Ensembles quantiles et fonction quantile en statistique et probabilités

On rencontre aussi le mot "fractile" au lieu de "quantile" ; il s'agit du même concept, les *quantiles* d'ordre α (α réel de $]0, 1[$) étant bien souvent définis pour des ordres fractionnaires, les plus courants étant : les *médianes* pour $\alpha = 1/2$, les *quartiles* pour $\alpha = k/4$, k=1,2,3, les *quintiles* pour $\alpha = k/5$, k=1,…, 4, les *déciles* pour $\alpha = k/10$, k=1,…,9, les *vingtiles* pour $\alpha = k/20$, k=1,…,19, et les *centiles* pour $\alpha = k/100$, k=1,…,99.

1. Ensembles quantiles et fonction quantile en statistique

Soit X un *caractère quantitatif (*ou *variable réelle),* c'est-à-dire, une application définie sur une population (ou un échantillon) de taille n notée $E = \{1, ..., n\}$ et à valeurs dans $\mathbb{R}$; $(X(1), ..., X(n))$ est le n-uplet des images des n éléments de E par X (appelé dans le secondaire "série statistique à une variable de taille n").

1.1 *Distribution d'effectifs et distribution de fréquences de X*

On note $X(E) = \{x_i \,;\, i = 1,..., r\}$ avec $x_1 < ... < x_r$ l'ensemble des images de E par X, $\{A_1,..., A_r\}$ avec $A_i = X^{-1}(\{x_i\})$, la partition engendrée par X, et $n_i = \text{Card}(A_i)$, effectif de A_i ; alors la *distribution d'effectifs* de X peut-être identifiée à l'ensemble $\{(x_i, n_i)\,;\, i = 1,..., r\}$.

On a $n = \sum_{i=1}^{r} n_i$ et si on pose $f_i = n_i / n$, fréquence de A_i, alors la *distribution de fréquences* de X peut être identifiée à l'ensemble $\{(x_i, f_i)\,;\, i = 1,..., r\}$.

1.2 *Fonction de répartition de X*

La *fonction de répartition* de X est l'application F définie, pour tout réel x par :

$$F(x) = \frac{1}{n}\text{Card}\left(\{k \in E\,;\, X(k) \le x\}\right)$$

Notation : Pour toute partie A de $\mathbb{R}$, on note $[X \in A]$ l'ensemble des éléments de E dont l'image par X appartient à A, c'est-à-dire, l'ensemble $X^{-1}(A)$, image réciproque par X de A.

On a alors, en notant Freq la proportion ou fréquence du sous-ensemble de E considéré :

$$\forall x \in \mathbb{R},\ F(x) = \frac{1}{n}\text{Card}\left([X \le x]\right) = \text{Freq}\left([X \le x]\right)$$

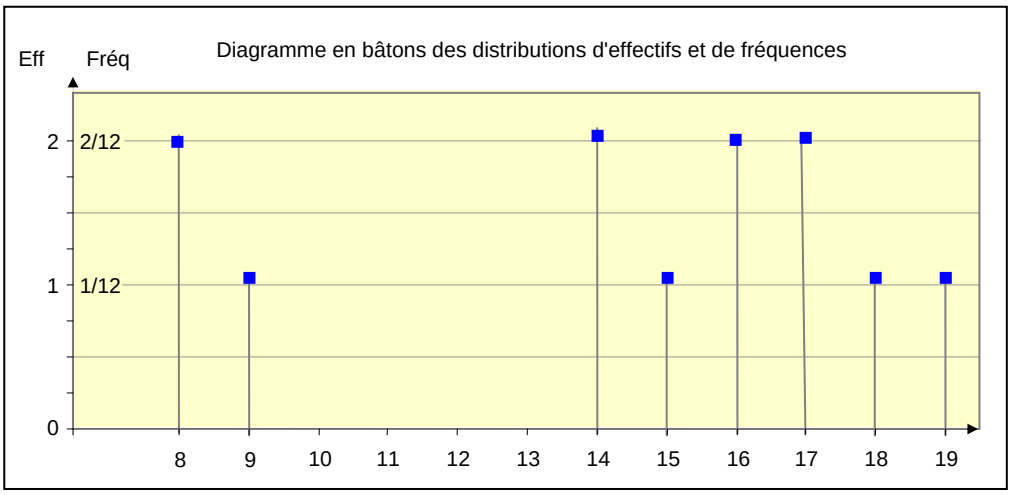
L'application F est en escalier, croissante, à valeurs dans [0, 1], continue sauf pour les éléments de $X(E)$ où elle est seulement continue à droite.

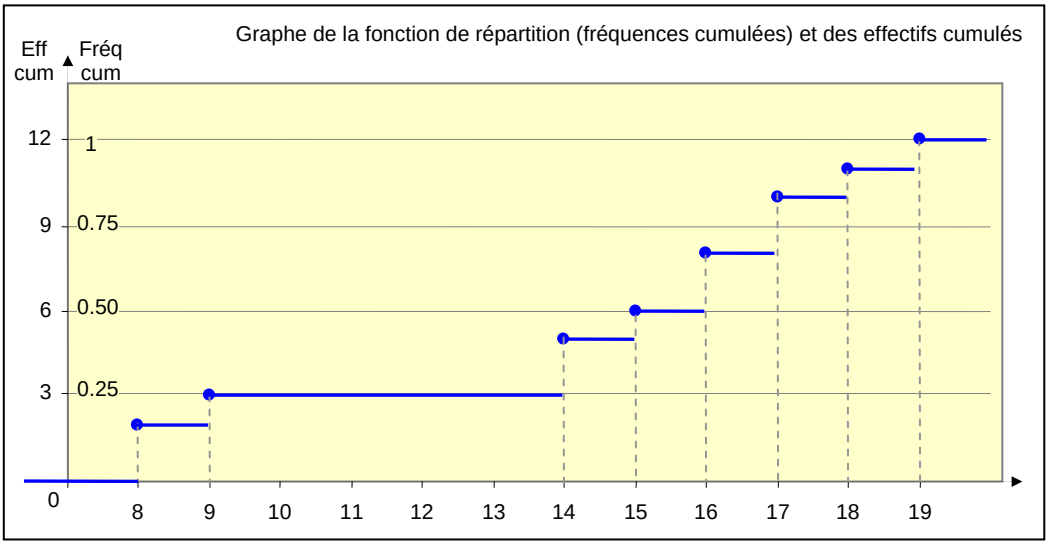
Si on pose, pour $j = 1,..., r$, $N_j = \sum_{i=1}^{j} n_i$ (effectifs cumulés) et $F_j = \sum_{i=1}^{j} f_i$ (fréquences cumulées), alors, en posant $x_{r+1} = +\infty$: $\forall x \in \mathbb{R},\ F(x) = \sum_{j=1}^{r} F_j \,\mathbf{1}_{[x_j, x_{j+1}[}(x)$.

