population donnée et l'expérience qui consiste à prélever un échantillon d'effectif n dans cette population. Il s'agit d'une expérience aléatoire. L'ensemble des résultats de cette expérience est l'ensemble de tous les échantillons d'effectif n qui peuvent être prélevés dans la population. A chaque échantillon x_1,x_2, . . . ,x_n (cf. convention précédente : ce sont des nombres), on peut associer, par exemple la valeur x_1; cette valeur est différente d'un échantillon à l'autre. L'application qui, à chaque échantillon, associe la 1ère valeur de cet échantillon est donc une variable aléatoire, que nous noterons X_1. Comme toute variable aléatoire, elle possède une distribution de probabilité, discrète ou continue : on l'appellera *distribution échantillonnée de* X_1. On définit de la même manière les distributions échantillonnées de X_2 (variable aléatoire qui, à chaque échantillon d'effectif n, associe la 2ème valeur de cet échantillon), X_3,X_4, . . . ,X_n.

La distribution échantillonnée de X_1 est identique à la distribution de la population (justifier). Lorsque l'échantillon est aléatoire et simple (ce que nous supposerons toujours dans la suite), il en sera de même des distributions échantillonnées de X_2,X_3, . . . ,X_n (justifier).

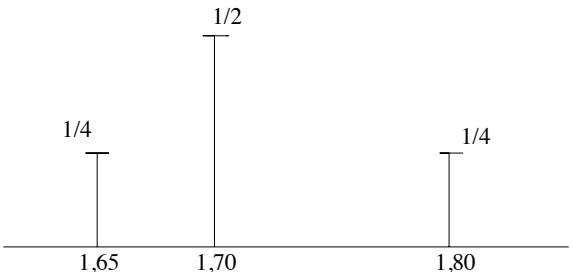
On peut aussi associer, à chaque échantillon d'effectif n, sa moyenne $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$. L'application qui, à chaque échantillon, associe sa moyenne est donc également une variable aléatoire, notée $\bar{X}$, qui possède une distribution échantillonnée. Il en est de

même pour les autres valeurs typiques de l'échantillon (variance S^2, moments M_k, ...).

Exemple

Soit une population de 4 individus A,B,C,D mesurant respectivement 1,65 m, 1,70 m, 1,70 m et 1,80 m.

Distribution de la population

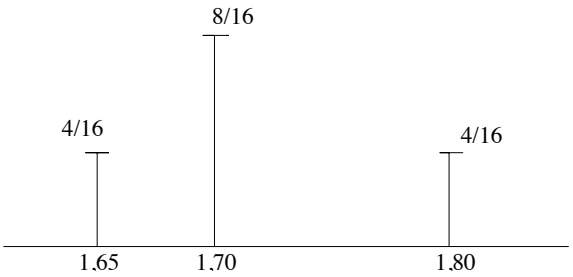


Liste de tous les échantillons d'effectif 2 pouvant être prélevés dans cette population

On effectue ici le prélèvement avec remplacement pour que l'échantillon soit simple; en pratique, les prélèvements se font sans remplacement mais les populations sont suffisamment grandes pour qu'on puisse supposer que l'échantillon est simple.

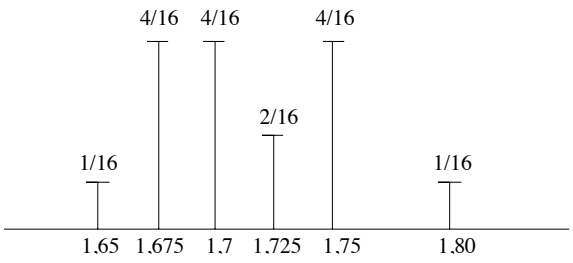
$$AA, AB, AC, AD, BA, BB, BC, BD, CA, CB, CC, CD, DA, DB, DC, DD.$$

Distribution échantillonnée de X_1 (c'est aussi celle de X_2)



Elle est identique à la distribution de la population.

Distribution échantillonnée de $\bar{X}$



1.2 Distribution échantillonnée de la moyenne $\bar{X}$

Supposons que la distribution de la population ait une moyenne μ et un écart-type σ. Il en est donc de même pour les distributions échantillonnées de X_1,X_2, . . . ,X_n et il résulte alors des propriétés de l'espérance mathématique que :

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \qquad \text{(justifier)} \\ &= \mu \end{aligned}$$

et il résulte des propriétés de la variance que

$$\begin{aligned} D^2(\bar{X}) &= D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) \qquad \text{(justifier)} \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

On voit donc que la variable $\bar{X}$ a une distribution centrée en μ dont la dispersion est d'autant plus faible que l'effectif de l'échantillon est grand.

Dans certains cas, on peut non seulement connaître les valeurs typiques de $\bar{X}$ mais aussi la forme de sa distribution.
Ainsi, *dans le cas particulier où les X_i sont des variables normales*, il résulte de la stabilité de la famille des variables normales que *$\bar{X}$ est aussi une normale* (de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$).

D'autre part, le théorème central-limite permet d'affirmer que pour n suffisamment grand, la variable $\bar{X}$ est approximativement normale, quelle que soit la distribution des X_i (en pratique, approximation satisfaisante à partir de $n = 30$), c'est-à-dire quelle que soit la distribution de la population. Remarquons encore que l'inégalité de BIENAYME-TCHEBYCHEFF permet d'écrire

$$P\left[|\bar{X} - \mu| \geq \frac{k\sigma}{\sqrt{n}}\right] \leq \frac{1}{k^2}$$

ou encore, en posant $\epsilon = \dfrac{k\sigma}{\sqrt{n}}$,

$$P\left[|\bar{X} - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2},$$

d'où

$$P\left[|\bar{X} - \mu| \geq \epsilon\right] \underset{n \to \infty}{\longrightarrow} 0.$$

On dit qu'il y a convergence en probabilité de $\bar{X}$ vers μ.

1.3 Distribution échantillonnée de la variance S^2

Rappelons que

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

On a donc

$$\begin{aligned} E(S^2) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - 2E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\bar{X}\right) + \frac{1}{n}\sum_{i} E(\bar{X}^2) \\ &= D^2(X_i) + \mu^2 - E(\bar{X}^2) \\ &= D^2(X_i) + \mu^2 - D^2(\bar{X}) - \mu^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

On voit qu'en moyenne la variance des échantillons est inférieure à la variance de la population, la différence étant cependant négligeable lorsque l'effectif de l'échantillon est grand.
On peut également démontrer que

$$D^2(S^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3},$$

μ_2 et μ_4 étant les moments centrés d'ordres 2 et 4 de la population. Si la population de référence est normale, on sait que $\mu_4 = 3\mu_2^2$ et on obtient

$$D^2(S^2) = \frac{2(n-1)\mu_2^2}{n^2} = 2\frac{n-1}{n^2}\sigma^4.$$

Dans certains cas, on peut connaître non seulement les valeurs typiques de S^2 mais aussi la forme de sa distribution.

