M1 IAD
Notes du cours RFIDEC (1)

Jean-Yves Jaffray

11 octobre 2006

1 Introduction

1.1 Statistique descriptive et statistique inférentielle

La compréhension de la plupart des phénomènes, qu'ils soient naturels, techniques, économiques ou sociaux, passe par leur observation approfondie, souvent complétée, lorsque c'est possible, par une expérimentation. Les données recueillies le sont généralement sous une forme brute, encombrante et peu parlante ; on leur fait donc subir diverses transformations afin d'en obtenir une représentation plus compacte et d'en dégager les caractéristiques sur lesquelles s'appuiera l'analyse du phénomène étudié. L'ensemble de ces méthodes de traitement des données constitue la *statistique descriptive*. On souhaite parfois aller plus loin et pouvoir séparer ce qui, dans les observations, est fortuit de ce qui y est fondamental ; on cherche souvent aussi à prévoir le déroulement de phénomènes futurs, dépendant éventuellement de certaines actions; ceci ne devient possible que dans le cadre d'un modèle probabiliste, nécessitant des hypothèses, parfois assez restrictives, sur les fluctuations possibles du phénomène étudié. C'est le domaine de la *statistique inférentielle*.

1.2 Statistique descriptive

La statistique descriptive cherche à synthétiser et résumer de la façon la plus efficace possible l'information pertinente contenue dans les données. Elle utilise abondamment des tableaux de nombres, mais aussi beaucoup de représentations visuelles : graphiques, courbes, etc.. Bien qu'elle fasse usage de caractéristiques (distributions, moyennes, dispersions, etc.) qui ont leurs analogues en calcul des probabilités, elle ne fait pas réellement appel au modèle probabiliste.
Les données disponibles concernant la *population* (l'ensemble des individus) étudiée la décrivent généralement à l'aide d'un certain nombre de variables. Face à une population de grande taille, on peut réduire la dimension des

données en regroupant les individus en classes plus ou moins homogènes : c'est le but des *méthodes de classification* ; un nombre excessif de variables peut être remplacé par un ensemble plus réduit de variables de synthèse : c'est ce que réalisent les *méthodes factorielles*.
La statistique descriptive se borne à mettre en évidence de l'information contenue implicitement dans les données et se refuse à toute extrapolation ; ses révélations peuvent éventuellement fournir des hypothèses concernant le phénomène observé et suggérer des corrélations ou même des relations causales entre les variables ; l'interprétation des représentations obtenues et la formulation de conclusions restent toutefois en dehors de l'étude des données proprement dite.

1.3 Statistique inférentielle

Plus ambitieuse, la statistique inférentielle a pour objectif d'arriver à des conclusions concernant un ensemble d'individus à partir des informations recueillies sur une partie d'entre eux. On distingue donc : la *population*, ensemble de tous les individus auxquels s'intéresse l'étude et l'*échantillon*, partie de cette population sur laquelle on possède de l'information.
Il se peut que l'échantillon soit *exhaustif*, c-à-d contienne toute la population, comme lors d'un *recensement* ; c'est rarement le cas, soit parce que ce recueil aurait un coût prohibitif, soit parce qu'il serait simplement impossible : population potentiellement infinie ou constituée en partie d'individus futurs (*prédiction*), échantillon détruit lors de son observation (*contrôle de qualité*).
Puisque l'on devra procéder à une extrapolation, des propriétés de l'échantillon à celle de la population, il est souhaitable que l'échantillon soit le plus *représentatif* possible, c-à-d reflète au mieux les caractéristiques pertinentes de la population considérée. Les biais éventuels résultant de l'influence de variables observables (âge, CSP, etc.) peuvent généralement être analysés et corrigés ; il n'en est pas de même pour l'influence de variables non-observées (souvent non-identifiées) ; on recueille donc de préférence, lorsque c'est possible, un *échantillon aléatoire* assurant à tous les membres de la population des chances égales d'être pris dans l'échantillon, en espérant que la distribution des variables non-observées y reproduira celle de l'ensemble de la population.
L'échantillonnage aléatoire a un autre avantage, théorique celui-ci, qui est de nous placer dans le cadre d'un modèle probabiliste et donc nous permettre d'évaluer, par exemple, les probabilités d'erreur d'un test d'hypothèses ou la probabilité que la valeur vraie d'un paramètre diffère de plus d'une quantitée donnée de sa valeur estimée. La statistique inférentielle est en fait le domaine d'application privilégié de l'outil probabiliste.
Une considération importante, qui n'est pas prise en compte explicitement par le modèle statistique est celle des coûts : recueillir des données est coûteux et ce coût croît avec la taille de l'échantillon ; se tromper aussi est coûteux et

le coût peut dépendre fortement du sens de l'erreur : il est dommage de donner des médicaments chers à quelqu'un qui n'en a pas besoin, mais il peut être catastrophique de ne pas soigner quelqu'un de malade. La *théorie de la décision statistique* introduit les coûts au sein d'un modèle probabiliste ; de plus, dans sa variante dite *modèle bayésien* (adjectif dérivé du nom de Bayes, probabiliste anglais), c'est une théorie purement probabiliste, toute incertitude étant évaluée sous forme d'une probabilité.

2 Statistique descriptive

2.1 Variables et données

Les caractéristiques susceptibles de différer d'un individu de la population à un autre sont appelées *variables* ou encore *caractères*. On distingue les *variables quantitatives* ou *numériques* qui ont pour valeurs des nombres (âge, revenu) et les *variables qualitatives*, encore appelées *variables catégorielles*, qui peuvent prendre leurs valeurs dans des ensembles sans structure particulière (lieu de naissance, nationalité) ; un type intermédiaire est constitué par les *variables ordinales*, qui sont des variables catégorielles dont les valeurs sont ordonnées naturellement (par exemple, l'appartenance politique de la gauche vers la droite).
Une variable quantitative est appelée *discrète* lorsqu'elle ne peut prendre un qu'un nombre fini ou dénombrable de valeurs et *continue* lorsque ses valeurs possibles constituent un continuum de nombres (le plus souvent un intervalle de nombres réels).
Une *observation* donne la valeur d'une ou plusieurs variables pour un individu ; l'ensemble des observations constitue les *données* ; les données sont donc *multidimensionnelles* lorsqu'elles concernent plusieurs variables et *unidimensionnelles* lorsqu'elles en concernent une seule ; elles sont de plus dites *quantitatives*, *qualitatives* ou *ordinales* en fonction de la nature des variables observées.

2.2 Description des données d'une variable quantitative

Soit X une variable quantitative (numérique) et $\mathcal{X}$ l'ensemble de ses valeurs possibles. On suppose qu'on a observé les valeurs de X pour n individus ; la forme brute des données consiste donc en une liste de n valeurs (non nécessairement distinctes)

$$x_1,..,x_i,..,x_n.$$

2.2.1 Tableaux statistiques

Cas d'une variable discrète Pour chaque valeur possible $x_l \in \mathcal{X}$ on ne retient que son *effectif*, n_l, c-à-d le nombre d'individus pour lesquels la variable a cette valeur ; on en déduit la *fréquence* $f_l = \frac{n_l}{n}$ de x_l. D'où un tableau de lignes :

x_l	n_l	f_l

On utilise aussi bien le tableau obtenu en transposant lignes et colonnes du précédent.

Exemple L'observation des résultats de $n = 360$ lancers successifs d'une paire de dés a donné les résultats suivants.

x_l	2	3	4	5	6	7	8	9	10	11	12
n_l	7	25	24	47	58	52	49	34	37	16	11
$100 f_l$	1.94	6.94	6.67	13.06	16.11	14.44	13.61	9.44	10.28	4.44	3.06

Pensez-vous que ces dés sont pipés? On ne pourra en fait tirer des conclusions précises que dans le cadre du modèle probabiliste.

Cas d'une variable continue On divise le domaine (supposé être un intervalle) $\mathcal{X}$ de X en k *classes*, intervalles semi-ouverts à droite (c-à-d dont la borne supérieure est exclue) $[e_{l-1},e_l[$ et note n_l l'effectif de la classe l ($l = 1,..,k$) et f_l sa fréquence. La *fréquence cumulée*, $F_l = \sum_{j=1}^{l} f_j$ indique la proportion d'individus pour lesquels la valeur de la variable est dans l'une des l premières classes, c-à-d vaut strictement moins que e_l.

On rassemble ces données dans un tableau de ligne générique :

$e_{l-1} - e_l$	n_l	f_l	F_l

ou une variante de ce tableau.

Le *centre* de la classe l est $c_l = \frac{e_{l-1}+e_l}{2}$ et son *amplitude* est $h_l = e_l - e_{l-1}$.