1.3 *Représentations graphiques*

Supposons que l'on ait 12 observations, qui, rangées dans l'ordre croissant, sont :
8, 8, 9, 14, 14, 15, 16, 16, 17, 17, 18, 19

x_i	n_i	f_i	N_i	F_i
8	2	2/12	2	2/12
9	1	1/12	3	3/12
14	2	2/12	5	5/12
15	1	1/12	6	6/12
16	2	2/12	8	8/12
17	2	2/12	10	10/12
18	1	1/12	11	11/12
19	1	1/12	12	1
Total	12	1		





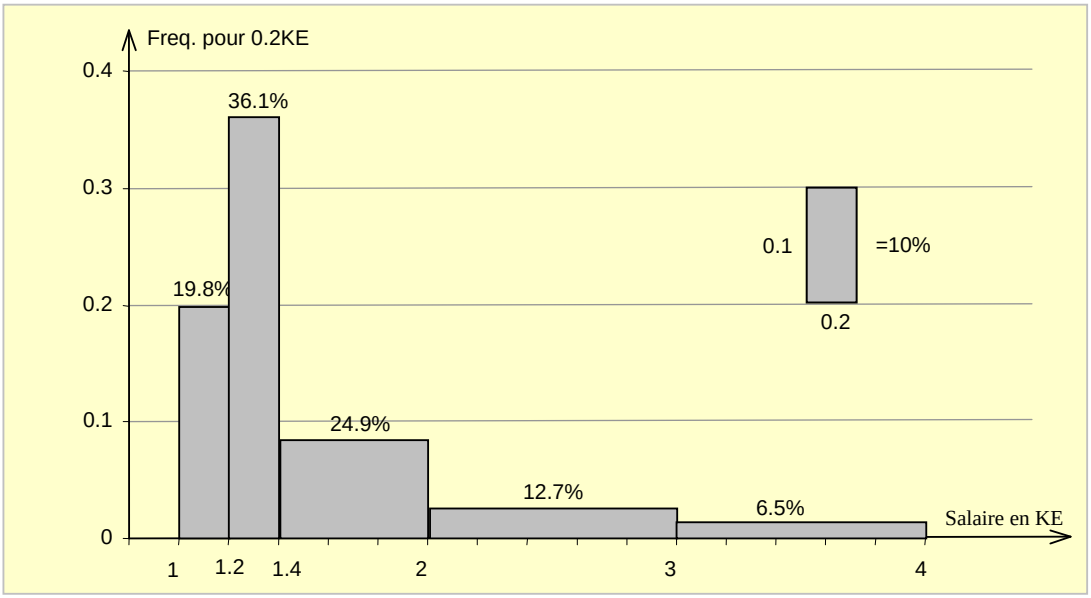
Remarque, cas des valeurs regroupées en classes
Lorsque les observations sont nombreuses et peuvent prendre n'importe quelle valeur d'un intervalle de $\mathbb{R}$, il est d'usage de regrouper les valeurs en classes (intervalles deux à deux disjoints dont la réunion est un intervalle contenant l'ensemble des observations) et de faire l'hypothèse que toutes les observations d'une même classe sont uniformément réparties dans la classe. La représentation graphique de la distribution des effectifs (ou des fréquences) des données groupées est alors appelée *histogramme* : des rectangles, dont les aires sont proportionnelles aux effectifs (et/ou aux fréquences), sont élevés au dessus des classes. La fonction de répartition est alors croissante, continue, affine par morceaux, d'image [0, 1].

Exemple

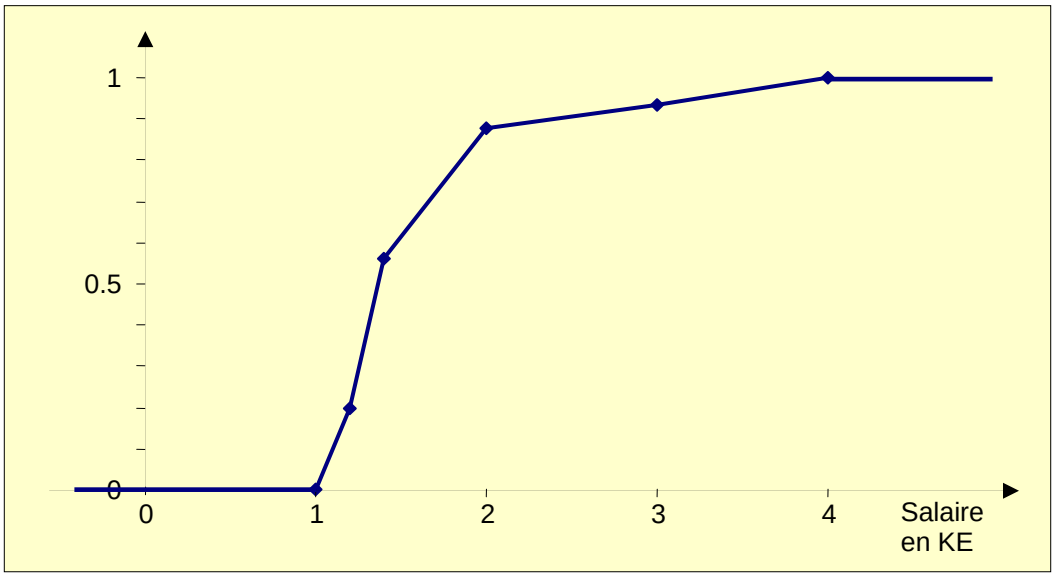
Dans une entreprise, la répartition des salaires mensuels (en milliers d'euros, KE) est donnée dans le tableau suivant :

salaire en milliers d'euros	[1 ;1.2[	[1.2;1.4[	[1.4; 2[	[2;3[	[3;4[
répartition (en %)	19.8	36.1	24.9	12.7	6.5

Histogramme des fréquences de la variable "salaire mensuel"



Représentation graphique de la fonction de répartition (fréquences cumulées) de la variable "salaire mensuel"



1.4 ***Introduction aux ensembles quantiles***

Reprenons l'exemple précédent :

8, 8, 9, 14, 14, 15, 16, 16, 17, 17, 18, 19

La *médiane* partage les observations en deux groupes d'effectifs égaux ; les *quartiles* partagent les observations en quatre groupes d'effectifs égaux (la médiane est donc aussi deuxième quartile).