Ainsi nous démontrons ci-dessous que *si la population est normale de moyenne μ et d'écart-type σ, alors la variable $\dfrac{nS^2}{\sigma^2}$ est une $\chi^2_{(n-1)}$.*

Démonstration

Soit les variables Y_i telles que

$$\begin{cases} Y_i = \dfrac{1}{\sqrt{i(i+1)}}(X_1 + X_2 + \ldots + X_i - iX_{i+1}), \quad i = 1, \ldots, n-1 \\ Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sqrt{n} \cdot \bar{X} \end{cases}$$

On démontrera, à titre d'exercice, que

$$\begin{cases} \displaystyle\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2 \\ E(Y_i) = 0, \quad \forall i = 1, \ldots, n-1, \\ D^2(Y_i) = \sigma^2, \quad \forall i \\ E(Y_i \cdot Y_j) = 0, \quad \forall i \neq j (\Rightarrow \text{ les } Y_i \text{ sont non-corrélées}) \\ Y_i \text{ est normale }, \forall i. \end{cases}$$

D'autre part, nous savons que des variables normales non-corrélées sont indépendantes (ce qui n'est pas le cas en général pour les autres variables). Il en résulte que les $\dfrac{Y_i}{\sigma}$ sont des $N(0,1)$ indépendantes. Or,

$$\begin{aligned} \frac{nS^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right] \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} Y_i^2 - Y_n^2 \right] = \sum_{i=1}^{n-1} \left(\frac{Y_i}{\sigma} \right)^2 . \end{aligned}$$

Par conséquent, $n\dfrac{S^2}{\sigma^2}$ est bien une $\chi^2_{(n-1)}$; de plus, cette variable est indépendante de Y_n et donc de $\bar{X}$.

Enfin, on peut démontrer que, pour n suffisamment grand, la variable S^2 est approximativement normale, quelle que soit la distribution des X_i (en pratique, approximation satisfaisante à partir de $n = 100$).

1.4 Distribution échantillonnée d'une fonction de $\bar{X}$ et S^2

Il résulte de ce qui précède que lorsque la population a une distribution normale, la variable $\sqrt{n-1}\dfrac{\bar{X}-\mu}{S}$ est une $t_{(n-1)}$ (justifier).

1.5 Distributions échantillonnées d'autres paramètres

On peut évidemment s'intéresser à la distribution échantillonnée de n'importe quelle variable aléatoire associée à l'échantillon. Nous serons amenés à en introduire dans la suite. La connaissance des distributions échantillonnées est en effet indispensable en inférence statistique : elles sont à la base de la construction des intervalles de confiance et de l'application des tests d'hypothèses (voir plus loin).

De façon générale, tous les moments centrés des échantillons et toutes les fonctions de ces moments sont des variables asymptotiquement normales. Ce résultat est très important : il signifie que pour des échantillons suffisamment grands, il suffit de connaître la moyenne et l'écart-type de la variable considérée pour avoir sa distribution échantillonnée complète et pouvoir ainsi appliquer les méthodes de l'inférence statistique.

Chapitre 2

Problèmes d'estimation

2.1 Introduction

On considère une population dont on voudrait estimer un paramètre h à partir d'un échantillon x_1,x_2, . . . ,x_n prélevé dans cette population. L'estimation que l'on va calculer sera donc une valeur dépendant de x_1,x_2, . . . ,x_n et variera avec l'échantillon choisi. L'application qui, à chaque échantillon, associe une estimation de h est à nouveau une variable aléatoire que nous appellerons *estimateur de* h, que nous noterons H, et qui possède une distribution échantillonnée.

Cet estimateur sera évidemment d'autant meilleur qu'il donnera le plus souvent des estimations proches de la valeur h. On demandera donc à l'estimateur de h d'avoir une distribution échantillonnée fortement concentrée autour de h. Cela signifie en particulier qu'on aura avantage à considérer un estimateur dont la distribution échantillonnée

1. a une espérance mathématique égale à h :

$$E(H) = h;$$

   on dit alors que l'estimateur est *sans biais*;

2. a une variance $D^2(H)$ la plus petite possible; de ce point de vue, on peut démontrer qu'à tout paramètre h de la population correspond un nombre $B(h)$ qui borne inférieurement $D^2(H)$; il n'y a donc pas moyen de trouver un estimateur H dont la variance soit strictement inférieure à $B(h)$: on dit alors que l'estimateur est *de variance minimum* (il n'existe pas toujours). Un estimateur sans biais et de variance minimum est appelé *estimateur efficace*. Un estimateur H sans biais

dont la variance tend vers $B(h)$ lorsque $n \to \infty$ est appelé *estimateur asymptotiquement efficace*.

La connaissance d'un estimateur efficace ne nous renseigne pas encore sur la précision de l'estimation qu'il fournit; bien que la variance de cet estimateur nous en donne une indication, il faudra faire appel à sa distribution échantillonnée pour pouvoir quantifier correctement cette précision. Ce sera l'objet du chapitre 3.

2.2 Estimation de la moyenne

La première variable qui se présente à l'esprit pour estimer la moyenne μ d'une population est évidemment la variable $\bar{X}$. Nous avons vu que c'était un estimateur sans biais ($E(\bar{X} = \mu)$ et que sa variance était d'autant plus petite que n était grand. En fait, on peut démontrer que $\bar{X}$ est un estimateur efficace de μ.

On pourrait également considérer, comme estimateur de μ, la variable $\widetilde{X}$ (variable qui, à chaque échantillon, associe sa médiane), mais il s'agit d'un estimateur de moins bonne qualité. Dans le cas d'une population normale, par exemple, il faut disposer d'environ 160 observations pour obtenir par l'intermédiaire de la médiane une estimation de qualité (dispersion) comparable à la moyenne de 100 observations.

2.3 Estimation de la variance

A nouveau, la première variable qui se présente à l'esprit, pour estimer la variance σ^2 d'une population, est la variable S^2. En réalité, ce n'est pas le meilleur estimateur. Nous avons vu en effet que ce n'était pas un estimateur sans biais $\left(E(S^2) = \dfrac{n-1}{n}\sigma^2\right)$: l'emploi de S^2 comme estimateur conduit, en moyenne, à une sous-estimation de σ^2. C'est la raison pour laquelle on utilise souvent, comme estimateur de σ^2, la variable

$$\frac{nS^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

qui est un estimateur sans biais de σ^2 (justifier); ce n'est pas un estimateur efficace mais on peut démontrer qu'il est asymptotiquement efficace.

