Statistiques

2. Statistique descriptive monovariée

2.1 Objectifs de la statistique descriptive monovariée

Partons du tableau de données individuelles tel qu'il a été défini au chapitre précédent, avec les individus en ligne et les variables en colonne.

Si on lit un tableau de données individuelles ligne par ligne, on retrouve pour chaque ligne la suite des modalités x_{ij} des différentes variables pour un même individu i, et l'on obtiendra ainsi une sorte de portrait de cet individu i qualifié par ses réponses aux différentes questions de l'enquête. Ceci correspond à une sorte de synthèse de la méthode **monographique**, avec toute la richesse de chaque "portrait", mais avec aussi son impossible généralisation.

Si au contraire on lit ce tableau colonne par colonne, on répond pour chaque variable à la question de la *répartition* des modalités sur la population étudiée. On perd alors de vue tout ce qui faisait l'unité, la cohérence, et la spécificité de chaque individu, visible dans son "portrait-ligne". Mais on gagne quelque chose de nouveau dans cette lecture verticale, qui est l'idée statistique de répartition ou distribution.

La statistique descriptive monovariée a pour objectif de résumer cette information sur la distribution d'une variable dans une population par des résumés graphiques et numériques qui dépendent du type de la variable. Résumer c'est bien sûr perdre de l'information, mais c'est aussi gagner de la pertinence. Il est impossible d'utiliser ou de transmettre à quelqu'un d'autre une information aussi riche que celle qui est dans un fichier de données individuelles sans en faire une synthèse. L'information utile n'est pas l'information brute du fichier, c'est celle qui permet de saisir une structure de la population dans ses grandes lignes, de connaître juste ce qu'il faut pour prendre les bonnes décisions, sans se perdre dans les détails. Il faut donc apprendre à résumer sans trahir et en perdant un minimum d'information.

Les techniques de résumé sont successivement celles de la table de fréquence, de sa représentation graphique, des caractéristiques de position et en particulier de valeurs centrales, celles de dispersion et de concentration. Mais attention, les résumés d'une même caractéristique sont multiples et il convient de bien connaître :

a) leur domaine de validité, et en particulier leur pertinence pour chaque *type de variable* (cf. leçon précédente),

b) leurs propriétés syntaxiques (formules) et sémantiques (significations), qui guident le choix de l'une ou l'autre comme résumé.

2.2 La table de fréquences

Le dépouillement d'une enquête passe en premier lieu par ce qu'on appelle une série de *tris à plat*, ou tris de profondeur 1, faits sur une seule variable à la fois. Un tel tri sur une colonne (une variable) du tableau des données individuelles (x_j) se faisait par marquage manuel (bâtons ou carrés de 5 bâtons) lorsque le dépouillement était manuel. Entre 1890 et 1950, il s'est fait par la mécanographie avec des données saisies sur cartes perforées. Aujourd'hui il se fait par un programme informatique qui lit le champ N°j de tous les enregistrements-individus et incrémente des compteurs différents pour chacune des k modalités de la variable. Le résultat d'un tel tri est une table de fréquence dans laquelle on a perdu l'information de *qui* a telle modalité de la variable pour ne retenir que *combien* ont cette modalité.

a) Si la question était une question ouverte, ce balayage peut conduire à une liste très longue de modalités différentes (certaines différences ne sont parfois que typographiques ou orthographiques) qu'il faut exploiter et reclasser "à la main" ou avec des outils d'analyse textuelle.

b) Si la question est qualitative mais précodée (nominale ou ordinale), ou encore quantitative (cardinale) discrète avec un petit nombre de modalités, le tri conduit à une table de fréquence qui, à la ième modalité x_i , fait correspondre l'effectif n_i de ceux qui ont cette modalité, ou bien encore la fréquence relative f_i définie comme rapport de cette effectif n_i à l'effectif total n : $f_i = n_i / n$. Attention une fréquence relative est toujours un nombre compris entre 0 et 1 qui s'exprime en général sous une forme décimale (par ex. 0,473, que l'on peut *dire* si l'on veut comme une fraction : 47,3/100 ou comme un pourcentage : 47,3%).

La table de fréquence est donc une table à autant de lignes (k) que de modalités et à deux colonnes au minimum : (x_i, n_i) ou (x_i, f_i) puisque l'on sait toujours passer de n_i à f_i. Mais on peut y rajouter des colonnes supplémentaires au fur et à mesure des calculs.

En particulier, on peut rajouter des cumuls, dans le cas où cela a un sens, c'est à dire pour une variable dont les modalités sont ordonnées. On appellera N le cumul des effectifs n et F le cumul des fréquences f :

$$N_i = \sum_{k=1}^{k=i} n_k$$

N_i est le cumul des n_k pour k variant de 1 à i.