8, 8, 9, 14, 14, 15, 16, 16, 17, 17, 18, 19

… plus précisément :

Définitions

un réel q_1 est *premier quartile* si au moins 25% des observations sont inférieures ou égales à q_1 et au moins 75% supérieures ou égales à q_1,

un réel q_2 est *deuxième quartile* (appelé aussi *médiane*) si au moins 50% des observations sont inférieures ou égales à q_2 et au moins 50% supérieures ou égales à q_2

un réel q_3 est *troisième quartile* si au moins 75% des observations sont inférieures ou égales à q_3 et au moins 25% supérieures ou égales à q_3

Tout réel de l'intervalle [9, 14] est alors *premier quartile*.
Tout réel de l'intervalle [15, 16] est *deuxième quartile* (ou *médiane*).
L'unique *troisième quartile* est 17.

Propriété : si l'on applique une transformation strictement monotone sur ces données $y = f(x)$, alors, pour k = 1, 2, 3, les ensembles quartiles (intervalles fermés) vérifient : $Q_{y,k} = f\left(Q_{x,k} \right)$ si f est croissante, $Q_{y,k} = f\left(Q_{x,4-k} \right)$ si f est décroissante.

Convention pour l'unicité : on choisit pour unique quartile le centre de l'intervalle.
Avec cette convention, si l'on applique une transformation affine sur les données $y = a\,x + b$, (a réel non nul) alors, pour k = 1, 2, 3, les quartiles vérifient : $q_{y,k} = a\,q_{x,k} + b$ si $a > 0$ et $q_{y,k} = a\,q_{x,4-k} + b$ si $a < 0$.

La convention adoptée au lycée est de prendre le milieu de l'intervalle pour la médiane, la borne inférieure de l'intervalle pour le premier et pour le troisième quartile (cf. programme en annexe et, ci-après, la définition des quantiles à partir de la fonction quantile). On conserve alors la première propriété (transformation affine strictement croissante) mais pas la seconde.

Dans le cas où le nombre d'observations n'est pas un multiple de 4 (i.e. $n = 4q + r$, avec $r = 1, 2, 3$) il existe un unique premier quartile et un unique troisième quartile qui correspondent bien aux définitions données dans les programmes. Il s'agit respectivement de la $(q + 1)^{\text{ème}}$ valeur en partant du début et en partant de la fin de la série des n valeurs rangées dans l'ordre croissant.

1.5 *Ensemble des quantiles d'ordre* α $(\alpha \in]0,1[)$ *et fonction quantile*

Définition : q est *quantile d'ordre* α $(\alpha \in]0,1[)$ si $\mathrm{Fr}(X \le q) \ge \alpha$ et $\mathrm{Fr}(X \ge q) \ge 1-\alpha$,

c'est-à-dire, si $\lim_{x \to q_-} F(x) \le \alpha \le F(q)$; l'*ensemble des quantiles d'ordre* α $(\alpha \in]0,1[)$ est donc

$$Q_\alpha = \left\{ q \in \mathbb{R} \; ; \lim_{x \to q_-} F(x) \le \alpha \le F(q) \right\}.$$

Les premiers, deuxièmes et troisièmes *ensembles quartiles* sont les ensembles quantiles d'ordre 0.25, 0.50 et 0.75, respectivement.
Les premiers, deuxièmes, …, neuvièmes *ensembles déciles* sont les ensembles quantiles d'ordre 0.1, 0.2, …, 0.9 respectivement.
Les *médianes* sont les quantiles d'ordre 0.5 donc aussi les deuxièmes quartiles et les cinquièmes déciles.

L'ensemble Q_α est un intervalle fermé non vide de $\mathbb{R}$; lorsqu'il n'est pas réduit à un singleton, pour avoir unicité, on choisit par convention pour quantile d'ordre α le centre de l'intervalle ; c'est ce qui est proposé au collège pour la médiane.
L'image réciproque par F du singleton $\{\alpha\}$ est :

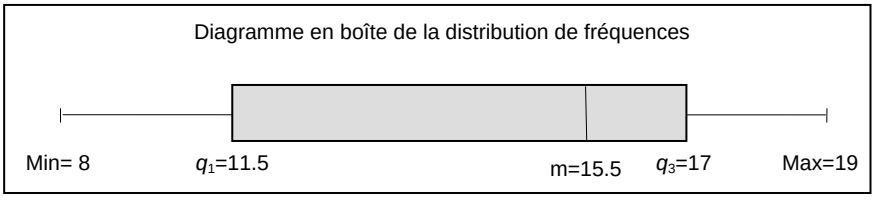
- soit un intervalle $[a, b[$ fermé à gauche, ouvert à droite, auquel cas l'ensemble des quantiles d'ordre α est l'intervalle fermé $[a, b]$ (en effet, on vérifie que b est aussi quantile d'ordre α) ; ce cas n'est possible que si $n\alpha$ est un entier
- soit l'ensemble vide, on a alors un unique quantile d'ordre α égal à $\inf \{ x \in \mathbb{R} \; ; F(x) \ge \alpha \}$; c'est le cas, en particulier, lorsque $n\alpha$ n'est pas un entier

Définition : la *fonction quantile* est définie par :

$$\forall \alpha \in]0,1[, \; q(\alpha) = \inf \{ x \in \mathbb{R} \; ; F(x) \ge \alpha \}.$$

On a : $\forall \alpha \in]0,1[, \; q(\alpha) = \inf (Q_\alpha)$
Le premier quartile et troisième quartile introduits au lycée sont $q(0.25)$ et $q(0.75)$.

Commentaires
1) Les conventions pour avoir unicité sont très nombreuses, aussi, les calculatrices et tableurs ne donnent pas toutes les mêmes valeurs pour les quartiles. Dans la pratique, pour des échantillons de taille importante, les quantiles fournis à partir de la fonction quantile donnent une information suffisante.
2) Le *diagramme en boîte* (ou "boîte et pattes", ou "boîtes et moustaches", "box plot" ou "box and whiskers plot" en anglais) est un résumé graphique de la distribution de fréquences construit à partir des valeurs minimale et maximale de la série et des trois quartiles. Cette présentation permet de comparer visuellement plusieurs distributions de fréquences.
Reprenons l'exemple avec, pour convention d'unicité, le centre des intervalles quartiles.
On a : $\min = 8, \; q_1 = 11.5, \; m = 15.5, \; q_3 = 17, \; \max = 19$ d'où le diagramme en boîte :



Des variantes de ce diagramme sont proposées ; on peut par exemple arrêter les "moustaches" au niveau de d_1 et d_9 plutôt que min et max et indiquer par des étoiles les observations extérieures à l'intervalle $[d_1, d_9]$.