2.4 Méthodes d'estimation

Nous avons vu quelles qualités on demandait à un estimateur et nous avons découvert, un peu au hasard, de bons estimateurs pour la moyenne et la variance d'une population. Le problème qui se pose maintenant est de savoir s'il existe des méthodes qui permettent d'obtenir de bons estimateurs des différents paramètres d'une population. La *méthode du maximum de vraisemblance* est une telle méthode. En effet, on démontre que, sous des conditions très générales, cette méthode fournit toujours des estimateurs de variance minimum ou asymptotiquement de variance minimum. De plus, leur distribution échantilllonnée est toujours asymptotiquement normale (la connaissance de la forme de la distribution échantillonnée d'un estimateur est fondamentale pour le chapitre 3). Par contre, les estimateurs fournis par cette méthode ne sont pas toujours sans biais. La méthode du maximum de vraisemblance est décrite dans le paragraphe suivant.

Il existe d'autres méthodes d'estimation qui peuvent être satisfaisantes dans certains cas : citons notamment *la méthode des moments* et *la méthode du χ^2-minimum*.

2.5 La méthode du maximum de vraisemblance

Le principe de la méthode est de choisir comme estimation du paramètre h, la valeur qui maximise la probabilité de provoquer l'apparition de l'échantillon observé.

Soit $x_1,x_2,\ldots,x_n$ un échantillon prélevé dans une population dont la distribution est discrète et dépend d'un paramètre inconnu h et soit $P(x_i;h)$ la probabilité que X_i soit égale à x_i $(i = 1,2,\ldots,n)$: cette probabilité dépend de h. L'échantillon étant simple et aléatoire, la probabilité d'obtenir l'échantillon $x_1,x_2,\ldots,x_n$ est donc égale à

$$P(x_1;h)\cdot P(x_2;h)\cdot \ \ldots\ \cdot P(x_n;h).$$

C'est cette fonction de h, appelée fonction de vraisemblance et notée $L(x_1,x_2,\ldots,x_n,h)$ que l'on maximise. La valeur $\widehat{h}$ de h qui maximise cette fonction est prise comme estimation de h et est appelée *estimation maximum de vraisemblance de h*.

En pratique, pour faciliter les calculs, on maximise plutôt $\log L$.

Dans le cas continu, la distribution de la population est caractérisée par une densité de probabilité $f(x;h)$ et la probabilité d'avoir $(x_1 \leq X_1 \leq x_1 + dx_1)$ et $(x_2 \leq X_2 \leq x_2 + dx_2)$ et ... et $(x_n \leq X_n \leq x_n + dx_n)$ est égale à

$$f(x_1;h) \cdot f(x_2;h) \cdot \ \ldots \ \cdot f(x_n;h)\, dx_1\, dx_2 \ldots dx_n.$$

La fonction

$$L(x_1,x_2,\ldots,x_n,h) = \prod_{i=1}^{n} f(x_i;h)$$

est la fonction de vraisemblance que l'on maximisera (ou son logarithme).

La méthode du maximum de vraisemblance s'étend sans difficultés à l'estimation simultanée de plusieurs paramètres. Dans ce cas, la fonction de vraisemblance dépend de ces différents paramètres : on la maximise en annulant ses dérivées partielles par rapport à ces paramètres et en résolvant le système obtenu (pour autant que L soit dérivable).

Exemple

Trouver les estimations maximum de vraisemblance de la moyenne et de l'écart-type d'une population $N(\mu,\sigma)$.

On a

$$f(x_i;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$L(x_1,x_2,\ldots,x_n;\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\log L = -n \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\begin{cases} \dfrac{\partial \log L}{\partial \mu} = 0 \\ \dfrac{\partial \log L}{\partial \sigma} = 0 \end{cases} \Rightarrow \begin{array}{l} \widehat{\mu} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \bar{x} \\ \widehat{\sigma} = \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu})^2} = s. \end{array}$$

Les estimations maximum de vraisemblance de la moyenne et de l'écart-type d'une population normale sont respectivement la moyenne et l'écart-type de l'échantillon.

Chapitre 3

Intervalles de confiance et tests d'hypothèses

3.1 La moyenne d'une population

3.1.1 Population normale dont l'écart-type σ est connu

1. **Intervalle de confiance**

   On a vu que si la population a une distribution normale de moyenne μ et d'écart-type σ, la variable $\bar{X}$ a une distribution normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Par conséquent, $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ a une distribution $N(0,1)$ et à tout nombre $\epsilon > 0$, on peut associer, grâce aux tables statistiques, un nombre $u_{\epsilon/2}$ tel que

$$P\left[-u_{\epsilon/2} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{\epsilon/2}\right] = 1-\epsilon,$$

   c'est-à-dire tel que

$$P\left[\bar{X} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\epsilon. \tag{1}$$

   Si ϵ est petit, l'intervalle aléatoire $\left[\bar{X} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$ contient donc la moyenne μ inconnue de la population avec une probabilité élevée.

   Pour chaque échantillon d'effectif n prélevé dans la population, on obtient un intervalle numérique $\left[\bar{x} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$, appelé *intervalle de confiance (ou de sécurité) de la moyenne*, qui fournit une estimation de la moyenne.

   L'égalité (1) nous assure que plus ϵ est petit, plus le nombre d'échantillons d'ef-

fectif n tels que $\left[\bar{x} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$ contient la valeur μ est grand.

A la limite, pour $\epsilon = 0$, on est sûr que l'intervalle de confiance contient la valeur μ, mais cette information est sans intérêt car pour $\epsilon = 0$, l'intervalle de confiance s'étend de $-\infty$ à $+\infty$. On se contente donc d'une valeur de ϵ positive mais petite (en général 5% ou moins) de façon à avoir un intervalle de longueur finie (et souvent très faible) qui ait de fortes chances de contenir μ.

Pour un ϵ fixé, la précision de l'estimation fournie par l'intervalle de confiance est d'autant plus grande que cet intervalle est peu étendu. Comme le montre la formule (1), une façon de réduire la longueur de cet intervalle consiste, ce qui est intuitivement évident, à augmenter n (cf. 2) ci-dessous). L'intervalle de confiance construit ci-dessus est symétrique par rapport à $\bar{x}$: on peut démontrer que, pour ϵ et n fixés, c'est l'intervalle de longueur minimum, mais il n'en sera pas toujours ainsi pour les autres paramètres.

2. **Détermination de** n

   Si l'on se fixe une erreur maximum d sur l'estimation de la moyenne, on aura, pour un ϵ fixé,

   $$u_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq d,$$

   et donc

   $$n \geq \frac{u_{\epsilon/2}^2\sigma^2}{d^2}.$$

   On obtient ainsi le nombre minimum d'éléments à prélever dans la population pour que l'échantillon ainsi constitué fournisse une estimation de μ avec une erreur qui ne dépasse d qu'avec une probabilité ϵ.

   On voit en particulier que pour diviser l'erreur maximum par 2, il faut quadrupler l'effectif de l'échantillon.