C'est le nombre d'individus ayant au plus la modalité x_i

$$F_i = \sum_{k=1}^{k=i} f_k = \sum_{k=1}^{k=i} \frac{n_k}{n} = \frac{N_i}{n} :$$

F_i est le cumul des f_k pour k variant de 1 à i. C'est la fréquence relative d'individus ayant au plus la modalité x_i.

	x_i	**n_i**	**f_i**	**N_i**	**F_i**
	Valeur	Effectif	Fréquence	Effectif cumulé	Fréquence cumulée
i=1	x_1	n_1	f_1	N_1	F_1
i	…	…	…	…	…
i=k	x_k	n_k	f_k	n	1
Total		n	1		

On pourrait aussi définir les effectifs et fréquences cumulées descendants de ceux qui ont une modalité supérieure à x_i. Comme ces nombres sont respectivement les compléments à n et à 1 de N_i et F_i, nous ne le ferons pas.

c) Si la variable est quantitative continue (ou quasi continue avec un grand nombre de modalités) cette table de fréquence aura un grand nombre de lignes (autant que d'individus à la limite) et elle constituera un très mauvais résumé de l'information. On préfère dans ce cas recoder la variables en regroupant les modalités observées par classes de valeurs. On parle alors de variable classée. Ces classes peuvent être prédéterminées (avant enquête) ou définies a posteriori en fonction des besoins de l'analyse. Il faut en choisir le nombre : classes grossières et peu nombreuses, fines et nombreuses. Il faut en choisir la largeur : classes de largeurs égales (ce qui facilite les calculs et les représentations) ou classes d'effectifs égaux (qui donne une représentation plus soucieuse des déséquilibres observés). Dans le cas d'une distribution uniforme (c'est à dire avec le même effectif par unité de largeur), et dans ce cas seulement, ces deux derniers choix se confondent.

Dans le cas d'une variable classée on définira n_i comme l'effectif de la classe de modalités $[x_{i-1}\ ,\ x_i]$ pour laquelle x_i est la borne supérieure (incluse) de la ième classe. On en déduit de la même façon qu'en b) la fréquence f_i de cette même classe ainsi que les effectifs cumulés N_i et les fréquences cumulées F_i.

2.3 Les représentations graphiques

L'objectif d'une représentation graphique est de traduire une distribution de grandeur en une impression visuelle synthétique. Il convient de bien choisir la règle sémantique du graphique – c'est à dire la correspondance entre objet arithmétique et objet géométrique - pour que cette traduction ne soit pas une trahison et procure une image non déformante de la réalité.

2.3.1 Fréquences

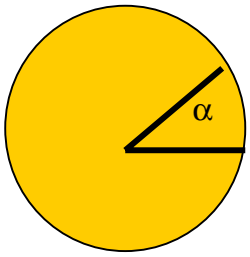
Les effectifs n_i des modalités d'une variable sont représentés par des éléments graphiques qui vont traduire leurs valeurs par des éléments géométriques de taille proportionnelle. La représentation graphique des fréquences relatives f_i sera la même que celle des effectifs n_i puisque ces deux séries de nombres sont elles mêmes proportionnelles entre elles. La représentation graphique dépend du type de l'échelle de mesure utilisée (voir encart).

a) Si l'on a affaire à une variable qualitative nominale, la seule propriété des modalités est de constituer une partition de catégories exclusives et complémentaires. On choisit l'image du **camembert**, ou en anglais de la tarte (**pie**), pour représenter cela, avec la propriété sémantique suivante : chaque effectif est représenté par une "part" c'est à dire un secteur dont l'angle (et par conséquent la surface) est proportionnel à ce nombre :

Effectif $n_i \rightarrow$ angle $\alpha_i = k\, n_i$

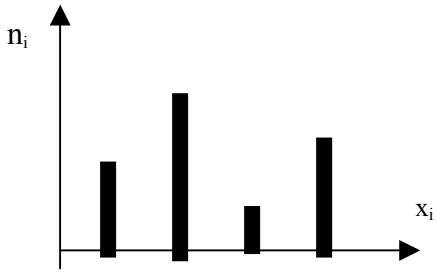
où k est un facteur de proportionnalité défini par :

$$\sum_i \alpha_i = 360° \;\heartsuit\; k\sum_i n_i = 360° \;\heartsuit\; kn = 360° \;\heartsuit\; k = \frac{360°}{n} \;\heartsuit\; \frac{\alpha_i}{360°} = \frac{n_i}{n}$$



L'angle qui représente l'effectif de la modalité x_i est dans la proportion f_i du cercle complet.

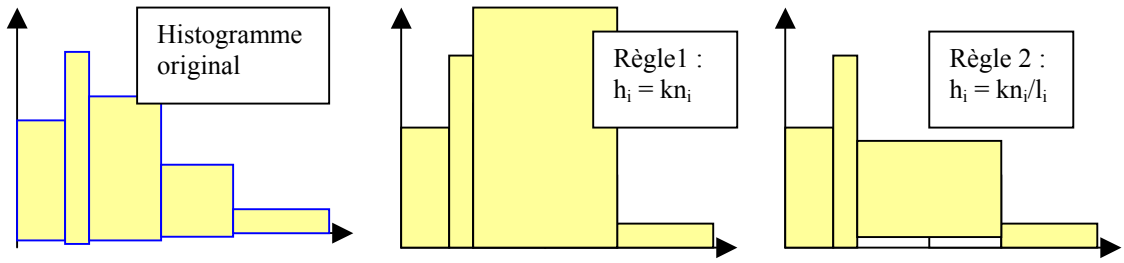
b) Si l'on a affaire à une variable qualitative ordinale il vaudra mieux abandonner cette représentation pour une autre qui sache traduire l'ordre des modalités. On privilégie alors une représentation en "**tuyaux d'orgue**" ou en "bâtons" dans un graphique cartésien à deux axes dont celui des abscisses traduit l'ordre des modalités et l'axe des ordonnées traduit la valeur de n_i ou f_i. **Attention :** le tableur Excel appelle ce graphique à tort un histogramme.