Le choix du nombre de classes est assez arbitraire et peu évident : un trop petit nombre de classes fait clairement perdre beaucoup d'information ; en revanche, un nombre élevé de classes peut donner une information non-significative. Par exemple, alors qu'une loi uniforme sur [0,1] donne la même probabilité d'avoir une observation dans chaque intervalle $[\frac{i}{100},\frac{i+1}{100}[$, il se peut que sur 100 observations, il y en ait jusqu'à 5 dans certains intervalles et 0 dans certains autres, ce qui ne suggèrerait pas du tout une distribution uniforme.

2.2.2 Graphiques statistiques

Cas d'une variable discrète On utilise un *diagramme en bâtons* associant, pour toute classe l, au point d'abscisse x_l un segment vertical de longueur f_l (et donc proportionnelle à n_l).

Cas d'une variable continue A chaque classe l on associe un rectangle ayant pour base le segment $[e_{l-1},e_l]$ et de hauteur égale à $\frac{f_l}{h_l}$, donc de surface égale à f_l. Le graphique ainsi obtenu est appelé *histogramme*. L'aire d'un histogramme vaut toujours 1.
Lorsque les classes sont d'égale amplitude (et seulement dans ce cas) les hauteurs des rectangles sont proportionnelles aux f_l et aussi aux n_l.

Remarque : Dans le cas de variables discrètes ayant beaucoup de modalités, on peut aussi opérer des regroupements en classes et tracer un histogramme, plus lisible dans ce cas qu'un diagramme en bâtons.

Il existe des techniques de lissage utilisant des *fenêtres mobiles* qui gomment les discontinuités de l'histogramme créées artificiellement par le découpage en classes. Lorsque les fréquences risquent d'être peu représentatives des probabilités sous-jacentes (s'il y en a), parce qu'il y a trop de classes pour la taille de l'échantillon, l'usage des fenêtres peut aussi compenser ce défaut.

Exemple (G.Saporta)
la mesure des hauteurs de 50 pièces usinées a donné les résultats suivants :

(1) 21.86	(18) 21.9	(35) 21.98
(2) 21.84	(19) 21.89	(36) 21.96
(3) 21.88	(20) 21.92	(37) 21.98
(4) 21.9	(21) 21.91	(38) 21.95
(5) 21.92	(22) 21.91	(39) 21.97
(6) 21.87	(23) 21.92	(40) 21.94
(7) 21.9	(24) 21.91	(41) 22.01
(8) 21.87	(25) 21.93	(42) 21.96
(9) 21.9	(26) 21.96	(43) 21.95
(10) 21.93	(27) 21.91	(44) 21.95
(11) 21.92	(28) 21.97	(45) 21.97
(12) 21.9	(29) 21.97	(46) 21.96
(13) 21.91	(30) 21.97	(47) 21.95
(14) 21.89	(31) 21.97	(48) 21.94
(15) 21.91	(32) 21.98	(49) 21.97
(16) 21.87	(33) 21.95	(50) 21.95
(17) 21.89	(34) 21.89	

Un histogramme (fig.1) permet d'y voir un peu plus clair, mais il faut un lissage par une méthode de noyau pour faire apparaitre une bi-modalité insoupçonnée (fig.2) : il y a des pièces de deux origines différentes.

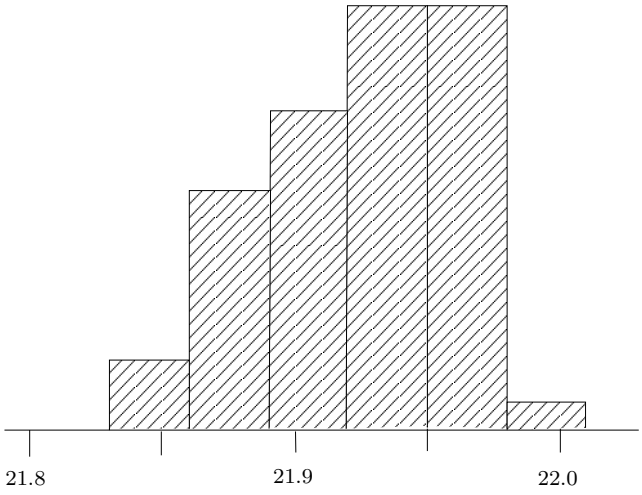

FIG. 1 – *histogramme*

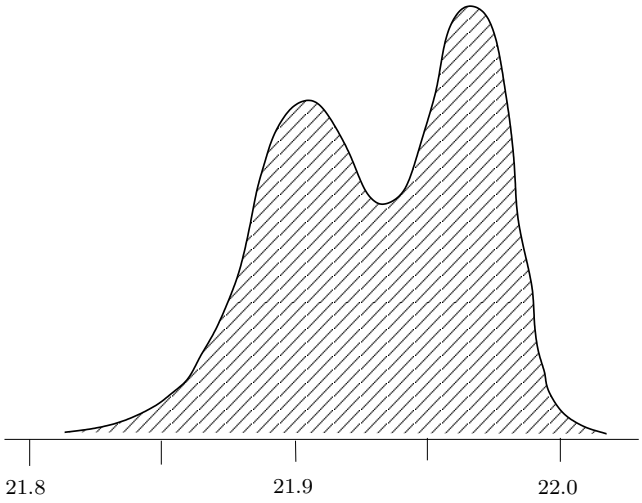

FIG. 2 – *lissage par méthode de noyau*

2.2.3 Caractéristiques statistiques

Une autre façon de fournir une idée de la *distribution* d'une variable X, c-à-d de la répartition de ses valeurs observées, consiste à donner ses caractéristiques essentielles : autour de quelle valeur la distribution est-elle centrée? est-on souvent loin de ce centre? la distribution est-elle symétrique ou pas? etc.

Caractéristiques de la tendance centrale : médiane, moyenne, mode

Médiane
Rangeons les observations par ordre croissant :

$$x_{(1)} \leq x_{(2)} \leq .. \leq x_{(i)} \leq .. \leq x_{(n)}.$$

Dans le cas d'un échantillon de taille n avec n *impair*,

$$M = x_{(\frac{n+1}{2})}$$

est telle qu'il y ait autant d'observations à sa gauche qu'à sa droite.
Dans le cas où n est *pair*, $x_{\frac{n}{2}}$ et $x_{(\frac{n}{2}+1)}$ ont toutes deux la propriété, plus faible, qu'au moins 50% des observations leur sont inférieures ou égales et, simultanément, au moins 50% des observations leurs sont supérieures ou égales.
On définit une *médiane* M d'une distribution par la propriété précédente ; lorsque M n'est pas unique (pour n pair), on privilégie en général le milieu de l'intervalle médian, qui est donné par :

$$M = \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2}.$$

Lorsqu'on ne connait qu'un histogramme des données, on ne peut déterminer que la *classe médiane*, $[e_{l-1},e_l[$, où l est telle que : $F_{l-1} < \frac{1}{2}$ et $F_l > \frac{1}{2}$; par interpolation linéaire, on prendra pour médiane :

$$M = e_{l-1} + h_l \frac{0{,}5-F_{l-1}}{f_l}$$

Notons que le concept de médiane a un sens pour toute variable ordinale ; il n'en est pas de même pour la moyenne, définie ci-après.
Moyenne
La *moyenne (arithmétique)* est définie comme :

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

.

Mode
Le *mode* est la valeur la plus fréquente dans le cas d'une distribution discrète. Dans le cas d'une distribution continue, connue par un histogramme, c'est la classe de plus grande hauteur, mais cette notion est alors peu significative puisqu'elle dépend du découpage en classes choisi.

Caractéristiques de dispersion : étendue, intervalle interquartile, écart-type et variance

Etendue

L'*étendue* est l'intervalle w séparant les valeurs extrêmes observées :

$$w = x_{max} - x_{min} \text{ où } x_{max} = x_{(n)} \text{ et } x_{min} = x_{(1)}$$

de l'échantillon ordonné.

Intervalle interquartile

Idéalement, les quartiles diviseraient les observation en quatre tranches égales. Comme on rencontre le même problème qu'avec la définition des médianes, on définit :

le premier quartile Q_1 par la propriété qu'au moins $\frac{1}{4}$ des observations sont inférieures ou égales à Q_1 et simultanément au moins $\frac{3}{4}$ des observations lui sont supérieures ou égales ;

et le troisième quartile Q_3 par la propriété qu'au moins $\frac{3}{4}$ des observations sont inférieures ou égales à Q_1 et simultanément au moins $\frac{1}{4}$ des observations lui sont supérieures ou égales

(le deuxième quartile Q_2 n'étant autre qu'une médiane M).

Comme il n'y a pas unicité, on prévilégie en général pour chaque quartile la demi-somme des valeurs extrêmes qu'il peut prendre.

L'*intervalle interquartile* est l'intervalle $Q_3 - Q_1$. C'est un indicateur de dispersion beaucoup moins instable que l'étendue.