3. **Test d'hypothèse**

On désire vérifier si la moyenne de la population est égale à une valeur donnée μ_0. On sait que, si cette hypothèse est vraie, alors

$$P\left[-u_{\epsilon/2} \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq u_{\epsilon/2}\right] = 1 - \epsilon,$$

ou encore

$$P\left[\mu_0 - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \epsilon.$$

Par conséquent, si l'hypothèse était vraie, la plupart des échantillons d'effectif n devraient avoir leur moyenne $\bar{x}$ comprise dans l'intervalle $\left[\mu_0 \pm u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$ (en choisissant un ϵ petit). Cette conclusion conduit à la règle de décision suivante : ayant prélevé un échantillon de moyenne $\bar{x}$ dans la population, on rejettera l'hypothèse $\mu = \mu_0$ si $\bar{x}$ n'appartient pas à l'intervalle $\left[\mu_0 - \mu_{\epsilon/e}\frac{\sigma}{\sqrt{n}}, \mu_0 + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$ et on acceptera l'hypothèse dans le cas contraire.

L'intervalle auquel $\bar{x}$ doit appartenir pour accepter l'hypothèse s'appelle *l'intervalle d'acceptation*.

4. **Erreurs de première et de deuxième espèces - Puissance du test.**

   La règle de décision précédente peut conduire à 2 types d'erreurs. L'erreur de première espèce est celle qui consiste à rejeter l'hypothèse alors qu'elle est vraie. La probabilité de commettre cette erreur est égale à la probabilité que $\bar{X}$ n'appartienne pas à l'intervalle d'acceptation, c'est-à-dire à ϵ : on la maîtrise donc facilement.

   L'erreur de deuxième espèce est celle qui consiste à accepter l'hypothèse alors qu'elle est fausse. Cette erreur est plus difficile à calculer (et est souvent négligée dans les applications).

   Dans le cas considéré ici, supposons que la vraie valeur de μ soit $\mu_1 \neq \mu_0$. Dans ce cas, $\bar{X}$ a une distribution normale de moyenne μ_1 et d'écart-type $\frac{\sigma}{\sqrt{n}}$ et la probabilité de commettre l'erreur de seconde espèce est donnée par

$$\beta = P\left[\mu_0 - \mu_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq N\left(\mu_1, \frac{\sigma}{\sqrt{n}}\right) \leq \mu_0 + \mu_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$$

ou encore

$$\beta = P\left[\frac{\mu_0 - \mu_1}{\sigma} \cdot \sqrt{n} - u_{\epsilon/2} \leq N(0,1) \leq \frac{\mu_0 - \mu_1}{\sigma} \cdot \sqrt{n} + u_{\epsilon/2}\right].$$

On appelle *puissance du test* la quantité $1-\beta$: c'est donc la probabilité de rejeter l'hypothèse alors qu'elle est fausse : c'est une fonction de μ_1 qui est minimum (et vaut ϵ) pour $\mu_1 = \mu_0$ et qui augmente avec l'écart $|\mu_0 - \mu_1|$. Lorsqu'il existe plusieurs tests possibles pour une hypothèse, on fixe le risque de première espèce (ϵ) et on sélectionne le meilleur test sur base de l'étude de la puissance : c'est l'objet de la théorie des tests statistiques, qui ne sera pas développée ici.

5. **Les tests unilatéraux**

   On désire vérifier si la moyenne de la population est supérieure ou égale à une valeur donnée μ_0.

   On sait que si $\mu = \mu_0$, alors

$$P\left[\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq -u_\epsilon\right] = 1 - \epsilon,$$

   ou encore

$$P\left[\bar{X} \geq \mu_0 - \mu_\epsilon\frac{\sigma}{\sqrt{n}}\right] = 1 - \epsilon.$$

   Par conséquent, si $\mu = \mu_0$, la plupart des échantillons d'effectif n devraient avoir la moyenne $\bar{x}$ supérieure ou égale à $\mu_0 - \mu_\epsilon\frac{\sigma}{\sqrt{n}}$, et a fortiori si $\mu \geq \mu_0$.

   Cette conclusion conduit à la règle de décision suivante : ayant prélevé un échantillon de moyenne $\bar{x}$ dans la population, on rejettera l'hypothèse $\mu \geq \mu_0$ si $\bar{x}$ est inférieur à $\mu_0 - \mu_\epsilon\frac{\sigma}{\sqrt{n}}$ et on acceptera l'hypothèse dans le cas contraire.

   Le raisonnement est analogue pour vérifier si $\mu \leq \mu_0$.

3.1.2 Population normale dont l'écart-type est inconnu

1. **Intervalle de confiance**

   Le raisonnement précédent ne s'applique plus ici puisque σ est inconnu, mais on se base sur le fait que lorsque la distribution de la population est normale, la variable

   $$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n-1}}}$$

   est une variable de Student à $n-1$ degrés de liberté.

   On peut donc associer à tout $\epsilon > 0$ un nombre $t_{\epsilon/2}$, lu dans les tables, tel que

   $$P\left[-t_{\epsilon/2} \leq \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n-1}}} \leq t_{\epsilon/2}\right] = 1 - \epsilon.$$

   Pour chaque échantillon d'effectif n prélevé dans la population, on obtient ainsi, pour la moyenne μ, l'intervalle de confiance

   $$\left[\bar{x} - t_{\epsilon/2}\frac{s}{\sqrt{n-1}}, \bar{x} + t_{\epsilon/2}\frac{s}{\sqrt{n-1}}\right].$$

2. **Test d'hypothèse**

   Si l'hypothèse $\mu = \mu_0$ est vraie on a

   $$P\left[-t_{\epsilon/2} \leq \sqrt{n-1}\frac{\bar{X} - \mu_0}{S} \leq t_{\epsilon/2}\right] = 1 - \epsilon.$$

   La règle de décision sera donc la suivante : ayant prélevé un échantillon de moyenne $\bar{x}$ et d'écart-type s dans la population, on rejettera l'hypothèse $\mu = \mu_0$ si

   $$|\sqrt{n-1}\frac{\bar{x} - \mu_0}{s}| > t_{\epsilon/2}$$

   et on l'acceptera dans le cas contraire.

   Les remarques concernant les erreurs de première et de deuxième espèces, la puissance du test et les tests unilatéraux s'appliquent encore dans ce cas-ci.

3.1.3 Population quelconque

Lorsqu'on ne connaît pas la distribution échantillonnée exacte de $\bar{X}$, on se base sur des résultats asymptotiques. On sait en effet que si n est suffisamment grand, la variable $\bar{X}$ a une distribution aproximativement normale. En se basant sur des raisonnements analogues aux précédents, on obtient donc les intervalles de confiance approchés

$$\left[\bar{x} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right] \text{ lorsque } \sigma \text{ est connu}$$

et

$$\left[\bar{x} - u_{\epsilon/2}\frac{s}{\sqrt{n}}, \bar{x} + u_{\epsilon/2}\frac{s}{\sqrt{n}}\right] \text{ lorsque } \sigma \text{ est inconnu}$$

(puisque pour n grand, s constitue une bonne approximation de σ).

Les tests d'hypothèses se réalisent comme précédemment.