c) Si la distribution représentée est celle d'une variable quantitative discrète, le principe de la représentation est encore celui du diagramme en **bâtons**, mais cette fois-ci l'axe des abscisses traduit plus qu'un ordre : il rend compte par une échelle appropriée d'une suite de valeurs numériques (souvent entières) dont les écarts ont un sens : l'écart entre 2 et 5 doit par exemple être triple de celui qui existe entre 1 et 2. **Attention** le tableur Excel appelle ce graphique à tort un histogramme.

d) Si la variable étudiée est (quasi) continue et que l'on a dû regrouper les modalités en nombre (très grand) infini dans des classes, il faut bien voir que l'on a perdu une partie de l'information du fichier en passant à la table des fréquences par classes. Le fait de représenter les effectifs par des **rectangles** est la conséquence d'une hypothèse implicite de répartition uniforme dans chaque classe que l'on substitue à l'information perdue.

Deux choix sont possibles. Le premier choix (règle 1) consiste à **représenter les effectifs par des rectangles de hauteur** proportionnelle à ces nombres. Une telle règle a cependant l'inconvénient majeur de donner une représentation graphique qui dépend de la façon dont on a fait les classes. Si l'on regroupe deux classes contiguës, leurs effectifs vont s'ajouter et la hauteur du rectangle correspondant de l'histogramme va augmenter de façon arbitraire puisque rien n'a changé dans la distribution des effectifs. Vice versa si l'on affine le découpage en divisant en deux une classe les effectifs et dont les hauteurs des rectangles seront abaissés artificiellement. Cette règle permet toute manipulation des représentations fournies : on pourrait déformer à loisir l'histogramme en jouant sur le découpage en classe.



En **représentant les effectifs par la surface des rectangles** (règle 2), on évite cette sensibilité du graphique au découpage en classe. En effet cette nouvelle règle consiste à prendre :

Surface rectangle = hauteur x largeur = h_i x l_i = kn_i donc $h_i = kn_i / l_i$

Ce qui revient à dire que la hauteur est cette fois-ci proportionnelle à l'effectif par largeur de classe, ce que l'on peut appeler la densité.

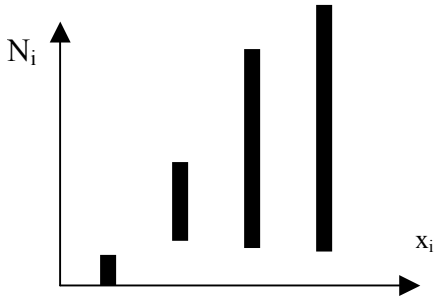
Avec cette nouvelle règle, le regroupement de deux classes conduit à remplacer deux rectangles de hauteur h1 et h2 par un seul rectangle dont la hauteur est moyenne entre les deux autres. La surface totale est la même dans les deux cas. C'est ce qu'on appelle le principe de conservation des aires. (voir aussi la simulation). Notons pour finir que le tableur Excel ne sait pas représenter un vrai histogramme dans le cas de classes de largeurs inégales. Il est nécessaire de recourir à un artifice intermédiaire pour y arriver.

L'histogramme prend aussi le nom de *courbe de densité empirique*. Le polygone des fréquences que l'on voit parfois dessiné en joignant les centres du côté haut des rectangles est le plus souvent sans signification. Aussi est-il préférable de ne pas utiliser cette représentation. Le seul intérêt de celle-ci est dans la courbe de densité continue qu'il peut représenter à la limite, lorsque le nombre des classes tend vers l'infini (et leur largeur vers zéro), comme on le verra en calcul des probabilités.

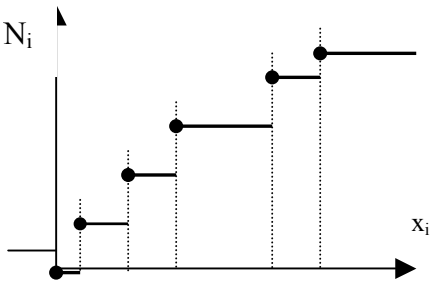
2.3.2 Fréquences cumulées

a)Le graphique des fréquences cumulées n'existe pas si celles-ci n'ont pas de signification, ce qui est le cas d'une variable nominale.