On peut aussi utiliser l'*intervalle interdécile* $D_9 - D_1$, où le *premier décile*, D_1, se définit comme Q_1 mais en remplaçant $\frac{1}{4}$ par $\frac{1}{10}$ et $\frac{3}{4}$ par $\frac{9}{10}$; définition symétrique pour le *dernier décile*, D_9.

Exemple

On recueille 12 observations d'une variable numérique discrète à valeurs dans {0,1,2,3,4,5,6}:

6	4	0	5	1	3	3	6	5	1	2	5

qui une fois ordonnées donne

0	1	1	2	3	3	4	5	5	5	6	6

D'où le diagramme en bâtons (fig.3), où l'on a indiqué médiane M et quartiles Q_1 et Q_3.

Ecart-type et variance

La *variance* s^2 est définie par :

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

et l'*écart-type* s n'est autre que sa racine carrée.

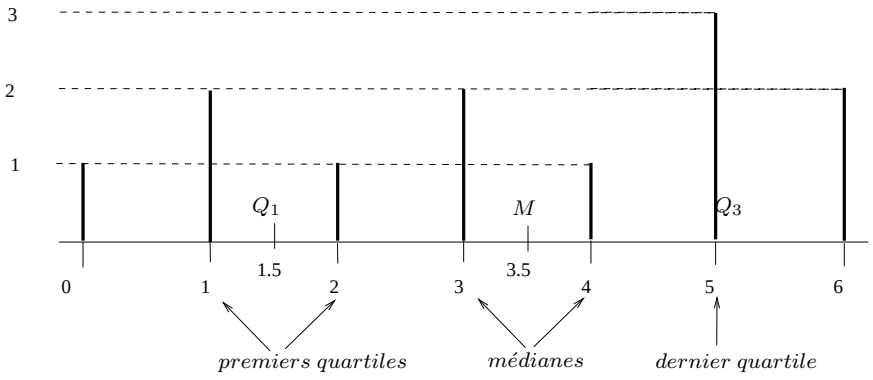

FIG. 3 – *Diagramme en bâtons, médiane et quartiles*

Relations entre les indicateurs de tendance centrale et de dispersion

La distance quadratique moyenne des observations à une valeur c

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

est minimale pour $c = \bar{x}$ et vaut donc alors s^2.

La distance de la valeur absolue moyenne

$$\frac{1}{n}\sum_{i=1}^{n} \mid x_i - c \mid$$

est, elle, minimale pour $c = M$.

2.3 Description de la liaison entre deux variables

2.3.1 Liaison entre deux variables quantitatives

Soit X et Y deux variables quantitatives (numériques) dont on a observé les valeurs conjointes (x_i, y_i) pour n individus ; on se demande s'il existe une liaison entre ces deux variables, c-à-d si les valeurs les plus élevées de X vont plutôt avec les valeurs les plus élevées de Y (liaison positive) ou au contraire avec ses valeurs les moins élevées (liaison négative) ou encore s'il n'y a pas de liaison apparente (indépendance). Introduisant les moyennes de X et de de Y,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, , \; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \, ,$$

on peut regarder, pour chaque i, si les écarts observés $x_i - \bar{x}$ et $y_i - \bar{y}$ sont ou non de même sens, donc le signe du produit $(x_i - \bar{x})(y_i - \bar{y})$.
La *covariance* observée

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

est alors un indicateur de la liaison moyenne entre X et Y, mais il n'est pas indépendant des unités choisies pour mesurer ces variables ; on normalise

donc en divisant par le produit des écarts-types de X et de de Y,

$$s_x = [\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{\frac{1}{2}}\,, s_y = [\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2]^{\frac{1}{2}}.$$

On obtient ainsi le *coefficient de corrélation linéaire*

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Le coefficient de corrélation linéaire est compris entre -1 et 1.
Le qualificatif "linéaire" vient de ce que les bornes sont atteintes si et seulement si Y dépend linéairement (et pas seulement fonctionnellement) de X :

$$y_i = a.x_i + b$$
$$(r = 1 \text{ si } a > 0\,;\, r = -1 \text{ si } a < 0).$$

Un coefficient de corrélation proche de 1 (ou -1) indique donc une forte dépendance, la moyenne de Y à X fixé, $X = x$, croissant (ou décroissant) à peu près linéairement avec x.
En revanche un coefficient de corrélation nul, ne correspond pas forcément à une indépendance de X et Y ; on peut très bien avoir $r = 0$ avec une dépendance fonctionnelle non-linéaire. Le coefficient r a en outre l'inconvénient d'être très sensible à la position des points les plus éloignés de $(\bar{x},\bar{y})$ qui peuvent être des points "aberrants" (parce qu'atypiques ou erronés).

Voici des exemples graphiques dans la fig. 4 (le nuage de points est supposé uniformément réparti dans la zone ombrée) :

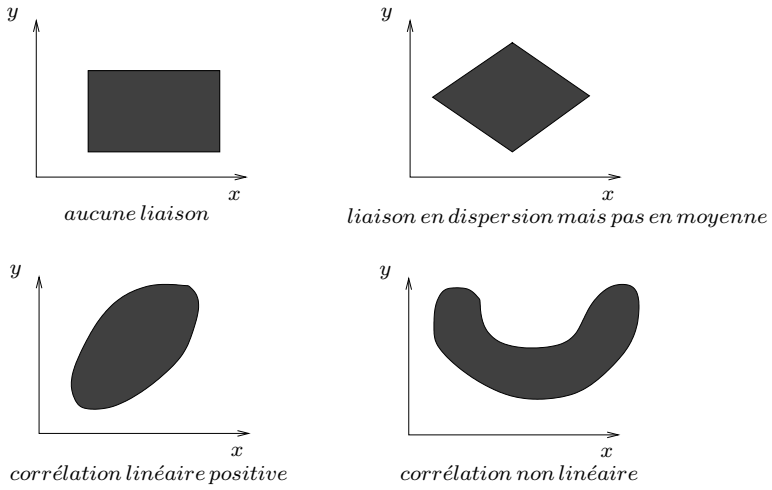

FIG. 4 – *Corrélations entre deux variables*

Moyennes, écart-types et coefficient de corrélation linéaire apportent de l'information sur le couple (X,Y), mais ne peuvent à eux seuls décrire complètement ces variables et leur liaison. Par exemple, elles ont des valeurs identiques

$$\bar{x} = 9;\ \bar{y} = 7.5;\ {s_x}^2 = 10;\ {s_y}^2 = 3.75;\ r = 0.82$$

pour les quatre échantillons de la fig. 5 :

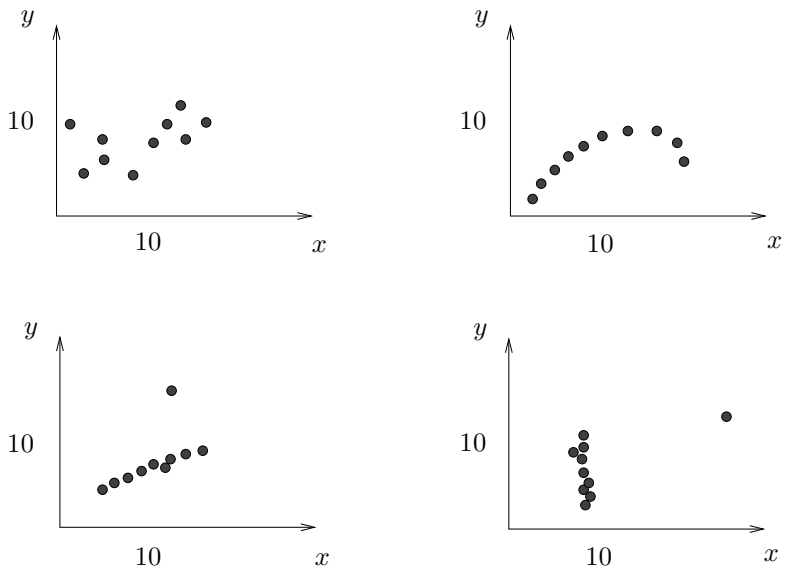

FIG. 5 – *Quatre échantillons de paramètres identiques*

2.3.2 Liaison entre deux variables ordinales

Une variable (qualitative) ordinale prend des valeurs qui peuvent être ordonnées : les opinions politiques vont de l'*extrême gauche* à l'*extrême droite* ; la violence d'un film de *très violent* à *sans aucune violence* ; le confort d'une voiture de *très confortable* à *très inconfortable*. On peut alors ranger les observations d'une telle variable de 1 à n dans l'ordre associé aux valeurs.

Pour étudier l'éventuelle liaison entre deux variables qualitatives, X et Y, on peut comparer les rangs respectifs, r_i et t_i, des valeurs x_i et y_i de chaque individu (ou objet) i. Voyons les deux principales caractéristiques utilisées.