3.2 La comparaison des moyennes de deux populations

3.2.1 Population normales d'écarts-types σ_1 et σ_2 connus

Si μ_1 et μ_2 sont les moyennes des 2 populations et si les échantillons sont prélevés indépendamment dans les 2 populations, alors

$$\bar{X}_1 - \bar{X}_2 = N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right), \qquad \text{(justifier)}$$

où n_1 et n_2 sont les effectifs des échantillons.

A tout $\epsilon > 0$, on peut donc associer un nombre $u_{\epsilon/2}$ tel que

$$P\left[-u_{\epsilon/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq u_{\epsilon/2}\right] = 1 - \epsilon,$$

d'où on peut déduire un intervalle de confiance pour $\mu_1 - \mu_2$.

D'autre part, si l'hypothèse $\mu_1 = \mu_2$ est vraie, on aura

$$P\left[-u_{\epsilon/2} \leq \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq u_{\epsilon/2}\right] = 1 - \epsilon,$$

d'où la règle de décision suivante : ayant prélevé un échantillon dans chaque population, on rejettera l'hypothèse $\mu_1 = \mu_2$ si

$$\left|\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| > u_{\epsilon/2}$$

et on l'acceptera dans le cas contraire.

3.2.2 Populations normales de même écart-type σ inconnu

Dans ce cas,

$$\bar{X}_1 - \bar{X}_2 = N\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{n_1 + n_2}{n_1 n_2}}\right) \qquad \text{(justifier)}$$

et

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} = \chi^2_{n_1+n_2-2} \qquad \text{(justifier)}$$

Ces deux variables étant indépendantes, on en déduit que

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2}}}$$

est une variable de Student à $n_1 + n_2 - 2$ degrés de liberté. A tout $\epsilon > 0$, on peut donc associer un nombre $t_{\epsilon/2}$ tel que

$$P\left[-t_{\epsilon/2} \leq T \leq t_{\epsilon/2}\right] = 1 - \epsilon,$$

d'où on peut déduire un intervalle de confiance pour $\mu_1 - \mu_2$. D'autre part, on rejettera l'hypothèse $\mu_1 = \mu_2$ si

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}} > t_{\epsilon/2}.$$

(Ce test est généralement appelé test de Student).

Lorsque les 2 populations n'ont pas le même écart-type, le problème est plus compliqué et il faut utiliser une autre variable pour effectuer le test : nous ne considérons pas ce cas ici.

3.2.3 Populations quelconques

Si n_1 et n_2 sont suffisamment grands (au moins 20), alors $\bar{X}_1 - \bar{X}_2$ a une distribution approximativement normale de moyenne $\mu_1 - \mu_2$ et d'écart-type $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, où on remplace σ_1 et σ_2 par s_1 et s_2 lorsqu'ils sont inconnus.

On rejette alors l'hypothèse $\mu_1 = \mu_2$ lorsque

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > u_{\epsilon/2}.$$

Comme dans le cas de la moyenne, on peut effectuer des tests unilatéraux, s'intéresser à la puissance des tests ou encore déterminer les effectifs des échantillons de manière à atteindre une précision donnée.

3.3 La variance d'une population

Si la population est normale de moyenne μ et d'écart-type σ, nous savons que

$$\frac{nS^2}{\sigma^2} = \chi^2_{n-1}.$$

A tout ϵ, on peut donc associer, grâce aux tables statistiques, deux nombres $k_{\epsilon/2}$ et $k_{1-\epsilon/2}$ tels que

$$\mathbb{P}\left[\frac{nS^2}{\sigma^2} \geq k_{\epsilon/2}\right] = \mathbb{P}\left[\frac{nS^2}{\sigma^2} \leq k_{1-\epsilon/2}\right] = \epsilon/2,$$

i.e. tels que

$$\mathbb{P}\left[k_{1-\epsilon/2} \leq \frac{nS^2}{\sigma^2} \leq k_{\epsilon/2}\right] = 1 - \epsilon,$$

d'où on déduit, pour σ^2, l'intervalle de confiance

$$\left[\frac{ns^2}{k_{\epsilon/2}}, \frac{ns^2}{k_{1-\epsilon/2}}\right]$$

et pour σ, l'intervalle de confiance

$$\left[\sqrt{\frac{ns^2}{k_{\epsilon/2}}}, \sqrt{\frac{ns^2}{k_{1-\epsilon/2}}}\right].$$

D'autre part, on rejette l'hypothèse $\sigma = \sigma_0$ si s^2 n'est pas dans l'intervalle

$$\left[\frac{\sigma_0^2 k_{1-\epsilon/2}}{n}, \frac{\sigma_0^2 k_{\epsilon/2}}{n}\right].$$

Si n est grand $(n > 30)$, on se base sur le résultat asymptotique selon lequel

$$\sqrt{2\chi^2_{n-1}} - \sqrt{2n-3} \approx N(0,1).$$

Si la population est quelconque, on se base sur les résultats suivants :

$$S^2 \approx N\left(\sigma^2, \sqrt{\frac{m_4 - \sigma^4}{n}}\right)$$

$$S \approx N\left(\sigma, \frac{1}{2s}\sqrt{\frac{m_4 - \sigma^4}{n}}\right)$$

3.4 La comparaison des variances de deux populations

Si les populations sont normales, alors la variable

$$\frac{\frac{n_1 S_1^2}{(n_1-1)\sigma_1^2}}{\frac{n_2 S_2^2}{(n_2-1)\sigma_2^2}}$$

a une distribution de Snedecor à $(n_1 - 1, n_2 - 1)$ degrés de liberté, ce qui permet de construire un intervalle de confiance pour le rapport $\frac{\sigma_2^2}{\sigma_1^2}$ et de tester l'hypothèse $\sigma_1^2 = \sigma_2^2$ (en testant l'hypothèse $\frac{\sigma_2^2}{\sigma_1^2} = 1$).

Si les populations sont quelconques, on se base sur le fait que $S_1^2 - S_2^2$ est approximativement normale (de quelle moyenne et de quel écart-type ?).

Exemples numériques

– On a prélevé un échantillon d'effectif 16 et de moyenne 0,828 dans une population normale d'écart-type 0,003; construire un intervalle de confiance pour la moyenne de cette population, au coefficient de sécurité 95%.

Réponse :

$$\left[\bar{x} - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $\bar{x} = 0{,}828$; $u_{\epsilon/2} = 1{,}96$; $\sigma = 0{,}003$ et $n = 16$.

– Quel est l'effectif minimum de l'échantillon à prélever pour que l'erreur sur l'estimation de la moyenne ne dépasse pas $0{,}001$, au coefficient de sécurité 95% ?

Réponse : il faut $u_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq 0{,}001$ où $\sigma = 0{,}003$ et $u_{\epsilon/2} = 1{,}96$.

– Tester, au niveau 5% l'hypothèse $\mu = 0{,}83$, où μ est la moyenne de la population.