4) La médiane *m* et l'écart inter-quartile $q_3 - q_1$ sont, respectivement, des indices de *position* et de *dispersion* de la distribution de fréquence de la variable considérée. En fait, la donnée dans l'ordre des cinq indices " $\min, q_1, m, q_3, \max$ " donne davantage d'informations et permet de construire le diagramme en boîte. Il est bien commode de noter " q_0, q_1, q_2, q_3, q_4 " ces cinq indices mais seuls q_1, q_2, q_3 sont des quartiles (contrairement aux "définitions" données par certains logiciels).

5) L'*indice de pauvreté* des personnes (ou des ménages) adopté par la Communauté Européenne est 60% du revenu médian : une personne (ou un ménage) dont le revenu est inférieur à 60% du revenu médian de l'ensemble des personnes (ou des ménages) est considérée pauvre. Il s'agit donc d'un indice de pauvreté relatif au mode de vie de l'ensemble de la population ; il ne donne aucune information sur les hauts revenus et sur les inégalités de revenus.

Pour mesurer les inégalités de revenus, on construit des indices à partir des déciles : les neuf déciles $d_1, ..., d_9$ partagent la population de ménages en dix sous-ensembles de fréquences égales 10%, …, 10% des revenus les plus bas aux revenus les plus hauts. Le rapport d_9 / d_1, appelé *rapport inter-décile*, ou le rapport du revenu moyen des 10% les mieux rémunérés sur le revenu moyen des 10% les moins bien rémunérés sont des *indices d'inégalités de revenus*.

Cf. en annexe un graphique, construit à partir de quantiles, présentant l'évolution de 1998 à 2006 des revenus des ménages les plus riches en France.

6) Les quartiles partagent la population en quatre sous-populations d'effectifs égaux. Ces sous-populations sont appelée par abus de langage, 1er quartile, 2ème quartile, 3ème quartile et 4ème quartile. De même pour les autres quantiles usuellement utilisés.

2. Ensembles quantiles et fonction quantile en probabilités

2.1 *Distribution de probabilité*

Soit *X* une variable aléatoire réelle qui peut être :

- soit discrète, de distribution de probabilité $\{(x_i, p_i); i \in I\}$, *I* fini ou dénombrable, on suppose (pour simplifier) que *I* est une section commençante de $\mathbb{N}^*$ ou $\mathbb{N}^*$ tout entier et que les réels x_i sont rangés dans l'ordre strictement croissant (ce qui élimine entre autres les distributions de probabilité sur $\mathbb{Z}$) ; on suppose de plus que le *support* de la distribution est $\{x_i ; i \in I\}$, c'est-à-dire que l'on a $\forall i \in I,\ p_i > 0$; on pourra reprendre les définitions dans le cas où la distribution est définie sur $\{0,1\}, \{0,...,n\}, \mathbb{N}$ ou même $\mathbb{Z}$

- soit continue, admettant une densité de probabilité *f* supposée (pour simplifier) strictement positive sur un intervalle ouvert *A* de $\mathbb{R}$ (*support* de la distribution), continue sauf, éventuellement, aux bornes de *A* ; on pourra reprendre les définitions dans le cas où le support *A* de $\mathbb{R}$ est réunion d'intervalles disjoints.

Pour tout intervalle *J* de $\mathbb{R}$, la probabilité de *J* est :

- $P_X(J) = \sum_{\{i \in I ; x_i \in J\}} p_i$ dans le cas discret,
- $P_X(J) = \int_{-\infty}^{+\infty} f(x) \mathbf{1}_J(x)\, dx$ dans le cas continu.

Dans les deux cas, si X est une variable aléatoire réelle définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, la distribution de probabilité P_X de X est définie pour tout intervalle J de $\mathbb{R}$ par $P_X(J) = P([X \in J])$ où $[X \in J]$ est le sous-ensemble de Ω, élément de $\mathcal{A}$, image réciproque de J par X.

2.2 *Fonction de répartition*

Définition : la *fonction de répartition* de X est l'application F définie par :

$$\forall x \in \mathbb{R},\ F(x) = P([X \leq x])$$

La fonction de répartition de X caractérise la loi de probabilité de X.

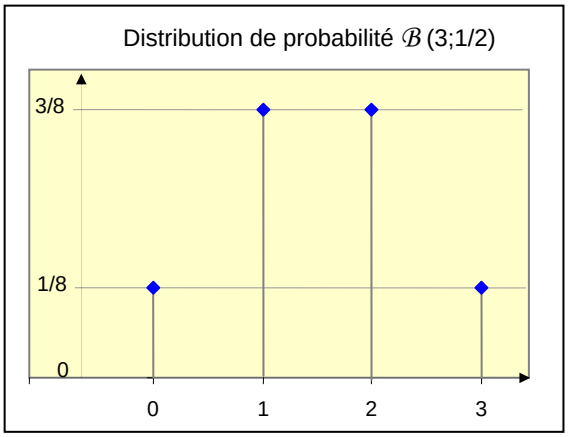
Dans le cas où X est discrète, $\forall x \in \mathbb{R},\ F(x) = P([X \leq x]) = \sum_{\{i \in I ; x_i \leq x\}} p_i$; l'application F est en escalier, croissante, à valeurs dans [0, 1], continue sauf pour les éléments du support $\{x_i ; i \in I\}$ de X où elle est seulement continue à droite ; $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$.

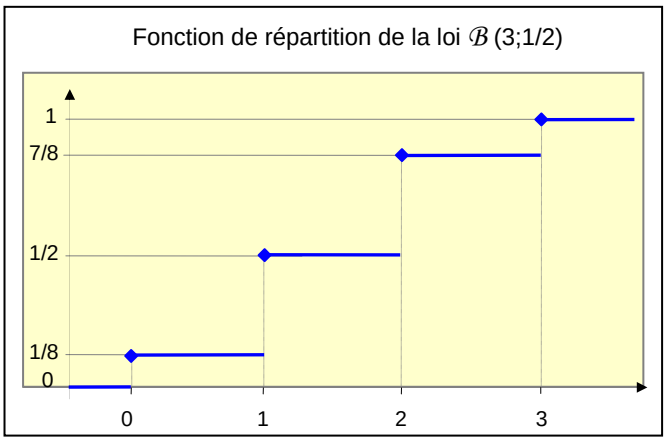
Dans le cas où X est continue, $\forall x \in \mathbb{R},\ F(x) = P([X \leq x]) = \int_{-\infty}^{x} f(x)\,dx$; l'application F est continue, dérivable et de dérivée égale à f (sauf éventuellement aux bornes de A), strictement croissante à valeurs dans [0, 1], $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$. Elle définit une bijection de A dans $]0, 1[$ dont on note abusivement F^{-1} la bijection réciproque.