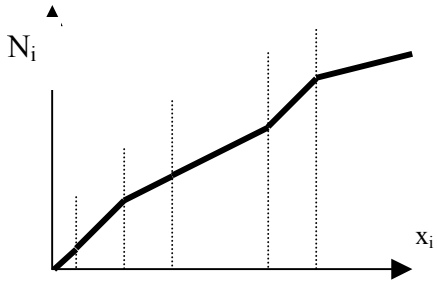
b) Dans le cas d'une variable ordinale, les fréquences cumulées ont un sens même si la variable est qualitative parce que l'on peut répondre à la question combien de personnes ont au plus telle modalité. Ce nombre N_i peut être représenté par un tuyau ou bâton de hauteur N_i. On obtient ainsi un diagramme en bâtons cumulés.



c) Dans le cas d'une variable quantitative discrète il en est de même, mais on peut aussi répondre à la même question pour des valeurs x_i intermédiaires entre deux valeurs observées. Par exemple "combien de personnes ont au plus 1,5 enfants" a pour réponse le même nombre que "combien de personnes ont au plus 1 enfants" : c'est le nombre de celles qui en ont 0 plus le nombre de celles qui en ont 1. Et cette réponse serait la même pour 1,2 ; pour 1,8 et pour 1,9999. Ce qui conduit à tracer un segment de droite horizontal entre 1 et 1,9999 avec un saut à une autre valeur dès que l'on arrive à 2 parce qu'il faudra rajouter aux précédent ceux qui ont deux enfants. Le graphique obtenu est alors celui d'une *courbe en escalier* avec discontinuité à droite à chaque valeur observée.

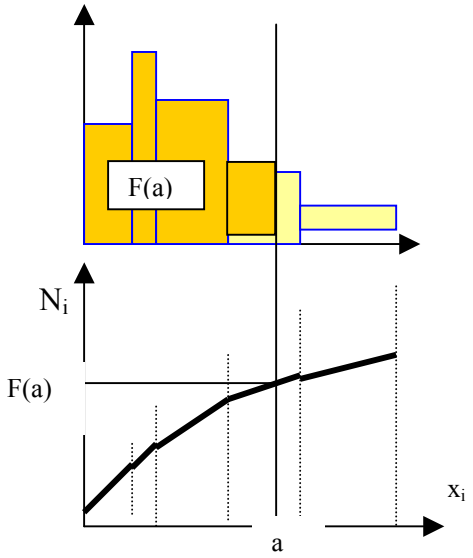


d) Dans le cas de données numériques classées (variable quasi continue), on peut encore faire correspondre aux valeurs x_i des *fins de classe* (et pas des milieux de classe) les effectifs cumulés N_i de tous ceux qui ont au plus cette valeur. Mais aucune courbe ne pourrait être tracé entre les points (x_i , N_i) si l'on ne faisait une hypothèse particulière. En effet cette courbe dépend de la répartition des individus dans la classe. Une concentration d'individus au début de la classe et la courbe des densités sera décroissante tandis que celle des cumuls sera convexe. A l'inverse s'ils se situent en majorité en fin de classe, la courbe de densité sera croissante et la courbe des cumuls sera concave. C'est seulement sous l'hypothèse (courante) d'une répartition uniforme dans la classe que l'on a une densité constante (rectangle) et une courbe de cumul linéaire : on accumule régulièrement des effectifs quand on avance dans la classe. Cette hypothèse permet alors de joindre les points (x_i , N_i) par des segments de droite et la courbe prend la forme d' *un polygone de fréquences cumulées*, ou *courbe de la fonction de répartition empirique.*



e) Lien entre densité et répartition :

Pour une valeur quelconque a la valeur F(a) de la courbe de répartition correspond à la part de la surface de l'histogramme (ou courbe de densité) située à gauche de cette même valeur a.



2.3.3 Valeurs cumulées

2.4 Les caractéristiques de position

L'idée centrale de cette section est celle de résumé numérique. Comment synthétiser une distribution statistique par quelques nombres bien choisis. En particulier comment définir le *milieu* d'une distribution de valeurs. Les astronomes du XVIIIème siècle ont utilisé ce terme (cf. encyclopédie méthodique) dans leurs recherches sur le milieu à prendre entre plusieurs observations, pour estimer le "lieu vrai" d'un corps céleste. Comment, de plusieurs mesures discordantes pour diverses causes d'erreur, peut on déduire une vraie valeur? Une question assez différente s'est posée en sciences sociales dans le cadre de la théorie des moyennes du belge Quetelet au milieu du XIXème siècle : comment décrire une population humaine ? En s'appuyant dit-il sur son centre de gravité, l'homme moyen. Les réponses à la question des astronomes peuvent alors être transposées en sciences sociales.

2.4.1 Mode

Le mode est la valeur la plus fréquente. La valeur "à la mode" en quelque sorte. Celle pour laquelle la densité est maximale. Elle est définie aussi bien pour une variable nominale que pour une variable ordinale ou cardinale (numérique). Il suffit de chercher dans la table de fréquence la fréquence maximale : la modalité correspondante est le mode.

Si la variable est (quasi) continue le mode correspondra au maximum de la densité. En fait l'information disponible si la variable est classée ne permet pas de déterminer une valeur modale mais seulement une *classe modale* : celle pour laquelle la densité (ou l'effectif par unité de largeur de classe n_i / l_i) est maximale.