Coefficient de corrélation des rangs de Spearman

C'est tout simplement le coefficient de corrélation linéaire appliqué aux rangs :

$$r_S = \frac{cov(r,t)}{s_r s_t}.$$

En appelant $d_i = r_i - t_i$ la différence des rangs d'un même objet i, on montre que le coefficient de Spearman s'écrit encore :

$$r_S = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n-1)^2}.$$

Il varie de -1, pour des classements inverses l'un de l'autre, à 1 pour des classements identiques.

Coefficient de corrélation des rangs de Kendall

Il y a $\frac{n(n-1)}{2}$ couples d'objets i et j ; pour chacun, on regarde si le classement de i par rapport à j est le même pour les deux variables ou pas et compte : $+1$ si les deux classements concordent (c-à-d $x_i < x_j \iff y_i < y_j$) et -1 s'ils sont discordants (c-à-d $x_i < x_j \iff y_i > y_j$).
En additionnant tous les $+1$ et -1, on obtient la somme S et après normalisation (c-à-d division par $\frac{n(n-1)}{2}$), le coefficient τ ("tau") de Kendall :

$$\tau = \frac{2S}{n(n-1)}.$$

Comme r_S, τ varie de -1, pour des classements inverses l'un de l'autre à 1 pour des classements identiques.

Exemple

Un oenologue a classé 12 vins de table rouges du 1^{er} au $12^{ème}$ (variable x_i). Un deuxième oenologue a un classement différent(variable y_i), comme le montre le tableau suivant :

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	3	7	5	1	6	2	10	4	12	11	9	8

Le coefficient de Spearman est $r_S = 0.60$ et celui de Kendall est $\tau = 0.36$.
La liaison est-elle significative? On ne pourra répondre à la question que dans le cadre d'hypothèses probabilistes.

2.3.3 Liaison entre deux variables qualitatives

Soit X et Y deux variables qualitatives, à r et s catégories. Le résultat des observations de n individus est présenté dans un tableau croisé, dit *tableau de contingence*, à r lignes et s colonnes ; la case de ligne i et colonne j indique le nombre d'individus, n_{ij}, pour lesquels on a observé les valeurs conjointes (x_i, y_i).
Les marges verticale et horizontale du tableau indiquent, respectivement, le nombre d'individus $n_{i.} = \sum_{j=1}^{s} n_{ij}$ pour lesquels $X = x_i$ et celui, $n_{.j} = \sum_{i=1}^{r} n_{ij}$ pour lesquels $Y = y_j$.

.

	y_1	y_2	.....	y_j	.....	y_s	
x_1	n_{11}	n_{12}				n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}				n_{2s}	$n_{2.}$
...							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{is}	$n_{i.}$
...							
x_r	n_{r1}	n_{r2}				n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	n

.

S'il y a indépendance, c-à-d s'il n'y a pas de liaison, entre les variables, on peut s'attendre à ce que les lignes du tableau, qui donnent les distributions observées de Y à $X = x_i$ donné, soient à peu près proportionnelles (elles le seraient exactement, s'il n'y avait pas les fluctuations d'échantillonnage) et donc aussi proportionnelles à la marge horizontale. Or

$$\frac{n_{ij}}{n_{i.}} = k, \forall j \;\Rightarrow\; \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

(par sommation des numérateurs et des dénominateurs), d'où : $\frac{n n_{ij}}{n_{i.} n_{.j}} = 1$.

Un indice d'écart à l'indépendance, le χ^2 .

La somme des écarts quadratiques à 1 des $\frac{n\, n_{ij}}{n_{i.} n_{.j}}$, pondérée par les fréquences (théoriques, sous hypothèse d'indépendance) des cellules, $\frac{n_{i.} n_{.j}}{n}$, s'écrit :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{i.} n_{.j}}{n} \left[\frac{n\, n_{ij}}{n_{i.} n_{.j}} - 1\right]^2$$

ou encore :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}}$$

On voit que χ^2 est nul (ou très voisin de 0) s'il y a indépendance. Sa borne supérieure est donnée par :

$$\chi^2 \leq n\, inf(s-1, r-1),$$

et est atteinte lorsqu'il y a dépendance fonctionnelle : Y fonction de X lorsque $\chi^2 = n(s-1)$ et X fonction de Y lorsque $\chi^2 = n(r-1)$.

Exemple

On cherche à évaluer l'effet à 6 mois sur une maladie de longue durée (rhumatismes) d'un nouveau médicament. La variable X (le médicament donné) peut prendre 3 valeurs : *nouveau*, *standard*, *placebo* ; l'état du patient après usage du médicament, variable Y peut prendre les valeurs : *guérison*, *amélioration*, *aucun effet*, *détérioration*. On observe 150 patients

ayant reçu le nouveau médicament pour 50 d'entre eux, le médicament standard pour 50 autres et le placebo pour les 50 restant. Les résultats sont donnés dans le tableau suivant :

	guérison	*amélioration*	*aucun effet*	*détérioration*	
nouveau	7	18	20	5	50
standard	5	23	19	3	50
placebo	2	9	33	6	50
	14	50	72	14	150

On trouve $\chi^2 = 17.50$. Comme il apparait que la $3^{ème}$ ligne (*placebo*) est assez différente des deux autres, on regarde aussi le tableau partiel :

	guérison	*amélioration*	*aucun effet*	*détérioration*	
nouveau	7	18	20	5	50
standard	5	23	19	3	50
	12	41	39	8	100

où $\chi^2 = 1.48$. Les effets des deux médicaments paraissent proches (et différents de celui du placebo), mais on ne pourra tirer des conclusions précises que dans le cadre du modèle probabiliste.

3 Le modèle probabiliste

3.1 Des fréquences aux probabilités

Dans leur *interprétation fréquentiste*, les probabilités sont des concepts abstraits, numériques, dont les valeurs sont estimées à partir de fréquences observées : par exemple, observant une suite de N tirages à *(P)ile* ou *(F)ace*, où il y a N_F *F)et constatant que le rapport* $\frac{n_F}{n}$ *parait tendre, avec des oscillations d'amplitude de plus en plus faibles, vers une limite proche de* $\frac{N_F}{N}$, *on va prendre ce rapport pour estimation de la probabilité deF :*
$P(F) =_{est} \frac{N_F}{N}$.

3.2 Propriétés des probabilités

Comme une fréquence, une probabilité est donc comprise entre 0 *et*1 *:*

$$0 \leq P(A) \leq 1 \ \ \forall A$$

Les fréquences sont additives sur les événements incompatibles, mais sous-additives en général ; par exemple :

Répartition de la population française par sexe et âge

(en millions)	*jeunes*	*adultes*	*seniors*	
hommes	*7.5*	*17.5*	*4*	*29*
femmes	*7.5*	*17.5*	*6*	*31*
	15	*35*	*10*	*60*

Il y a $\frac{17.5}{60}$ *d'hommes adultes et* $\frac{6}{60}$ *de femmes seniors, donc* $\frac{17.5+6}{60} = \frac{23.5}{60}$ *de gens qui sont l'un ou l'autre ; en revanche, alors qu'il y a* $\frac{29}{60}$ *d'hommes et* $\frac{10}{60}$ *de seniors, il n'y a que* $\frac{29+10-4}{60} = \frac{35}{60}$ *de gens qui sont l'un ou l'autre, car il faut retrancher les* $\frac{4}{60}$ *d'hommes seniors, qui sont l'un et l'autre.*

Estimant la probabilité qu'une personne tirée au hasard soit un homme, événement H, *par la fréquence des hommes dans la population* $P(H) =_{est} \frac{29}{60}$ *et faisant de même pour les autres événements,* $F(emme)$, $E(nfant)$, $A(dulte)$, $S(enior)$, *ainsi que pour leurs intersections et unions* $H \cap E$, $H \cup E$, *etc., on aura donc* $P([H \cap A] \cup [F \cap S]) = P(H \cap A) + P(F \cap S)$, *mais* $P(H \cup S) = P(H) + P(S) - P(H \cap S)$.