Réponse : on accepte l'hypothèse si

$$\mu - u_{\epsilon/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + u_{\epsilon/2}\frac{\sigma}{\sqrt{n}},$$

où $\mu = 0{,}83$; $u_{\epsilon/2} = 1{,}96$; $\sigma = 0{,}003$; $n = 16$ et $\bar{x} = 0{,}828$

– On a prélevé un échantillon d'effectif 20, de moyenne 0,7 et de variance 0,01 dans une population normale; construire des intervalles de confiance pour la moyenne et la variance de cette population, au coefficient de sécurité 99%.

Réponses : pour la moyenne : $\left[0{,}7 - 2{,}861 \cdot \frac{0{,}1}{\sqrt{19}}; 0{,}7 + 2{,}861 \cdot \frac{0{,}1}{\sqrt{19}}\right]$

pour la variance : $\left[\frac{20 \times 0{,}01}{38{,}58}, \frac{20 \times 0{,}01}{6{,}84}\right]$.

3.5 Coefficients de corrélation et de régression

On a les résultats asymptotiques suivants :

a) $\frac{1}{2} \ln \frac{1+R}{1-R} \approx N\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{\sqrt{n-3}}\right)$,

où R est la variable aléatoire qui à tout échantillon bidimensionnel d'effectif n associe la valeur du coefficent de corrélation et ρ est le coefficient de corrélation de la population.

b) si $\rho = 0$, alors $\frac{\sqrt{n-2} \cdot R}{\sqrt{1-R^2}}$ se comporte asymptotiquement comme une variable de Student à $n-2$ degrés de liberté.

c) $\frac{S_1\sqrt{n-2}}{S_2\sqrt{1-R^2}}(B_{12} - \beta_{12})$ se comporte asymptotiquement comme une variable de Student à $n-2$ degrés de liberté, où B_{12} est la variable aléatoire qui à tout échantillon bidimensionnel d'effectif n associe la valeur du coefficient de régression de la 1ère variable par rapport à la 2ème.

Les principes de construction d'intervalles de confiance et de réalisation de tests sont les mêmes que précédemment (voir illustrations aux exercices).

3.6 Les pourcentages

Soit une population d'individus dels que chacun possède une certaine propriété A avec une probabilité inconnue notée p. On désire obtenir un intervalle de confiance pour p (exemple d'application : les sondages).

A chaque échantillon d'effectif n prélevé dans cette population, on peut associer le nombre d'individus de l'échantillon qui ont la propriété A. On obtient ainsi une variable aléatoire F qui, si la population est infinie, a une distribution binomiale de paramètres n et p. (Si la population est finie, F a une distribution hypergéométrique mais dès que l'effectif de la population vaut au moins 10 fois celui de l'échantillon, elle est pratiquement identique à la binomiale).

Le pourcentage $\frac{F}{n}$ d'individus ayant la propriété A est donc une variable aléatoire telle que

$$\begin{cases} P\left[\frac{F}{n} = \frac{k}{n}\right] = \binom{n}{k} p^k (1-p)^{n-k} \\ E\left[\frac{F}{n}\right] = p \\ D^2\left[\frac{F}{n}\right] = \frac{p(1-p)}{n} \end{cases} \qquad \text{(justifier)}$$

Pour de petits échantillons, on dispose de tables basées sur ces relations et permettant de traiter de manière exacte les problèmes d'intervalles de confiance et des tests d'hypothèse (voir exercices).

Pour de grands échantillons, on peut approcher la distribution de $\frac{F/n - p}{\sqrt{\frac{pq}{n}}}$ par une $N(0,1)$ et on obtient alors, pour p, l'intervalle de confiance suivant :

$$(1) \qquad \frac{n}{n + u_{\epsilon/2}^2} \left[\frac{F}{n} + \frac{u_{\epsilon/2}^2}{2n} \pm u_{\epsilon/2} \sqrt{\frac{F(n-F)}{n^3} + \frac{u_{\epsilon/2}^2}{4n^2}} \right].$$

Pour n grand, on se contentera de l'intervalle approché

$$\left[\frac{F}{n} \pm u_{\epsilon/2} \sqrt{\frac{F(n-F)}{n^3}} \right].$$

On peut également appliquer le résultat asymptotique suivant : la variable aléatoire 2 arc sin $\sqrt{\frac{F}{n}}$ se comporte asymptotiquement comme une variable normale de moyenne 2 arc sin $\sqrt{p}$ et d'écart-type $\frac{1}{\sqrt{n}}$.

Rappel : $u_{\epsilon/2}$ est la valeur, lue dans les tables, telle que

$$P\left[N(0,1) > u_{\epsilon/2}\right] = \epsilon/2.$$

3.7 La comparaison de deux pourcentages

Soient 2 populations d'individus dans lesquelles les probabilités qu'un individu possède une certaine propriété A sont respectivement p_1 et p_2. Si n_1 et n_2 sont les effectifs des échantillons prélevés dans ces 2 populations et si F_1 et F_2 sont les variables aléatoires qui représentent les nombres d'individus possédant la propriété A dans chacun de ces échantillons, on sait que
$\frac{F_1}{n_1} - \frac{F_2}{n_2}$ se comporte comme une normale (justifier). On en déduit la règle suivante pour tester l'égalité de p_1 et p_2 : on rejette cette hypothèse si la valeur 0 n'appartient pas à l'intervalle

$$\frac{F_1}{n_1} - \frac{F_2}{n_2} \pm u_{\epsilon/2}\sqrt{\frac{F_1+F_2}{n_1+n_2}\left(1-\frac{F_1+F_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

et on l'accepte dans le cas contraire.

Chapitre 4

Tests d'ajustement et de normalité

Les tests d'ajustement consistent à vérifier si une population donnée possède ou non une distribution théorique fixée, à partir d'échantillons prélevés dans cette population. Ces tests permettent en particulier de vérifier si une population donnée possède ou non une distribution normale, mais, dans ce cas particulier, on dispose aussi d'autres méthodes appelées tests de normalité.

4.1 Test d'ajustement dans le cas d'une distribution théorique complètement spécifiée

Le test consiste à partitionner la population en sous-ensembles S_i et à associer à chaque S_i les 2 nombres suivants :

p_i = probabilité qu'un individu prélevé au hasard dans la population appartienne au sous-ensemble S_i, calculée sur base de la distribution théorique proposée;

n_i = nombre d'individus appartenant à S_i parmi un échantillon d'effectif n prélevé dans la population.

Si la population a effectivement la distribution théorique proposée, on peut s'attendre à ce que les quantités p_i et $\frac{n_i}{n}$ soient proches les unes des autres. Le test est donc basé sur une mesure de l'écart entre ces quantités. Cette mesure est définie comme suit :

$$\delta = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

où r est le nombre des sous-ensembles S_i; elle vaut 0 ssi $\frac{n_i}{n} = p_i$, $\forall i$ et elle est d'autant plus grande que les nombres $\frac{n_i}{n}$ sont éloignés des p_i. D'autre part, on peut démonter que si la population a effectivement la distribution théorique proposée, alors la variable aléatoire Δ (qui associe la valeur δ à chaque échantillon) a un comportement asymptotiquement égal à celui d'une $\chi^2_{(r-1)}$, d'où le test suivant, connu sous le nom de *test χ^2 d'ajustement* : ϵ étant fixé, on rejette l'hypothèse selon laquelle la population a la distribution proposée si $\delta > k_\epsilon$, où k_ϵ est la valeur lue dans les tables telles que

$$P\left[\chi^2_{(r-1)} > k_\epsilon\right] = \epsilon,$$

et on accepte l'hypothèse dans le cas contraire.