2.4.2 Moyenne arithmétique

a) La moyenne arithmétique est la valeur fictive de la grandeur étudiée qui caractériserait chaque individu, si l'on répartissait également le total de toutes les valeurs entre tous les individus. Elle résulte donc d'une simple division du total des valeurs de la variable par le nombre d'individus.

Cette moyenne, notée $\bar{x}$ peut s'exprimer de deux façons :

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$: ici la somme porte sur les *n* individus *i* qui varient de 1 à *n*

$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j x_j$: ici la somme porte sur les *k* modalités *j* de la variable, et

en général *k* est bien plus petit que *n*. mais il faut pondérer chaque valeur x_j par le nombre de fois n_j où on l'a observée.

Soit par exemple la suite de 10 nombres : 2, 5, 7, 2, 8, 12, 2, 5, 2, 5

Le premier calcul donne $\bar{x}$ = (2+5+7+2+8+12+2+5+2+7) /10 = 52/10 = 5,2

Le second calcul donne $\bar{x}$ = (4*2 + 2*5 + 2*7 + 1*8 + 1*12) /10 = 52/10 = 5,2

b) Propriétés de la moyenne arithmétique

- Elle n'est définie que si l'addition des modalités a un sens, ce qui est le cas pour des variables numériques (quantitatives discrètes ou continues)
- La moyenne en général "ne tombe pas juste". Ce n'est pas forcément une valeur possible. C'est une fiction.

- Elle est très sensible et donc peu robuste : l'ajout d'un un individu à valeur exceptionnellement faible ou forte modifie de façon importante sa valeur.
- Linéarité 1 : Si j'ajoute une constante *b* à toutes les valeurs x_i la moyenne augment de b : $\overline{x+b} = \overline{x} + b$
- Linéarité 2 : Si je multiplie toutes les valeurs x_i par une constante *a* la moyenne est multipliée par a : $\overline{ax} = a\overline{x}$
- Exhaustivité : Si je connais les moyennes partielles $\overline{x}_A$ et $\overline{x}_B$ de deux groupes disjoints A et B d'effectif n_A et n_B alors je peux en déduire la moyenne générale du groupe $A \cup B$:

$$\overline{x} = \frac{n_A \overline{x}_A + n_B \overline{x}_B}{n}$$

2.4.3 Moyennes généralisées

La moyenne arithmétique n'est pas toujours la moyenne qui a les bonnes propriétés.

a) Soit une somme de 1000 F placée pendant un an à 5% puis un an à 15% puis un an à 25%. Le taux moyen est-il de (5+15+25)/3 = 15%? Non! Il doit vérifier :

1000 (1+0,05)(1+0,15)(1+0,25) = 1000 $(1+t_m)^3$

soit $(1+t_m) = [(1+0,05)(1+0,15)(1+0,25)]^{1/3}$

Ce qui donne t_m = 14,7%.

La formule précédente s'écrit : $\log(1+t_m)$ = (Log 1,05 + Log 1,15 + Log 1,25)/3

soit "*Log de la moyenne = moyenne arithmétique des Log des valeurs*"

La formule précédente définit une *moyenne géométrique*, utile chaque fois que l'on cherche une moyenne de taux ou de grandeurs qui sont en croissance (quasi) exponentielle.

b) Soit un avion parcourant un carré de côté 100 km à la vitesse de 100 km/h sur le premier côté, 200 km/h sur le second, 300 km/h sur le troisième, et 400 km/h sur le quatrième. Sa vitesse moyenne est-elle (100 + 200 + 300 + 400)/4 = 250 km/h ?

Non. Sa vitesse moyenne est définie par le quotient d'une distance totale par une durée totale. La distance totale est de 400 km. La durée totale est :

$$\frac{100}{100} + \frac{100}{200} + \frac{100}{300} + \frac{100}{400} = 1h + 30mn + 20mn + 15mn = 2h05mn = 2{,}083h$$

Sa vitesse moyenne est donc :

$$V_m = \frac{400}{\frac{100}{100} + \frac{100}{200} + \frac{100}{300} + \frac{100}{400}} = \frac{400}{2{,}083} = 192km/h$$

Remarquons que la formule précédente peut s'écrire :

$$\frac{1}{v_m} = \frac{\frac{1}{100}+\frac{1}{200}+\frac{1}{300}+\frac{1}{400}}{4}$$

soit "*inverse de la moyenne = moyenne arithmétique des inverses".*

Ceci définit une *moyenne harmonique*, utile chaque fois qu'on cherche une moyenne de rapports.

c) Quel serait le champ moyen entre 3 champs carrés de côté 2, 3, 5 hectomètres?