Donc, les probabilités vérifient toujours

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \forall A,B$$

d'où leur sous-additivité

$$P(A \cup B) \leq P(A) + P(B) \quad \forall A,B$$

ainsi que leur additivité sur les événements incompatibles

$$P(A \cup B) = P(A) + P(B) \quad \forall A,B \text{ tels que } A \cap B = \emptyset$$

En fait, on suppose de plus que les probabilités sont σ-additives ("sigma"-additives), c-à-d que

$$\forall A_n, n \in \mathbb{N},\ \textit{tels que}\ A_n \cap A_m = \emptyset,\ P(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} P(A_n).$$

3.2.1 Donnée d'une loi de probabilité

Connaître une loi de probabilité, c'est être capable d'associer à tout événement sa probabilité, ce qui parait a priori complexe. La propriété d'additivité peut être mise à profit pour simplifier les données nécessaires :

Cas discret *Dans ce cas, il existe un nombre fini ou dénombrable d'événements élémentaires $e_k, k \in\ K \subseteq \mathbb{N}$, ayant (par définition) les propriétés suivantes :*
i) ils forment une partition de l'événement certain, Ω, c-à-d qu'ils sont deux à deux incompatibles et ont pour union Ω (autrement dit : un et un seul d'entre eux sera réalisé) : $e_k \cap e_l = \emptyset\ \ \forall k \neq l$ et $\bigcup_{k=1}^{\infty} e_k = \Omega$;
ii) tout autre événement est l'union d'une partie d'entre eux :
$\forall A, \exists K(A)$ tel que $A = \bigcup_{k\in K(A)} e_k$.
Il résulte alors de la σ-additivité des probabilités que :
$\forall A,\ P(A) = \sum_{k\in K(A)} P(e_k)$.
Les probabilités élémentaires déterminent donc complètement la loi P.
Cela fait encore beaucoup d'information en général. C'est pourquoi on fait souvent l'hypothèse, qu'il faut encore justifier, que les probabilités élémentaires sont données par une expression générique connue.

Cas continu *Dans ce cas, il y a une infinité non-dénombrable d'événements élémentaires et chacun est de probabilité nulle. En revanche, la probabilité que la réalisation d'une variable uni-dimensionnelle X soit dans un intervalle I est positive et égale à la valeur de l'intégrale sur I d'une fonction, positive ou nulle, p, dite densité (de probabilité) de la loi de X :*

$$P(X \in I) = \int_I p(x)\ dx$$

Ici, c'est la connaissance de la fonction p qui est suffisante pour pouvoir calculer la probabilité de n'importe quel événement lié à X. Si p a une expression analytique connue, dépendant de certains paramètres, il ne reste qu'à estimer ces paramètres.

La considération des intervalles de la forme $]-\infty,x[$ conduit à la fonction de répartition, F, de la variable X

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(x')\ dx'$$

Une fonction de répartition F est croissante et vérifie

$$\lim_{x \leftarrow -\infty} F(x) = 0 \,;\, \lim_{x \leftarrow +\infty} F(x) = 1$$

La densité de probabilité de X étant la fonction dérivée de sa fonction de répartition, la donnée de F est équivalente à celle de P.

On peut aussi définir la fonction de répartition d'une variable discrète

$$F(x) = P(X < x) = \sum_{k:x_k<x} p_k$$

mais alors que pour une loi continue F est dérivable et donc continue, pour une loi discrète F fait un saut en tout x_k.

Pour les variables bi-dimensionnelles, ou couples de variables, continues, (X,Y) la densité de probabilité et la fonction de répartition sont des fonctions de deux variables, dont les valeurs sont donc notées, respectivement, $p(x,y)$ et $F(x,y)$ et vérifient :

$$\forall x,y, F(x,y) = P(X < x, Y < y) = \int \int_{\{(x',y'):x'<x,y'<y\}} p(x,y)dxdy$$

3.2.2 Caractéristiques d'une loi de probabilité

On retrouve pour les lois de probabilité les mêmes caractéristiques que pour les distributions d'observations :

Soit X une variable de loi P, de fonction de répartition F et, selon qu'elle est discrète ou continue, de probabilités élémentaires p_k ou de densité de probabilité p.

Tout M tel que simultanément $P(X \leq M) \geq \frac{1}{2}$ et $P(X \geq M) \geq \frac{1}{2}$ est une Médiane de P.

L'espérance mathématique ou moyenne de X, notée $E(X)$ est donnée par $E(X) = \sum x_k p_k$ (X discrète) ou $E(X) = \int x\, p(x)dx$ (X continue).
Noter que l'espérance mathématique n'existe pas toujours, car une série et une intégrale peuvent diverger.

L'espérance mathématique a les propriétés générales suivantes (conséquences immédiates de celles des sommes et des intégrales) :

$$E(aX + b) = aE(X) + b$$
$$\forall X,Y, E(X + Y) = E(X) + E(Y)$$

Le mode Mo de P (pas toujours unique) est caractérisé par $p(Mo) = \max_k p(x_k)$ *(X discrète) ou* $p(Mo) = \max_x p(x)$ *(X continue).*

La caractéristique de dispersion la plus utilisée est la variance, notée $V(X)$ *ou* σ^2 *donnée par* $\sigma^2 = \sum[x_k - E(X)]^2 p_k$ *(X discrète) ou* $\sigma^2 = \int [x - E(X)]^2 \, p(x)dx$ *(X continue).*

La variance de$X$ est donc la moyenne des écarts quadratiques entre les valeurs prises par X et son espérance $E(X)$. La racine carrée de la variance, σ, est l'écart-type.
Noter que variance et donc écart-type n'existent pas toujours.

Propriétés générales de la variance :

$$V(X) = E(X^2) + E(X)^2$$
$$\forall X,Y, V(X+Y) = V(X) + V(Y) + 2cov(X,Y)$$

où $cov(X,Y)$ *est la covariance de X et Y donnée, si X et Y sont des variables discrètes et* $p_k = P(X = x_k, Y = y_k)$*, par*

$$cov(X,Y) = \sum [x_k - E(X)][y_k - E(Y)]p_k$$

et si ce sont des variables continues, de densité de probabilité $p(x,y)$*, par*

$$cov(X,Y) = \int [x - E(X)][y - E(Y)]p(x,y)dxdy$$

3.2.3 Exemples de lois de probabilité

Lois discrètes à support fini

Loi de BERNOULLI ***de paramètre*** p
Variable X à support $\mathcal{X} = \{0,1\}$ *telle que* $P(X = 1) = p$ *et* $P(X = 0) = 1 - p$.

$$E(X) = p \, ; \, V(X) = p(1-p)$$

Loi binomiale de paramètres n ***et*** p***,*** $\mathcal{B}(n,p)$
Supposons que vous tiriez n fois de suite une boule d'une urne contenant deux sortes de boules, les unes marquées d'un 1*, les autre marquées d'un* 0 *; supposons en outre que les tirages se font avec remise, c-à-d en remettant à chaque fois la boule tirée. Si vous pensez que la proportion de boules marquées* 1 *dans l'urne est p, une hypothèse naturelle est que vous avez à chaque tirage, indépendamment des résultats des tirages précédents, la même probabilité p de tirer une boule marquée* 1 *et donc* $(1 - p)$ *de tirer une boule marquée* 0*. La probabilité d'obtenir une séquence donnée de* 1 *et de* 0 *comportant k boules* 1 *et* $(n - k)$ *boules* 0 *est alors le produit de ces probabilités et vaut donc* $p^k(1 - p)^{(n-k)}$.

L'ordre des 1 *et des* 0 *dans une séquence n'a donc pas d'influence sur sa probabilité, qui ne dépend que du nombre de* 1 *; si la variable* X *à laquelle on s'intéresse est justement le nombre de* 1 *tirés,* $P(X = k)$*, probabilité que ce nombre vaille* k*, est la somme des probabilités de toute les séquences comportant un nombre de* 1 *égal à* k *: comme il y en a* $\binom{n}{k}$*, on obtient*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}, k = 0, \cdots, n$$

La loi de X *est appelée loi binomiale de paramètres* n *et* p *et notée* $\mathcal{B}(n,p)$*.*

$$E(X) = np\,;\, V(X) = np(1-p)$$

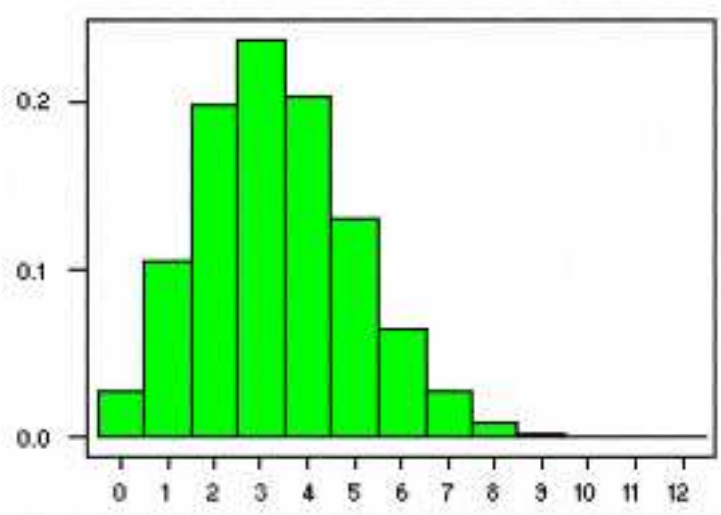

FIG. 6 – *loi binomiale* $\mathcal{B}(20, \frac{1}{6})$

Lois discrètes à support dénombrables

Loi de POISSON ***de paramètre*** λ**,** $\mathcal{P}(\lambda)$

Une variable discrète X *suit la loi de* POISSON *de paramètre* $\lambda > 0$ *lorsque les événements élémentaires* $e_k =$ *"*$X = k$ *",* $k \in \mathbb{N}$ *ont pour probabilités :*

$$P(X = k) = \frac{\exp\{-\lambda\}.\lambda^k}{k!}.$$

En télécommunications, le nombre d'appels téléphoniques sur un réseau pendant une période donnée satisfait à une loi de POISSON*. Lorsqu'on le sait, il n'y a plus qu'un nombre à estimer, le paramètre* λ*, qui est à la fois l'espérance mathématique et la variance de la loi :*

$$E(X) = \lambda\,;\, V(X) = \lambda$$

Loi géométrique de paramètre p

C'est la loi de la variable donnant le nombre d'essais nécessaires pour que se produise un événement de probabilité p.