En pratique, les sous ensembles S_i seront choisis de telle sorte que $np_i > 5$ (ce qui peut toujours s'obtenir en regroupant des sous-ensembles ne vérifiant pas cette condition); d'autre part, le calcul de δ est plus simple lorsqu'on utilise la formule suivante :

$$\delta = \sum_{i=1}^{r} \frac{n_i^2}{np_i} - n$$

(démontrer qu'elle est équivalente à la formule précédente).

4.2 Test d'ajustement dans le cas d'une distribution théorique non complètement spécifiée

Lorsque les paramètres de la distribution théorique ne sont pas tous fixés, on commence par les estimer par la méthode du maximum de vraisemblance. On associe ensuite à chaque S_i les 2 nombres suivants :

$\widehat{p}_i$ = probabilité qu'un individu prélevé au hasard dans la population appartienne au sous-ensemble S_i, calculée sur base de la distribution théorique proposée après estimation des paramètres non fixés;

n_i = nombre d'individus appartenant à S_i parmi un échantillon d'effectif n prélevé dans la population.

On peut démontrer que la variable $\widehat{\Delta}$ qui associe à chaque échantillon la valeur

$$\widehat{\delta} = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i}$$

se comporte asymptotiquement comme une χ^2_{r-s-1}, où s est le nombre de paramètres qui ont dû être estimés.

Le test s'effectue comme dans le cas précédent.

4.3 Exemple

On a observé le nombre d'individus qui ont subi $0,1,2,\ldots$ accidents de la route au cours d'une période déterminée.

Tester, au niveau 5%, que le nombre d'accidents suit une loi de Poisson.

Nombre d'accidents	0	1	2	3	5
Nombre d'individus n_i	102	59	31	8	1

L'estimation maximum de vraisemblance du paramètre λ de la loi de Poisson est la moyenne de l'échantillon (cf. exercices sur la méthode du max. de vraisemblance)

$$\Rightarrow \lambda = \bar{x} = 0{,}746.$$

Si le nombre d'accidents suivait une loi de Poisson de paramètre $0{,}746$, on aurait la distribution suivante :

Nombre d'accidents	0	1	2	3	> 3
Probabilité p_i	0,475	0,353	0,132	0,033	0,007

Pour respecter la condition $np_i \geq 5$, on regroupe les 2 dernières catégories et on obtient

$$\delta = \sum_{i=1}^{4} \frac{n_i^2}{np_i} - n = 3{,}42,$$

à comparer à la valeur $5{,}99$ lue dans la table de la χ^2 à 2 degrés de liberté $\Rightarrow$ on accepte l'hypothèse.

4.4 Tests de normalité

On peut tester la normalité d'une population à l'aide du test χ^2 d'ajustement, mais il existe également d'autres méthodes.

En particulier, si la population est normale, on sait que

$$\gamma_1 = \gamma_2 = 0.$$

On doit donc s'attendre à ce que les paramètres g_1 et g_2, calculés à partir d'un échantillon prélevé dans cette population, aient des valeurs proches de 0. Le problème est donc ramené aux tests $\gamma_1 = 0$ et $\gamma_2 = 0$, pour lesquels il existe des tables statistiques particulières.

Signalons encore l'existence du "papier de probabilité" ou "papier probit" : il s'agit d'un papier gradué dont l'échelle des ordonnées est telle que les fonctions de répartition des variables normales se représentent par des droites. La représentation sur ce papier de la fonction de distribution d'un échantillon permet donc de se faire une idée de la normalité de la population dont il est issu.

Chapitre 5

Tests d'homogénéité

Les tests d'homogénéité consistent à vérifier si k populations données possèdent ou non la même distribution théorique, à partir d'échantillons prélevés dans ces populations.

5.1 Test d'homogénéité dans le cas d'une distribution théorique complètement spécifiée

On partitionne le domaine de la variable étudiée en sous-ensembles S_i $(i = 1,2,\ldots,r)$ auxquels on associe les nombres suivants :

p_i = probabilité qu'un individu prélevé au hasard dans la population appartienne au sous-ensemble S_i, calculée sur base de la distribution théorique proposée;

$n_{i\ell}$ = nombre d'individus appartenant à S_i parmi un échantillon d'effectif n_ℓ prélevé dans la population ℓ.

Le test, analogue au test d'ajustement, se base sur la quantité

$$\delta = \sum_{\ell=1}^{k} \sum_{i=1}^{r} \frac{(n_{i\ell} - n_\ell p_i)^2}{n_\ell p_i},$$

qui est d'autant plus proche de 0 que l'hypothèse est vraie. La variable Δ se comporte asymptotiquement comme un $\chi^2_{k(r-l)}$.

5.2 Distribution théorique non complètement spécifiée

cf. 4.2.

5.3 Test d'homogénéité dans le cas d'une distribution théorique non spécifiée

Le test se fait comme précédemment, après estimation des nombres p_i par $\frac{1}{n}\sum_{\ell=1}^{k} n_{i\ell}$.

La variable Δ se comporte alors asymptotiquement comme une $\chi^2_{(k-1)(r-1)}$.

Chapitre 6

Test d'indépendance

Considérons une population dont les individus peuvent être classés suivant deux critères : on se propose de tester si ces 2 critères peuvent être considérés comme indépendants.

Soient $S_i (i = 1,2,\ldots,q)$ et $S'_j (j = 1,2,\ldots,r)$ les classes correspondant aux 2 critères, p_{ij} la probabilité qu'un individu appartienne aux classes S_i et S'_j, $p_{i\cdot}$ et $p_{\cdot j}$ étant les probabilités marginales.

En cas d'indépendance, on sait que

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j} :$$

c'est l'hypothèse à tester.

Soit un échantillon d'effectif n et appelons n_{ij} le nombre d'individus appartenant à $S_i \cap S'_j$, $n_{i\cdot}$ le nombre d'individus appartenant à S_i et $n_{\cdot j}$ le nombre d'individus appartenant à S'_j.

Si les deux critères de classification sont indépendants, on peut s'attendre à ce que

$$\frac{n_{ij}}{n} \approx \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}.$$

Le test est donc basé sur une mesure de l'écart entre ces quantités. Cette mesure est définie comme suit :

$$\delta = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}.$$

Elle a un comportement asymptotique égal à celui d'une $\chi^2_{(q-1)(r-1)}$.