Ce n'est pas un champ de côté (2+3+5)/3 = 3,33 hm. Mais un champ dont la surface est la moyenne des surfaces des trois carrés, donc tel que :

$$a^2 = \frac{2^2+3^2+5^2}{3} \text{ ♥ } a = \sqrt{\frac{38}{3}} = 3{,}56$$

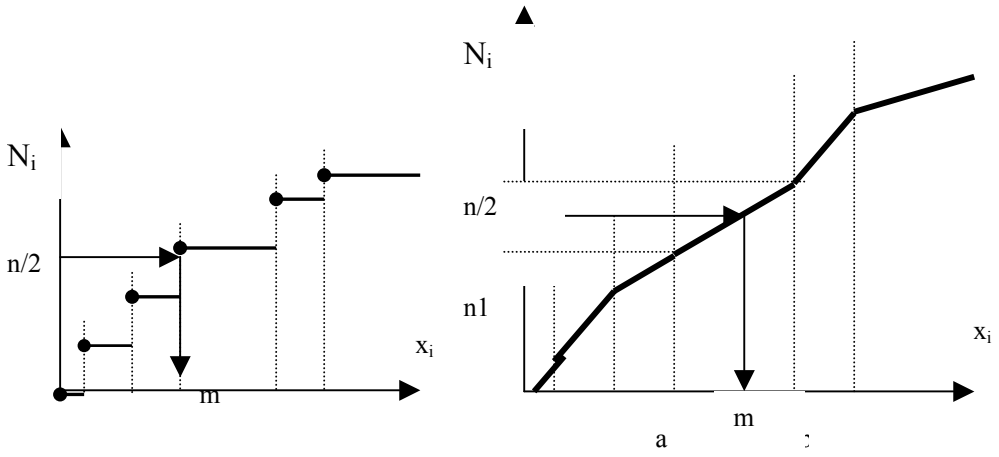
soit "*carré de la moyenne = moyenne arithmétique des carrés".* a est la moyenne quadratique des nombres 2, 3 et 5.

d) En généralisant ces trois exemples, on voit que l'on peut construire des moyennes généralisées sur le principe "*truc de la moyenne = moyenne arithmétique des truc des valeurs*" où "truc" désigne une transformation des données par une fonction monotone du type logarithme, inverse, carré…

2.4.4 Médiane

a) Si l'on ordonne la population étudiée des n individus par valeur croissante de la variable étudiée, avec d'éventuels ex-aequo, l'individu médian divise la population étudiée en deux parties de même effectif. Si n est impair, c'est l'individu numéro (n+1)/2. Si n est pair, on peut hésiter entre l'individu numéro n/2 ou (n/2+1). Pour de grands effectifs c'est peu important.

On appelle médiane *m* la valeur de la variable pour l'individu médian. On peut la définir formellement comme la solution *m* de l'équation *F(m) = 0,5* dans laquelle F est la fonction de répartition. Concrètement on peut l'obtenir sur le graphique de cette fonction (établi à partir des fréquences cumulées) de la façon suivante :



Dans le cas d'une variable ordinale ou quantitative discrète, la médiane est une valeur de la table de fréquences. Dans le cas d'une variable classée (graphique de droite)

on ne peut en principe que situer la médiane dans une classe. L'hypothèse de répartition uniforme dans la classe permet cependant de préciser la valeur de la médiane par un calcul d'interpolation linéaire : si a et b sont les bornes de la classe médiane, et n1 et n2 les effectifs cumulés correspondants, alors on a :

$$\frac{m-a}{b-a} = \frac{n/2 - n1}{n2 - n1} \quad \text{d'où } m = a + \frac{(n/2 - n1)}{(n2 - n1)}(b-a)$$

b) Propriétés de la médiane :

- Elle est définie pour toute variable dont les modalités sont ordonnées.
- Elle est toujours une valeur concrète et observable.
- Elle est plus robuste que la moyenne : l'ajout d'un individu à valeur exceptionnellement faible ou forte ne change quasiment pas la valeur de la médiane.
- Toute transformation monotone (linéaire ou non) sur la variable (qui ne change pas l'ordre des modalités) donne le même individu médian et se répercute donc par la même transformation sur la médiane.
- La médiane d'un groupe ne peut être déterminée à partir des médianes de deux sous-groupes.

2.4.5 Autres fractiles

De la même façon que nous avons défini la médiane, nous pouvons définir :

- les quartiles : il y en a 3, notés Q1, Q2, Q3, qui découpent la population étudiée en quatre parts d'effectif égaux (à 25% de l'effectif total).
- les déciles : il y en a 9, notés D1, D2, …D9, qui découpent la population étudiée en dix parts d'effectif égaux (à 10% de l'effectif total).
- les centiles : il y en a 99, notés C1, C2, …C99, qui découpent la population étudiée en cent parts d'effectif égaux (à 1% de l'effectif total).
- les fractiles (ou quantiles) d'ordre α, notés $x^{(\alpha)}$ tels que $F(x^{(\alpha)}) = \alpha$: c'est la valeur de x telle que la proportion des individus qui ont au plus cette valeur est α.

2.5 Les caractéristiques de dispersion

Le salaire des hommes est non seulement plus élevé en moyenne que celui des femmes, il a plus de chance de s'éloigner beaucoup de la moyenne. Si le premier résumé numérique d'une distribution cherche à déterminer son "milieu", il faut le compléter nécessairement par une mesure de dispersion.