2.3 *Représentation graphique*

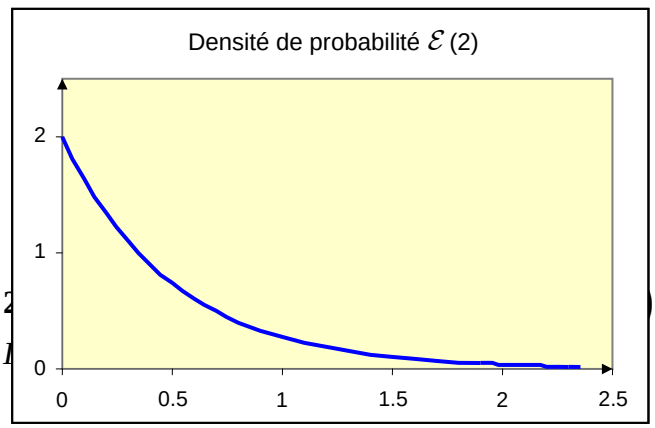
Exemples : (i) Comme variable aléatoire réelle discrète, on peut reprendre l'exemple introductif et considérer que la distribution de fréquences est une distribution de probabilité.

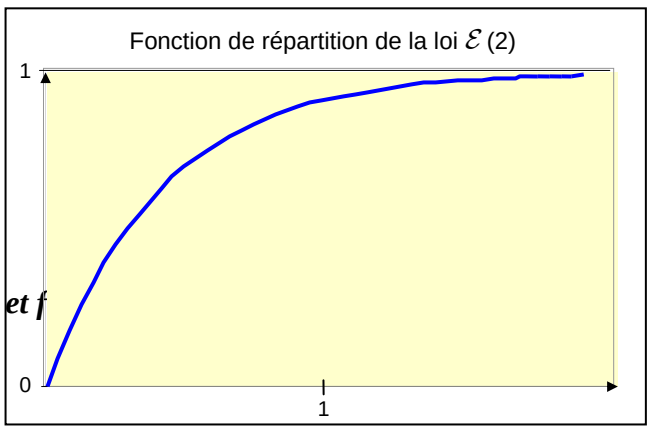
(ii) On peut aussi considérer une loi discrète usuelle, par exemple la loi binomiale $\mathcal{B}(3 ; 1/2)$





(iii) Comme variable aléatoire réelle continue, prenons l'exemple de la loi exponentielle $\mathcal{E}(2)$





Par définition, le réel *m* est médiane de la v.a.r. *X* (ou de sa loi de probabilité) si
$P(X \le m) \ge 0.5$ et $P(X \ge m) \ge 0.5 \Leftrightarrow P(X < m) \le 0.5 \le P(X \le m) \Leftrightarrow \lim_{x \to m_-} F(x) \le 0.5 \le F(m)$

L'ensemble des médianes de *X* est donc : $M = \left\{ m \in \mathbb{R} \ ; \lim_{x \to m_-} F(x) \le 0.5 \le F(m) \right\}$

Généralisation

Par définition, ***l'ensemble des quantiles d'ordre*** α $(\alpha \in]0,1[)$ est :

$$Q_\alpha = \left\{ q \in \mathbb{R} \ ; \lim_{x \to q_-} F(x) \le \alpha \le F(q) \right\}.$$

Les premiers, deuxièmes et troisièmes *ensembles quartiles* sont les ensembles quantiles d'ordre 0.25, 0.50 et 0.75, respectivement.
Les premiers, deuxièmes, …, neuvièmes *ensembles déciles* sont les ensembles quantiles d'ordre 0.1, 0.2, …, 0.9 respectivement.
Les *médianes* sont les quantiles d'ordre 0.5 donc aussi les deuxièmes quartiles et les cinquièmes déciles.

L'ensemble Q_α est un intervalle fermé non vide de $\mathbb{R}$.

Dans le cas où *X* est discrète, l'image réciproque par *F* de $\{\alpha\}$ est :

- soit un intervalle [*a*, *b*[fermé à gauche, ouvert à droite, auquel cas l'ensemble des quantiles d'ordre α est l'intervalle fermé [*a*, *b*] (en effet, on vérifie que *b* est aussi quantile d'ordre α)
- soit l'ensemble vide, on a alors un unique quantile d'ordre α égal à $\inf \{ x \in \mathbb{R} \ ; F(x) \ge \alpha \}$.

Dans le cas où *X* est continue et admet une densité de probabilité *f* continue et strictement positive sur un intervalle ouvert *A* de $\mathbb{R}$, alors l'image réciproque par *F* de $\{\alpha\}$ est le singleton $\left\{ F^{-1}(\alpha) \right\}$.

Définition : La ***fonction quantile*** est définie par :

$$\forall \alpha \in]0,1[, \ q(\alpha) = \inf \{ x \in \mathbb{R} \ ; F(x) \ge \alpha \}.$$

On a : $\forall \alpha \in]0,1[, \ q(\alpha) = \inf (Q_\alpha)$

Dans le cas où *X* est discrète de distribution de probabilité $\{ (x_i, p_i) \ ; i \in I \}$, *I* section commençante de $\mathbb{N}^*$ ou $\mathbb{N}^*$ tout entier et les réels x_i rangés dans l'ordre strictement croissant, alors, en posant $P_0 = 0$ et $\forall i \in I, \ P_i = \sum_{j=1}^{i} p_j$ (probabilités cumulées), on a :

$$\forall \alpha \in]0,1[, \ q(\alpha) = \sum_{i \in I} x_i \mathbf{1}_{[P_{i-1}, P_i[}(\alpha).$$

Dans le cas où *X* est continue et admet une densité de probabilité *f* continue et strictement positive sur un intervalle ouvert *A* de $\mathbb{R}$, alors

$$\forall \alpha \in]0,1[, \ q(\alpha) = F^{-1}(\alpha).$$

C'est la raison pour laquelle la fonction quantile peut être considérée comme une application inverse généralisée de *F*.

2.5 ***Utilisation de la fonction quantile pour la simulation de la loi de probabilité de X***

Proposition
Soit *U* une variable aléatoire réelle de loi uniforme continue sur $]0,1[$, c'est-à-dire, admettant pour densité de probabilité l'indicatrice de $]0,1[$.
Soit *X* une variable aléatoire réelle dont on note q_X la fonction quantile et $Y = q_X(U)$, alors *Y* et *X* ont même distribution de probabilité.