$$P(X = k) = p(1-p)^{(k-1)}, k = 1, \cdots, n, \cdots$$

$$E(X) = \frac{1}{p}\,;\, V(X) = \frac{1-p}{p^2}$$

Lois continues

Loi exponentielle de paramètre λ

Elle a pour support $\mathbb{N}_+^*$ *et pour densité*

$$p(x) = \lambda \exp\{-\lambda x\}, x > 0\,;$$

d'où,

$$E(X) = \frac{1}{\lambda}\,;\, V(X) = \frac{1}{\lambda^2}$$

La durée de vie de nombreux composants suit une loi exponentielle, d'où son importance en fiabilité : λ y est appelé taux de défaillance ; $E(X) = \frac{1}{\lambda}$ est le temps moyen entre défaillances.

Loi normale de paramètres m ***et*** σ, $\mathcal{N}(m,\sigma)$

Une famille de lois continues qui joue un rôle très important, car elle contient souvent une très bonne approximation de la loi réelle, est celle des lois normales, dites encore lois de Laplace-Gauss *; c'est une famille à deux paramètres, m et σ, de densité positive sur tout $\mathbb{R}$:*

$$p(x) = \frac{1}{\sqrt{2\pi}.\sigma} \exp\{-\frac{1}{2}(\frac{x-m}{\sigma})^2\};$$

On montre que

$$E(X) = m\,;\, V(X) = \sigma^2$$

Les deux paramètres sont donc l'espérance, m, de la loi et son écart-type, σ. La forme de la courbe représentant la densité de probabilité d'une loi normale lui a valu le nom de "courbe en cloche" (fig. 7)

Cette famille de lois se généralise au cas de plusieurs dimensions. En particulier, un couple de variables (X,Y) suit une
loi normale bi-dimensionnelle *lorsqu'elle a pour densité dans* $\mathbb{R}^2$

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$
$$\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-m_x}{\sigma_x})^2 - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + (\frac{y-m_y}{\sigma_y})^2]\},$$

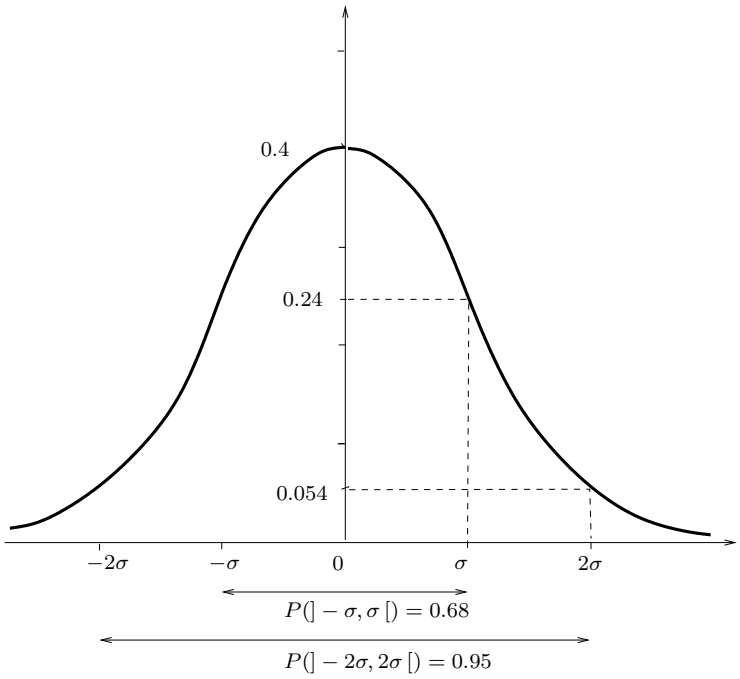

FIG. 7 – *densité de la loi normale* $\mathcal{N}(0,\sigma)$

où m_x *et* σ_x *sont l'espérance et l'écart-type de* X*,* m_y *et* σ_y *ceux de* Y *et* ρ *est le coefficient de corrélation linéaire entre* X *et* Y*. Par définition,*

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

On peut donc prendre comme 5^{eme} *paramètre de la loi soit* ρ *soit* $cov(X,Y)$*.*

3.3 Probabilités conditionnelles

La probabilité de réalisation que l'on attribue à un événement dépend de l'information dont on dispose, en particulier de ce que l'on peut déjà savoir à propos d'autres événements. Reprenons l'exemple précédent de

La population française (suite) *La proportion de femmes dans la population française est de $\frac{31}{60}$ (ce qui est proche de $\frac{1}{2}$) ; c'est donc aussi la probabilité pour qu'une personne qui y est tirée au hasard soit une femme. En revanche, parmi les seniors, la proportion de femmes est de $\frac{6}{10}$ (ce qui est nettement plus élevé) ; si l'on apprend que la personne tirée est senior, alors la probabilité que ce soit une femme doit maintenant être estimée à $\frac{6}{10}$.*

Pour prendre en compte l'impact de l'information sur les probabilités, il faut utiliser le concept de probabilité conditionnelle : Etant donné deux événements A et B, où B a une probabilité positive, on appelle probabilité de A conditionnellement à B (ou "si B", ou encore "sachant B"), la quantité :

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

Considérant alors une partition $B_k, k \in K$ et remarquant que

$$A = A \cap \Omega = A \cap [\cup_{k \in K} B_k] = \cup_{k \in K} [A \cap B_k],$$

il vient, par additivité sur les événements incompatibles,

$$P(A) = \sum_{k \in K} P(A \cap B_k)$$

dont on déduit, en utilisant la définition ci-dessus, la formule des probabilités composées

$$P(A) = \sum_{k \in K} P(A/B_k) P(B_k).$$

La probabilité de A est donc la moyenne de ses probabilités conditionnelles aux B_k, pondérées par les probabilités de ces événements.

Des formules précédentes, on peut alors tirer la formule de Bayes, *connue aussi sous le nom de formule de probabilité des causes. Elle tire son nom d'une situation générique où la réalisation d'un événement A,* l'*effet (symptôme, panne, etc.) est susceptible d'être entrainée par la présence de l'une ou l'autre de diverses causes (maladies, pièces déficientes, etc.) : événements B_k.*
$P(A/B_k)$ est donc la probabilité que la k^{eme} cause entraine l'effet considéré, et
$P(B_k/A)$ la probabilité, cet effet ayant été observé, quil soit dû à la k^{eme}

cause.
Partant de la formule des probabilités composées, plusieurs applications successives de la définition d'une probabilité conditionnelle amènent à l'expression

$$P(B_{k_0}/A) = \frac{P(B_{k_0} \cap A)}{P(A)} = \frac{P(A \cap B_{k_0})}{P(A)} = \frac{P(A/B_{k_0})P(B_{k_0})}{\sum_{k \in K} P(A/B_k)P(B_k)}.$$

qui constitue la formule de BAYES.
Voyons un exemple.

Diagnostic médical *Deux maladies, dont l'une est nettement plus fréquente que l'autre, peuvent se manifester par un même symptôme, très souvent présent avec la première maladie, beaucoup moins avec la seconde, jamais présent sinon.*
Un malade présente ce symptôme ; que doit-on diagnostiquer?
Notons les événements comme suit : B_1=“$1^{ère}$ *maladie présente*” ; B_2=“$2^{ème}$ *maladie présente*” ; A=“*présence du symptôme*” ; A^c=“*son absence*”. *Les données sont les suivantes :*

	A	A^c	
B_1	*80*	*20*	*100*
B_2	*100*	*800*	*900*
	180	*820*	*1000*

Les calculs donnent :
$P(A/B_1) = \frac{8}{10}$; $P(A/B_2) = \frac{1}{9}$; $P(B_1) = \frac{1}{10}$; $P(B_2) = \frac{9}{10}$;
d'où : $P(B_1/A) = \frac{4}{9}$ *et* $P(B_2/A) = \frac{5}{9}$.
Donc, la maladie ayant très peu de chance d'être accompagnée du symptôme est néanmoins la plus probable, parce qu'elle est beaucoup plus répandue que l'autre.
Un biais de jugement souvent observé base rate fallacy consiste à conclure à la présence de la maladie que le symptôme accompagne le plus souvent, en oubliant qu'il faut aussi prendre en compte les raretés relatives des maladies.