2.5.1 Etendue

L'étendue est tout simplement l'écart (valeur maxi – valeur mini). C'est une mesure très frustre de la dispersion qui a l'avantage d'être facile à calculer mais d'être peu robuste parce que beaucoup trop sensible aux valeurs extrêmes.

2.5.2 Intervalle inter-quartiles, inter-déciles

Pour remédier à ce dernier défaut on lui préfère par exemple l'intervalle inter-déciles D9-D1 qui est l'étendue des 80% de la population obtenue après suppression des valeurs inférieures à D1 et supérieures à D9. Cette mesure est plus robuste mais elle est insensible à des modifications de valeurs internes à cet intervalle.

2.5.3 Ecart absolu moyen. Ecart-type

a) L'idée est donc de prendre "une moyenne des écarts au milieu de la distribution". Remarquons que nous ne pouvons pas prendre "la moyenne arithmétique des écarts à la moyenne" : car celle-ci est toujours nulle car les écarts à la moyenne se compensent exactement:

$$\frac{1}{n}\sum_i (x_i - \bar{x}) = \frac{1}{n}\sum_i x_i - \frac{1}{n}\sum_i \bar{x} = \bar{x} - \bar{x} = 0$$

b) Une solution est de prendre une moyenne des écarts absolus à un certain milieu a. Or on peut montrer que le milieu qui minimise cette moyenne des écarts est la médiane. On peut donc prendre comme mesure *l'écart absolu moyen à la médiane* :

$$EAMM = \frac{1}{n}\sum_i |x_i - m|$$

c) Une autre solution qui a des propriétés mathématiques plus intéressantes est de prendre une *moyenne quadratique des écarts* à un certain milieu. C'est la moyenne arithmétique qui minimise cette moyenne. La moyenne quadratique des écarts à la moyenne ou *écart-type* ("standard deviation" en anglais) s'écrit :

$EQMM = \sigma_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$. La quantité $\sigma_x^{\,2} = \frac{\sum_i (x_i - \bar{x})^2}{n}$ s'appelle *variance*.

Dans la formule précédente, la somme porte sur les n observations de x. Si l'on veut regrouper ces n observations par valeurs, on écrira :

$\sigma_x^2 = \frac{\sum_j n_j (x_j - \bar{x})^2}{n}$ dans laquelle n_j représente l'effectif de la modalité x_j.

Enfin, si l'on fait les calculs à la main ou à la calculette, on préférera la formule suivante obtenue en développant $\left(x_j - \bar{x}\right)^2$:

$\sigma_x^2 = \frac{\sum_j n_j x_j^2}{n} - \bar{x}^2$: *moyenne des carrés moins carré de la moyenne*.

e) Propriétés de la variance et de l'écart-type :

- L'écart-type s'exprime dans la même unité que la variable et s'interprète comme une moyenne des fluctuations autour de la moyenne. La variance n'a pas d'interprétation concrète.
- L'écart-type d'une constante est zéro.
- Translation : Si l'on ajoute b à toutes les valeurs de x, variance et écart-type ne changent pas : $\sigma_{x+b} = \sigma_x$
- Homothétie : Si l'on multiplie toutes les valeurs de x par *a*, l'écart-type est multiplié par a (et la variance par a^2) : $\sigma_{ax} = a\sigma_x$
- Si l'on a deux groupes A et B de données, la variance totale de x se décompose en *variance intra-classes* et *variance inter-classes* : la première est la somme des variances à l'intérieur de chaque groupe autour de la moyenne de chaque groupe. La seconde est la variance des moyennes de groupe par rapport à la moyenne générale.

f) Coefficient de variation.

Si l'on veut comparer les dispersions dans deux populations (par exemple homme et femme) on est gêné par la propriété d'homothétie : si les salaires des hommes sont 1,5 fois ceux des femmes à poste égal, alors la moyenne et l'écart type seront pour les hommes 1,5 fois ceux des femmes. Pour remédier à cet effet d'échelle, on cherche ce que serait les dispersions si les moyennes étaient les mêmes : on se ramène à la même moyenne en comparant les rapports écart-type / moyenne des deux populations. C'est ce rapport que l'on appelle coefficient de variation :

$CV = \frac{\sigma_x}{\bar{x}}$

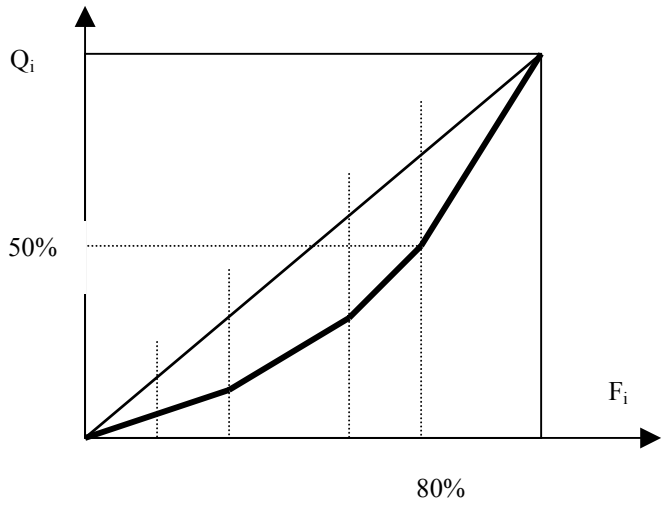
2.6 Concentration

La table de fréquence d'un variable continue peut être complétée par le calcul des valeurs cumulées de la façon suivante :