Preuve
Cas où *X* est discrète
L'ensemble des valeurs de *Y* est $\{x_i\ ;\ i \in I\}$ et on a :
$P_Y(\{x_1\}) = P(U \in]0,\ p_1]) = p_1$ et, pour j > 1, $P_Y(\{x_j\}) = P\left(U \in \left]\sum_{i=1}^{j-1} p_i, \sum_{i=1}^{j} p_i\right]\right) = p_j$.
Cas où *X* est continue
On désigne par F_X et F_Y les fonctions de répartition de *X* et de *Y*
Soit $y \in \mathbb{R}$, $F_Y(y) = P(Y \le y) = P\left(F_X^{-1}(U) \le y\right) = P\left(U \le F_X(y)\right) = F_X(y)$.
X et *Y* ont même fonction de répartition donc même distribution de probabilité.

A partir d'un générateur de nombres pseudo-aléatoires (calculatrice ou tableur) on peut générer un échantillon de *n* observations indépendantes $\{u_1,...,u_n\}$ de la loi uniforme continue *U* sur $]0,1[$ et en déduire un échantillon de *n* observations indépendantes $\{q_X(u_1),...,q_X(u_n)\}$ de la loi de probabilité de *X*, ceci dès qu'il est possible d'écrire explicitement la fonction quantile de *X*.

Applications

1) Soit $P = \{(-1, 1/2), (0, 1/4), (2, 1/4)\}$ la loi de probabilité à générer.
Soit *u* une observation de la loi uniforme continue sur]0, 1[obtenue avec un générateur de nombres pseudo-aléatoires.
On pose $x = -1$ si $u \in]0, 0.5[$, $x = 0$ si $u \in [0.5, 0.75[$, $x = 2$ si $u \in [0.75, 1[$.
Alors *x* est une observation de la loi de probabilité *P*.

2) Loi binomiale $\mathcal{B}(3\,;0.25)$
On remarquera que si *U* suit une loi uniforme continue sur $]0,1[$, alors $X = \mathrm{E}(U + 0.25)$, partie entière de *U* + 0.25, suit une loi de Bernoulli de paramètre 0.25 et la somme de 3 v.a.r. indépendantes de même loi de Bernoulli de paramètre 0.25 suit une loi binomiale $\mathcal{B}(3\,;0.25)$.
Aussi, à partir de (u_1, u_2, u_3) observations "indépendantes" de la loi uniforme continue sur $]0,1[$, $x = \mathrm{E}(u_1 + 0.25) + \mathrm{E}(u_2 + 0.25) + \mathrm{E}(u_3 + 0.25)$ est une observation de la loi binomiale $\mathcal{B}(3\,;0.25)$.

3) Loi exponentielle de paramètre 2
La fonction de répartition définit une bijection de $\mathbb{R}_+^*$ sur $]0,1[$: $u = 1 - e^{-2x}$, la fonction quantile n'est autre que la bijection réciproque : $x = -\ln(1-u)/2$.
A toute observation *u* d'une loi uniforme continue sur]0, 1[, $x = -\ln(1-u)/2$ est une observation de la loi de probabilité exponentielle de paramètre 2.
On remarquera que $x = -\ln u/2$ fait aussi bien l'affaire car si *U* suit une loi uniforme sur]0, 1[, il en est de même de $1 - U$ réciproquement.

3) Loi normale *N*(0, 1)
Nous n'avons pas d'écriture explicite de la fonction de répartition (et donc de la fonction quantile) de la loi normale *N*(0, 1) ; c'est la raison pour laquelle ses valeurs sont tabulées ou données par la calculatrice ou l'ordinateur à partir de calculs approchés.
On montre cependant que si *U* et *V* sont des v.a.r. indépendantes en probabilité chacune de loi uniforme continue sur $]0,1[$, alors $X = \sqrt{-2\ln(U)}\cos(2\pi V)$ suit une loi normale centrée réduite (il en est de même de $Y = \sqrt{-2\ln(U)}\sin(2\pi V)$ et les v.a.r. *X* et *Y* sont indépendantes en probabilité).
On peut aussi simuler une loi de probabilité approchée de la loi *N*(0, 1) : par exemple, si $(U_i)_{i=1,\ldots,12}$ sont des v.a.r. indépendantes en probabilité, de même loi uniforme continue sur $]0,1[$, alors $X = \sum_{i=1}^{12} U_i - 6$ suit approximativement une loi *N*(0, 1).

ANNEXE 1

Les quantiles, programme et document d'accompagnement

Programme de 1ère S
Contenu

Diagramme en boîte ; intervalle interquartile. Influence sur l'intervalle interquartile d'une transformation affine des données.

Document d'accompagnement au programme de 1ère S (29/12/00)
Annexes

Médiane (empirique) : on ordonne la série des observations par ordre croissant ; si la série est de taille *2n+1*, la médiane est la valeur du terme de rang *n+1* dans cette série ordonnée ; si la série est de taille *2n*, la médiane est la demi-somme des valeurs des termes de rang *n* et *n+1* dans cette série ordonnée.
La définition de la médiane n'est pas figée : certains logiciels et certains ouvrages définissent la médiane comme étant le second quartile ou le cinquième décile : dans la pratique de la statistique, les différences entre ces définitions sont sans importance, on évitera tout développement la dessus qui ne serait pas une réponse individuelle à une question d'un élève.

Premier quartile (empirique) : c'est le plus petit élément *q* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 25% des données soient inférieures ou égales à *q*.

Troisième quartile (empirique) : c'est le plus petit élément *q'* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 75% des données soient inférieures ou égales à *q'*.
Intervalle interquartile : intervalle dont les extrémités sont le premier et le troisième quartile.
Écart interquartile : différence entre le troisième et le premier quartile.

Premier décile (empirique) : c'est le plus petit élément *d* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 10% des données soient inférieures ou égales à *d.*
Neuvième décile (empirique) : c'est le plus petit élément *d'* des valeurs des termes de la série, ordonnées par ordre croissant, tel qu'au moins 90% des données soient inférieures ou égales à *d'*.
Intervalle interdécile : intervalle dont les extrémités sont le premier et le neuvième décile.