Remarque : il faut être prudent au niveau de l'interprétation : le rejet de l'hypothèse d'indépendance ne signifie pas nécessairement qu'il existe une relation de cause à effet entre les deux caractéristiques considérées.

Pour le calcul, on utilisera la formule équivalente :

$$\delta = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{n\, n_{ij}^2}{n_{i.}n_{.j}} - n$$

Chapitre 7

Les tests non paramétriques

On qualifie de non paramétriques les tests qui utilisent des variables aléatoires dont les distributions sont indépendantes des distributions des populations, tels que, par exemple, les tests d'ajustement, d'homogénéité et d'indépendance. Ce sont en général des tests relativement simples parce qu'ils font intervenir soit des variables indicatrices (indiquant l'appartenance ou non à des classes), soit des rangs (numéros d'ordre des valeurs observées rangées par ordre croissant). A titre d'exemple, nous présentons deux tests non paramétriques permettant de comparer les distributions de 2 populations quant à leur position, comme le fait le test de comparaison de 2 moyennes (qui, lui, est un test paramétrique).

7.1 Le test des médianes

Ayant prélevé un échantillon dans chaque population, on calcule la médiane de l'ensemble des valeurs contenues dans les deux échantillons ainsi que, pour chaque échantillon, la proposition d'observations inférieures à cette médiane. Si les distributions des deux populations sont analogues quand à leur position, on peut s'attendre à ce que les propositions obtenues soient proches l'une de l'autre. On effectue donc le test en comparant ces proportions (cf. avant : comparaison de 2 pourcentages).

7.2 Le test des rangs (WILCOXON et WHITE)

Ayant prélevé un échantillon dans chaque population, on range l'ensemble des valeurs contenues dans les deux échantillons par ordre croissant et on calcule les sommes des rangs relatives aux deux échantillons, soit z_1 et z_2. On sait que

$$z_1 + z_2 = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1),$$

où n_1 et n_2 sont les effectifs des 2 échantillons.

D'autre part, si les distributions des 2 populations sont analogues quant à leur position, on peut s'attendre à ce que z_1 et z_2 soient proportionnelles à n_1 et n_2, c'est-à-dire à ce que

$$z_1 \approx \frac{n_1}{n_1 + n_2}(z_1 + z_2) \approx \frac{1}{2}n_1(n_1 + n_2 + 1)$$

et

$$z_2 \approx \frac{n_2}{n_1 + n_2}(z_1 + z_2) \approx \frac{1}{2}n_2(n_1 + n_2 + 1).$$

Le principe du test consiste donc à rejeter l'hypothèse d'identité des 2 distributions lorsque la valeur observée z_1 s'écarte trop de la valeur attendue. Pour des effectifs suffisamment élevés $(n_1 + n_2 > 30)$, on se base sur le fait que $Z_1 \sim N(\frac{1}{2}n_1(n_1 + n_2 + 1), \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}})$.

Pour des effectifs plus petits, on dispose de tables spéciales. Le test des rangs a été étendu au cas de plus de deux populations (test des rangs de KRUSKAL et WALLIS).

Chapitre 8

Introduction à l'analyse de variance

L'analyse de variance à un facteur a pour but de comparer les moyennes de plusieurs populations *supposées normales et de même variance*, à partir d'échantillons aléatoires, simples et indépendants les uns des autres. On peut la considérer comme une généralisation du test de Student (cf. comparaison de 2 moyennes).

Soit $n_i (i = 1,2,\ldots,p)$ l'effectif de l'échantillon extrait de la $i^{\text{ème}}$ population et x_{ik} la $k^{\text{ème}}$ observation de cet échantillon.

En posant

$$\begin{aligned} n &= \sum_{i=1}^{p} n_i, \\ \bar{x}_i &= \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik}, \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik} = \frac{1}{n} \sum_{i=1}^{p} n_i \bar{x}_i, \end{aligned}$$

on obtient (justifier) :

$$\sum_{i=1}^{p} \sum_{k=1}^{n_i} (x_{ik} - \bar{x})^2 = \sum_{i=1}^{p} n_i (\bar{x}_i - x)^2 + \sum_{i=1}^{p} \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_i)^2.$$

Cette relation montre que la somme des carrés des écarts par rapport à la moyenne générale (appelée somme des carrés des écarts totale et notée $S\,C\,E_t$) peut être décomposée

en la somme des carrés des écarts entre échantillons (appelée somme des carrés des écarts factorielle et notée $S\,C\,E_f$) plus la somme des carrés des écarts dans les échantillons (appelée somme des carrés des écarts résiduelle et notée $S\,C\,E_r$).

Lorsqu'il existe des différences importantes entre les moyennes des populations. on doit s'attendre à ce que SCE_f soit beaucoup plus importante que SCE_r : c'est le rapport entre ces 2 quantités qui servira à faire le test (le rejet de l'hypothèse d'égalité des moyennes se fera pour les valeurs élevées de ce rapport). Pour préciser la technique du test, il faut donc connaître la distribution échantillonnée du rapport $\dfrac{SCE_f}{SCE_r}$ ou d'une fonction de ce rapport.

Les populations étant supposées *normales* et de *même variance* σ^2 (ce sont les deux hypothèses de base de l'analyse de variance), on en déduit que si les populations ont même moyenne, alors $\dfrac{SCE_t}{\sigma^2}$ est une valeur observée d'une $\chi^2_{(n-1)}$.

D'autre part, que les populations aient même moyenne ou non, $\dfrac{1}{\sigma^2}\displaystyle\sum_{k=1}^{n_i}(x_{ik}-\bar{x}_i)^2$ est une valeur observée d'une $\chi^2_{(n_i-1)}$, de sorte que $\dfrac{CSE_r}{\sigma^2}$ est une valeur observée d'une $\chi^2_{(n-p)}$.

Enfin, on peut démontrer que SCE_f et SCE_r sont indépendantes, d'où il résulte que $\dfrac{SCE_f}{\sigma^2}$ est une valeur observée d'une $\chi^2_{(p-1)}$, lorsque les populations ont même moyenne.

Finalement, si les populations ont même moyenne, alors $\dfrac{\dfrac{1}{p-1}SCE_f}{\dfrac{1}{n-p}SCE_r}$ est une valeur observée d'une variable de Snedecor $F_{p-1,n-p}$.

On rejettera donc l'hypothèse d'égalité des moyennes des populations lorsque la valeur observée est supérieure à la valeur F_ϵ, abscisse laissant à sa droite une masse de probabilité égale à ϵ dans une distribution de Snedecor à $p-1$ et $n-p$ degrés de liberté.

L'analyse de variance est particulièrement utilisée dans les disciplines où l'on étudie les effets de différents traitements sur les individus d'une population (agronomie, médecine,...).

L'analyse de variance à plusieurs facteurs concerne le cas où les populations étudiées peuvent être classées suivant plusieurs facteurs (**exemple :** étude de l'effet de différents traitements sur les individus d'une population classés suivant le sexe et la classe d'âges).

Table des matières