3.4 Indépendance

3.4.1 Indépendance de deux événements

On dit que deux événements, A et B sont indépendants lorsque :

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

Cette égalité est trivialement vérifiée lorsque $P(B) = 0$; lorsque $P(B) > 0$, la probabilité conditionnelle $P(A/B)$ existe et est égale à $P(A)$; en fait :

Les trois propriétés suivantes sont équivalentes :
(i) $P(A \cap B) = P(A).P(B)$ *(les événements* A *et* B *sont indépendants) ;*
(ii) $P(B) = 0$ *ou* $P(A/B) = P(A)$*;*
(iii) $P(A) = 0$ *ou* $P(B/A) = P(B)$*.*

Ainsi deux événements sont indépendants lorsqu'observer l'un des deux n'apporte pas d'information sur l'autre, en ce sens que sa probabilité ne change pas.

En fait l'indépendance n'est pas une propriété du couple (A,B) *mais du couple* $(\{A,A^c\},\{B,B^c\})$ *car l'on montre facilement que :*
Lorsque deux événements A *et* B *sont indépendants,* A *et* B^c*, ainsi que* A^c *et* B *et que* A^c *et* B^c *le sont aussi.*

3.4.2 Indépendance de deux variables

L'extension naturelle de cette propriété aux couples de variables est donc la suivante :

Deux variables discrètes X *et* Y *sont indépendantes lorsque* $\forall x,\forall y$*, les événements* $X = x$ *et* $Y = y$ *sont indépendants, ce qui s'exprime indifféremment par chacune des propriétés suivantes:*
(i) $\forall x,\forall y, P(X = x \cap Y = y) = P(X = x).P(Y = y)$ *;*
(ii) $\forall x,\forall y\, t.q.\, P(Y = y) > 0, P(X = x \,/\, Y = y) = P(X = x)$ *;*
(iii) $\forall y,\forall x\, t.q.\, P(X = x) > 0, P(Y = y \,/\, X = x) = P(Y = y)$*.*

Deux variables continues X *et*Y *sont indépendantes lorsque* $\forall I,\forall J$*, intervalles, les événements* $X \in I$ *et* $Y \in J$ *sont indépendants.*

Il suffit en fait que les fonctions de répartition, F_X*,* F_Y *de* X *et* Y *et* $F_{(X,Y)}$ *du couple satisfassent*

$$\forall x,y, F_{(X,Y)}(x,y) = F_X(x) \times F_Y(y)$$

ou encore que les densités de probabilité p_X*,* p_Y *de* X *et* Y *et* $p_{(X,Y)}$ *du couple satisfassent*

$$\forall x,y, p_{(X,Y)}(x,y) = p_X(x) \times p_Y(y).$$

Propriété
La covariance de deux variables indépendantes X *et* Y *est toujours nulle ; montrons-le pour deux variables continues :*

$$cov(X,Y) = \int [x - E(X)][y - E(Y)]p_{(X,Y)}(x,y)dxdy$$

Si $p_{(X,Y)}(x,y) = p_X(x) \times p_Y(y)$*, l'intégrale double se décompose en produit de deux intégrales simples, nulles toutes les deux :*

$$cov(X,Y) = \int [x - E(X)]p_X(x)dx \int [y - E(Y)]p_Y(y)dy = 0 \times 0 = 0$$

La réciproque n'est pas vraie : par exemple le couple de variables discrètes (X,Y) *prenant chacun des* 4 *couples de valeurs* $(-1,0), (0,-1), (0,1)$ *et* $(1,0)$ *avec probabilité* $\frac{1}{4}$ *satisfait* $E(X) = E(Y) = 0$ *et* $cov(X,Y) = E(XY) = 0$*, bien que* X *et* Y *ne soient pas indépendantes puisque la loi de* Y *sachant que* $X = -1$ *(et celle sachant que* $X = 1$*) est la loi certaine en* 0*, alors que sachant que* $X = 0$*, c'est la loi donnant* $Y = -1$ *et* $Y = 1$ *avec équiprobabilité. Cependant, pour* (X,Y) *suivant une loi normale à deux dimensions, la nullité de* $cov(X,Y)$ *entraine l'indépendance de* X *et* Y *; en effet, il suffit de remarquer qu'alors* $p(x,y)$ *se factorise en*

$$p(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x} \times \exp\{-\frac{1}{2}(\frac{x - m_x}{\sigma_x})^2\} \times \frac{1}{\sqrt{2\pi}\sigma_y} \times \exp\{-\frac{1}{2}(\frac{y - m_y}{\sigma_y})^2\}$$

3.4.3 Indépendance mutuelle de n variables

Soit n *variables,* $(X_1,X_2,\cdots,X_k,\cdots,X_n)$*; on dit qu'elles sont mutuellement indépendantes lorsque tout événement lié à une partie d'entre elles est indépendant de tout événement lié à toute autre partie disjointe de la précédente.*
L'indépendance mutuelle est la généralisation naturelle de l'indépendance de deux variables, car on peut montrer que :
des variables discrètes $(X_1,\cdots,X_k,\cdots,X_n)$ *sont mutuellement indépendantes lorsque*

$$\forall x_k \, P(\cap_{k=1}^n X_k = x_k) = \prod_{k=1}^n P(X_k = x_k)\,;$$

des variables continues $(X_1,\cdots,X_k,\cdots,X_n)$ *sont mutuellement indépendantes lorsque*

$$\forall x_k \, (k = 1,\cdots,n) \, p_{X_1,\cdots,X_k,\cdots,X_n}(x_1,\cdots,x_k,\cdots,x_n) = \prod_{k=1}^n p_{X_k}(x_k).$$

L'indépendance mutuelle de n *variables entraîne leur indépendance deux à deux, mais la réciproque n'est pas vraie ; par exemple, les trois variables de* Bernoulli X,Y *et* Z*, toutes de paramètre* $p = \frac{1}{2}$ *et satisfaisant*
$P(X = x,Y = y,Z = z) = \frac{1}{16}$ *si* $x + y + z$ *vaut* 0 *ou* 2 *et*
$P(X = x,Y = y,Z = z) = \frac{3}{16}$ *si* $x + y + z$ *vaut* 1 *ou* 3
sont deux à deux indépendantes mais pas complètement indépendantes puisque $P(X = 1,Y = 1,Z = 1) = \frac{3}{16} \neq \frac{1}{8} = P(X = 1)P(Y = 1,Z = 1)$.

3.5 Convergences et lois-limites

Son caractère fondamentalement imprévisible n'empêche pas le hasard de présenter certaines régularités dont l'exploitation est le fondement de la statistique inférentielle. Il s'agit essentiellement de propriétés asymptotiques de suites de variables $(X_n)_{n\in\mathbb{N}}$.

3.5.1 Convergence en probabilité et loi faible des grands nombres

La suite de variables $(X_n)_{n\in\mathbb{N}}$ *converge en probabilité vers la constante* a *lorsque pour tout* $\epsilon > 0$ *la probabilité que l'écart absolu entre* X_n *et* a *dépasse* ϵ *tend vers* 0 *quand* n *tend vers l'infini :*

$$\lim_{n\to\infty} P(\mid X_n - a \mid \geq \epsilon) = 0$$

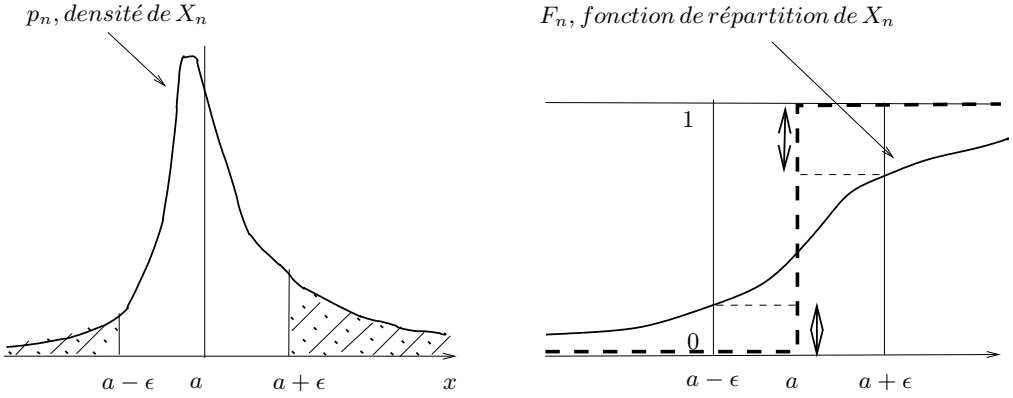

FIG. 8 – *Convergence en probabilité : les aires hachurées sous la densité et la somme des longueurs des segments* $[0, F_n(a-\epsilon)]$ *et* $[F_n(a+\epsilon), 1]$ *tendent vers* 0 *quand* $n \to \infty$

Une version de la loi faible des grands nombres s'énonce ainsi :
Lorsque $(X_n)_{n\in\mathbb{N}}$ *est une suite de variables de même loi, d'espérance* m*, possédant une variance* σ^2 *et deux à deux indépendantes, la suite des variables* $\bar{X}_n = \frac{\sum_{k=1}^n X_k}{n}$ *converge en probabilité vers* m.