ANNEXE 2

Pyramide des âges, pyramide des salaires

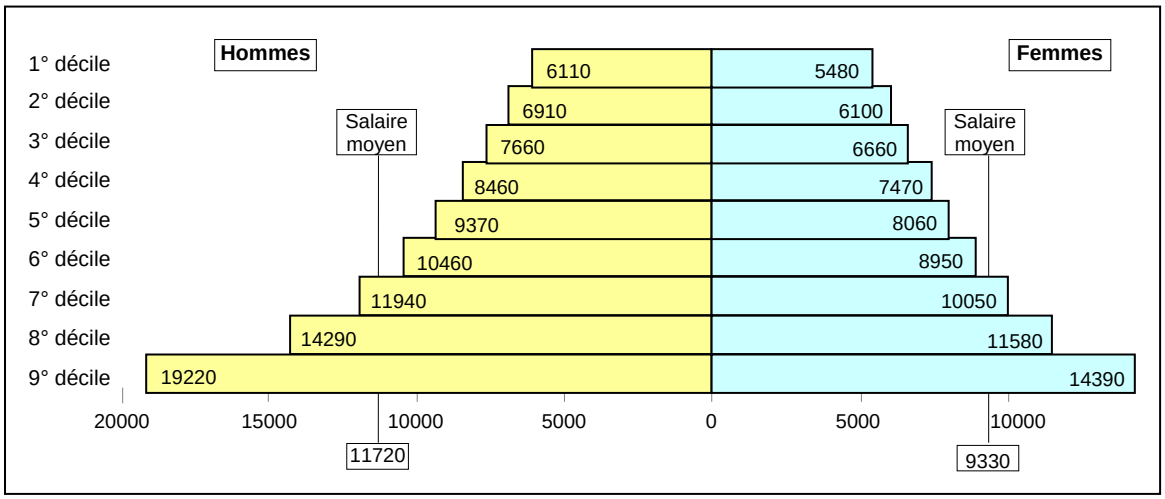
Pyramide des âges
La pyramide des âges d'une population donnée est un graphique très utilisé par les démographes présentant en face à face deux histogrammes, celui des hommes et celui des femmes. On rappelle qu'un histogramme est une représentation de la distribution des effectifs d'une variable réelle (ici l'âge) dont les valeurs sont regroupées en classes (classes d'âge d'amplitude 1 an, 5 ans ou 10 ans). Il faut convenir de regrouper dans une dernière classe toutes les personnes ayant dépassé un certain âge que l'on se fixe comme limite inférieure de la dernière classe.
La pyramide des âges de la population française en 2008 n'a plus la forme d'une pyramide.
Cf. http://www.ined.fr/fr/tout_savoir_population/animations/pyramide_ages/, une animation intitulée "de la pyramide à la toupie" sur le site de l'INED (Institut National d'Etudes Démographiques).

Pyramide des salaires
On peut de la même manière construire une pyramide des salaires, par exemple, la pyramide des salaires des salariés français à temps complet en 1998 (hommes/femmes) dont on nous donne les informations suivantes.
Le salaire moyen est de 9330 F pour les femmes et de 11 720 F pour les hommes.
Les déciles du salaire mensuel en francs des femmes et des hommes sont donnés dans le tableau suivant.

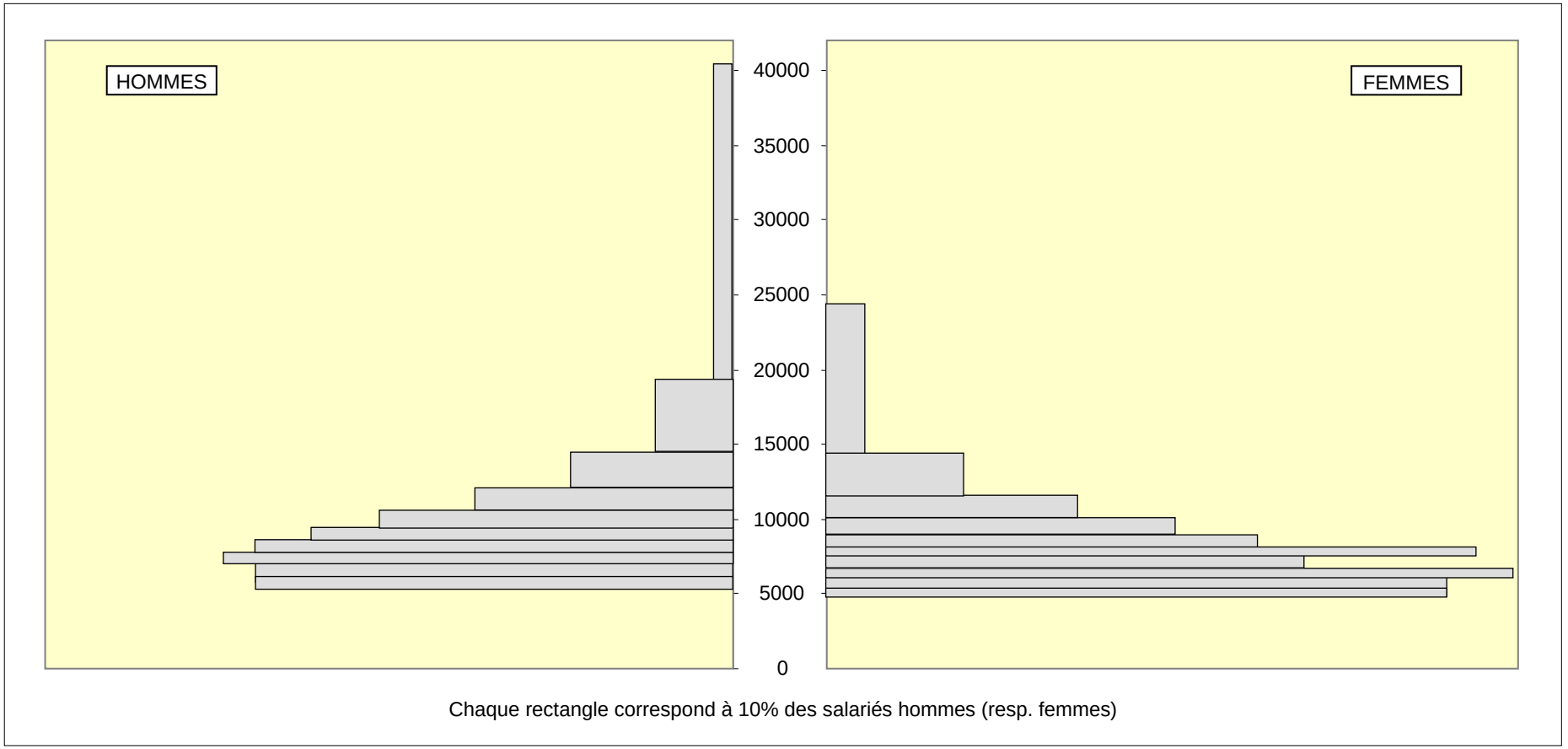
	Hommes	Femmes
d1	6110	5480
d2	6910	6100
d3	7660	6660
d4	8460	7470
d5	9370	8060
d6	10460	8950
d7	11940	10050
d8	14290	11580
d9	19220	14390