La variable $\bar{X}_n$ *est appelée moyenne empirique, car elle vaut* $\bar{x}_n = \frac{\sum_{k=1}^n x_k}{n}$ *lorsque l'échantillon observé est une réalisation* $(x_1, \cdots, x_k, \cdots, x_n)$ *de* $(X_1, \cdots, X_k, \cdots, X_n)$.
On montre facilement que
$E(\bar{X}_n) = m$ *et* $V(\bar{X}_n) = \frac{\sigma^2}{n}$.
Il existe des versions plus générales de la loi faible ; en fait, on n'a pas besoin de supposer que les X_n *suivent la même loi mais seulement que leurs espérances,* m_n*, et variances,* σ_n^2*, satisfont*

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n m_k}{n} = m \text{ et } \lim_{n\to\infty} \frac{\sum_{k=1}^n \sigma_k^2}{n^2} = 0 \ .$$

3.5.2 Convergence presque sûre et loi forte des grands nombres

Un deuxième concept de convergence, dont on peut démontrer qu'il est plus exigeant que le précédent, est celui de convergence presque sûre.

La suite de variables $(X_n)_{n\in\mathbb{N}}$ converge presque sûrement vers la constante a lorsqu'il y a une probabilité 1 que la suite des réalisations des X_n tende vers a :

$$P(\lim_{n\to\infty} X_n = a) = 1$$

La loi forte des grands nombres s'énonce ainsi :
Lorsque $(X_n)_{n\in\mathbb{N}}$ est une suite de variables de même loi, d'espérance m, possédant une variance et mutuellement indépendantes, la suite des variables $\bar{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$ converge presque sûrement vers m.
Autrement dit, les données obtenues ne peuvent pas vous tromper sur la valeur de m ...à condition toutefois que la taille de l'échantillon que vous pouvez recueillir ne soit pas limitée.

Noter que par rapport à la loi faible, on a renforcé l'hypothèse d'indépendance. Comme pour la loi faible, il existe une extension à des variables X_n ne suivant pas forcément la même loi, mais telles que leurs espérances, m_n, et variances, σ_n^2, satisfont
$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} m_k}{n} = m$ *et* $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{\sigma_k^2}{n^2} = 0$.

3.5.3 Convergence en loi, lois-limites et théorème central-limite

Introduisons d'abord le concept de convergence en loi.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables de fonctions de répartition F_n et X, variable de fonction de répartition F. La suite X_n converge en loi vers X lorsque $F_n(x)$ tend vers $F(x)$ en tout point de continuité de F. On écrit alors $X_n \overset{loi}{\to} X$.

Les résultats qui suivent montrent que la loi normale apparait comme une loi-limite (noter que c'est une loi continue, limite dans les deux cas de lois discrètes).

On démontre en effet que si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables binomiales mutuellement indépendantes $\mathcal{B}$(n,p) , où le paramètre p est le même pour toutes les lois, alors la suite des variables centrées réduites associées, $\frac{X_n - E(X_n)}{\sigma_{X_n}} = \frac{X_n - np}{\sqrt{np(1-p)}}$ tend en loi vers la loi normale centrée réduite :

$$\frac{X_n - np}{\sqrt{np(1-p)}} \overset{loi}{\to} \mathcal{N}(0,1).$$

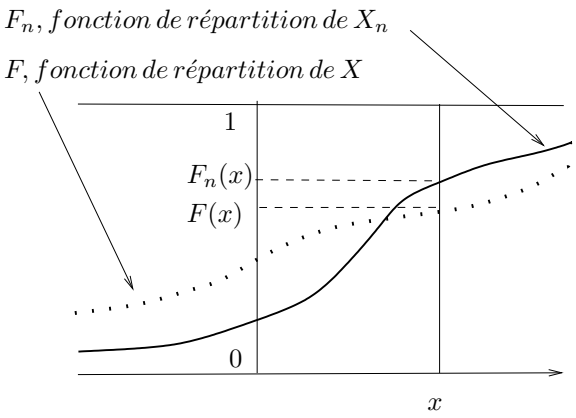

FIG. 9 – *Convergence en loi : en tout x où $F(x)$ est continue, $F_n(x)$ tend vers $F(x)$ quand $n \to \infty$*

De même, si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables mutuellement indépendantes suivant des lois de POISSON *de paramètres λ_n, $\mathcal{P}(\lambda_n)$, où $\lambda_n \to \infty$, la suite des variables centrées réduites associées, $\frac{X_n - E(X_n)}{\sigma_{X_n}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$ tend en loi vers la loi normale centrée réduite :*

$$\frac{X_n - \lambda_n}{\sqrt{\lambda_n}} \overset{loi}{\to} \mathcal{N}(0,1).$$

Nous avons vu que les lois des grands nombres nous disaient quand la moyenne empirique $\bar{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$ tendait vers m, mais ne nous apprenaient rien de précis sur l'allure de la loi de $\bar{X}_n$ pour n grand ; il existe en fait une propriété générale, le théorème central-limite, qui s'énonce ainsi :

Etant donné une suite de variables mutuellement indépendantes, $(X_n)_{n\in\mathbb{N}}$, de même loi d'espérance μ et d'écart-type σ, la suite des moyennes empiriques centrées réduites $\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$ tend en loi vers la loi normale centrée réduite :

$$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \overset{loi}{\to} \mathcal{N}(0,1).$$

3.5.4 Théorème fondamental de la statistique

Supposons que l'échantillon observé $(x_1, \cdots, x_k, \cdots, x_n)$ soit une réalisation $(x_1, \cdots, x_k, \cdots, x_n)$ d'une suite de n variables mutuellement indépendantes, de même loi de fonction de répartition F, $(X_1, \cdots, X_k, \cdots, X_n)$. A tout nombre x, on peut associer le nombre d'observations de valeur inférieures à x,

$$F_n(x) = \mid \{k \,:\, x_k < x\} \mid ;$$

$F_n(x)$ est lui-même aléatoire et l'application $x \mapsto F_n(x)$ est une "fonction aléatoire", appelée fonction de répartition empirique de l'échantillon. La variable

$$\Delta_n = \sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid$$

est elle-même une variable aléatoire.

GLIVENKO *a démontré que*

$$P(\lim_{n \to \infty} \Delta_n = 0) = 1\,,$$

c-à-d que, avec une probabilité 1*, la fonction de répartition empirique d'un échantillon converge uniformément vers la fonction de répartition de la population dont il est tiré, quand la taille de l'échantillon tend vers l'infini.*

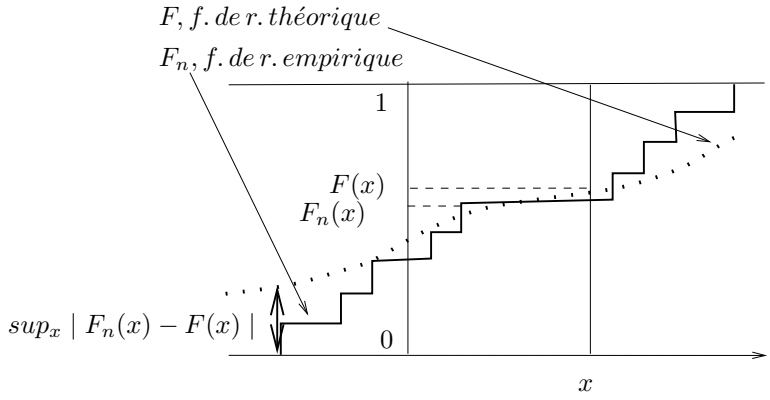

FIG. 10 – *Théorème de* GLIVENKO*: la f. de r. empirique F_n converge uniformément vers tend vers la f. de r. théorique F quand $n \to \infty$*

Ce résultat, connu sous le nom de théorème fondamental de la statistique, est extrêmement important car il nous dit qu'il est presque sûr qu'en recueillant un échantillon suffisamment grand nous obtenions une information aussi précise que nous le souhaiterons sur la fonction de répartition de la population, donc sur sa loi de probabilité et sur toutes ses caractéristiques (la loi des grands nombres nous renseignait seulement sur son espérance mathématique).

Ce résultat est surprenant puisque les fluctuations aléatoires font que l'on aurait bien pu recueillir des données en apparence assez différentes de celles que l'on a obtenues ; le théorème de GLIVENKO *nous dit que peu importe! Tous les échantillons, s'ils sont assez grands, apportent une information exacte et complète!*