On se propose de construire une pyramide constituée, pour chaque histogramme (les hommes à gauche, les femmes à droite), de dix classes, les déciles formant les limites des classes.
Il reste à convenir de la borne inférieure de la première classe et de la borne supérieure de la dernière classe que l'on note respectivement d_0 et d_{10}. Pour la borne inférieure, on conviendra que les rectangles représentant les deux premières classes ont même longueur (pour ne pas dire hauteur puisque, pour une pyramide, l'axe des salaires est vertical). On trouvera 5310 pour les hommes et 4860 pour les femmes.
Pour la borne supérieure, on utilisera l'information concernant la moyenne ; en effet, en notant c_i le centre de la $i^{\text{ème}}$ classe et f_i la fréquence (égale à 10% pour chaque classe) la moyenne vérifie : $\bar{x} = \sum_{i=1}^{10} f_i\, c_i = 0.1 \sum_{i=1}^{10} c_i$ avec $c_i = \frac{1}{2}\left(d_{i-1} + d_i\right)$. On en déduit d_{10}. On trouvera 40250 pour les hommes et 24260 pour les femmes.
Enfin, si on note a_i l'amplitude de la $i^{\text{ème}}$ classe ($a_i = d_i - d_{i-1}$), comme la surface hachurée du rectangle correspondant à cette classe doit représenter 10% de la population, on en déduit que la longueur l_i du rectangle correspondant à la $i^{\text{ème}}$ classe est proportionnelle à $1/a_i$.
La pyramide est donnée page suivante (on n'a pas d'information sur la proportion des hommes et des femmes dans l'ensemble des salariés, la pyramide représente les répartitions dites conditionnelles des salariés hommes et des salariées femmes).

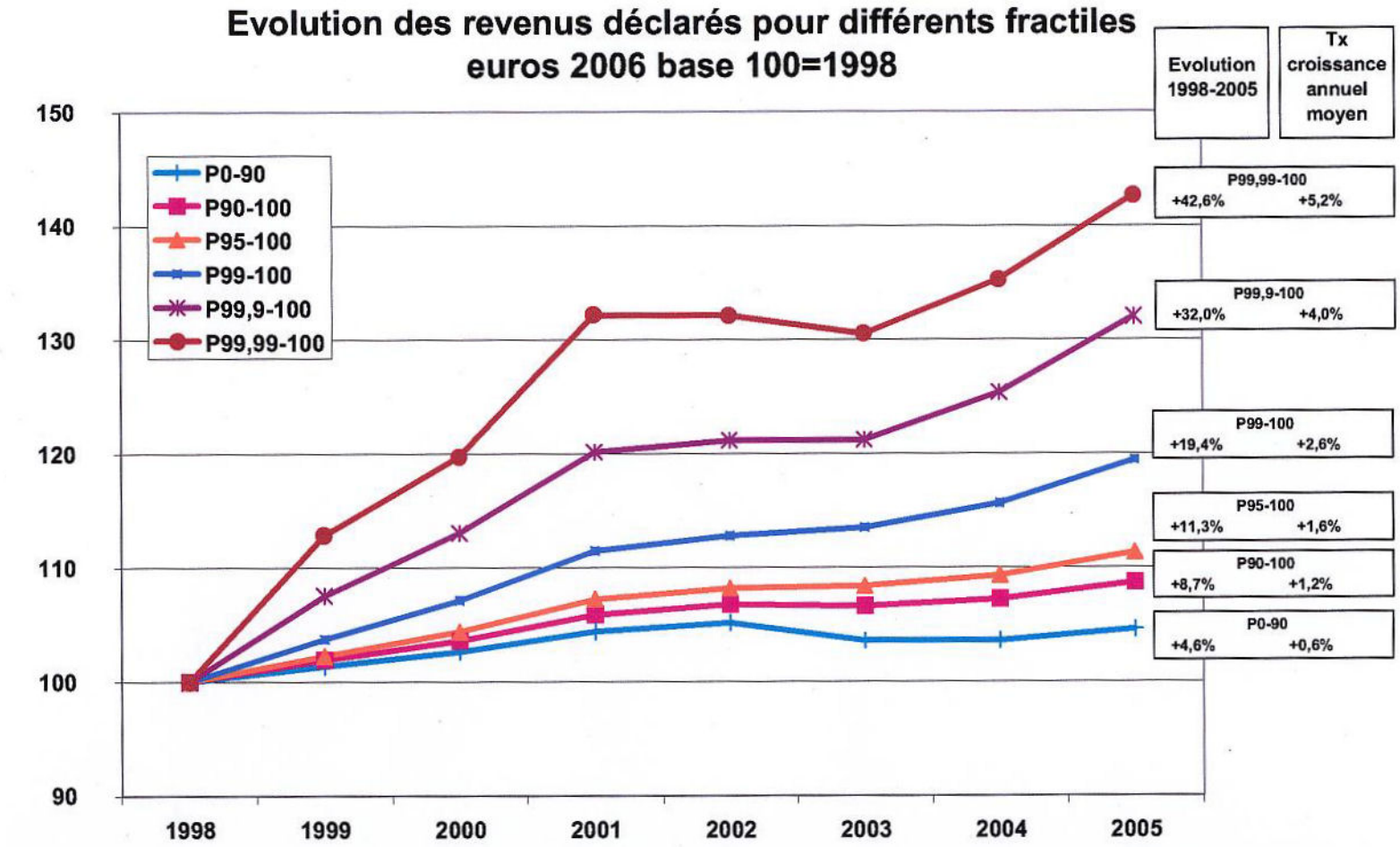
Voici une pyramide, construite avec les mêmes données, rencontrée dans plusieurs manuels du secondaire. Qu'en pensez-vous ?



Distribution des salariés à temps complet en1998 (salaire mensuel en F)



Chaque rectangle correspond à 10% des salariés hommes (resp. femmes)



Note : le fractile P90-100 correspond aux 10% des foyers les plus riches (3,5 millions de foyers sur 35 millions), le fractile P95-100 au 5% des foyers les plus riches, etc. Le fractile P99,99-100 correspond aux 0,01% des foyers les plus riches (3 500 contribuables les plus riches sur 35 millions)

Extrait de Camille LANDAIS "Les hauts revenus en France (1998-2006) Une explosion des inégalités ?
1er juillet 2007 Ecole d'Economie de Paris

1er juillet 2007 Ecole d'Economie de Paris