	$\overline{x}_i$	$\mathbf{n_i}$	$\mathbf{f_i}$	$\mathbf{N_i}$	$\mathbf{F_i = \Sigma f_i}$	$v_i = n_i \overline{x}_i$	$q_i = f_i \overline{x}_i$	$\mathbf{Q_i = \Sigma q_i / \overline{x}}$
	Valeur moyenne de la classe i	Effectif de la classe i	Fréquence relative	Effectif cumulé	Fréquence cumulée	Valeur de la classe i	Valeur par individu	Valeur relative cumulée
TOTAL		n	1			Val. Tot.	Moy. $\overline{x}$	

Les quantités $v_i = n_i \overline{x}_i$ représentent à chaque ligne de la table les valeurs totales de la classe i. Si par exemple x est un salaire, cette quantité est la masse salariale de la classe i. Leur somme est la valeur totale sur la population, dans l'exemple la masse salariale totale. q_i est obtenue en divisant v_i par n; son total est la moyenne arithmétique. Enfin la quantité $\mathbf{Q_i}$ est un cumul des $\mathbf{q_i}$ rapportés à leur somme : elle donne à la ligne i la part de la valeur totale attribuable à la classe i . Il est alors intéressant de porter sur un graphique pour chaque classe ces parts Q_i de la valeur totale cumulées par les classes les plus pauvres (jusqu'à la ième) en regard des parts de la population F_i qu'elles représentent.

En joignant ces points par des segments de droite (conformes à l'hypothèse de répartition uniforme dans chaque classe) on obtient une courbe polygonale dite *courbe de concentration de Lorentz*.



Cette courbe permet par exemple de dire que 80% des plus pauvres n'ont que 50% de la masse salariale, ou encore que 20% des plus riches ont 50% de la masse salariale.

Cette courbe occupe une position intermédiaire entre la diagonale, correspondant à un égalitarisme total, et l'axe horizontal avec saut final au point (100%, 100%) qui correspondrait à la concentration maximale des biens entre les mains d'un seul individu.

Le statisticien italien Gini a proposé sur cette base le calcul d'un coefficient de concentration dit de Gini égal au rapport des surfaces comprises entre courbe et diagonale dans le cas observé et dans le cas d'une concentration maximale; soit en remarquant que la surface du triangle OAB est ½ :

G = 2 fois (surface comprise entre la coure de Lorentz et la diagonale)

Ainsi G varie de 0 à 1 quand la courbe passe de la diagonale à l'axe horizontal.

La valeur de x telle que Q(x) = 50% s'appelle la médiale. Puisque la médiale est toujours supérieure à la médiane (et égale à la médiane dans le seul cas limite de l'égalitarisme), on peut aussi prendre pour mesure de la concentration l'écart entre médiale et médiane.

Quelques points importants de statistique descriptive monovariée

Histogramme : Nom réservé au graphique représentant des données classées. Le principe dans ce cas est que l'on représente les fréquences (absolues ou relatives) par des *surfaces* de rectangles. Dit autrement, *les hauteurs sont proportionnelles aux effectifs par largeur de classe.*

Polygone des fréquences cumulées : Les points ont pour abscisse les fins de classe et pour ordonnées les fréquences cumulées en fin de classe. Ces points sont liés par des segments de droite, parce que l'on fait une hypothèse de répartition uniforme dans chaque classe. Pour n'importe quelle valeur *a* de la variable, la fréquence cumulée F(a) représente la surface de l'histogramme à gauche de *a*.

Courbe de concentration : Les points du graphique ont pour abscisse **F(s)** = part cumulée des individus jusqu'à x = s **,** et pour ordonnée la quantité **Q(x)** = $\sum_{i=1}^{s} f_i x_i / \bar{x}$ qui est la part du cumul total de la variable (revenu, surfaces…) pour ces individus.

Quantiles ou fractiles : ils donnent lieu à une modalité (ou valeur) *observée* dans le cas de variables ordinales ou cardinales discrètes. Dans le cas de variables continues classées et *dans ce cas seulement*, on peut en avoir une estimation par interpolation linéaire : si F(a) et F(b) sont les fréquences cumulées des bornes de la classe [a b] entre lesquelles se trouve le quantile cherché Q_α (tel que $F(Q_\alpha) = \alpha$) , alors $(Q_\alpha - a)/(b-a) = (\alpha - F(a))/(F(b)-F(a))$

Moyenne arithmétique : moy (x) **=** $\bar{x} = \frac{\sum_i n_i x_i}{n} = \sum_i f_i x_i$

Propriétés : moy(ax) = a moy(x)

Moy (x + y) = moy (x) + moy (y)

Moyenne générale = moyenne pondérée des moyennes de classes

Moyennes généralisées : f(Moy) = moy (f(x))

. f = logarithme : moyenne géométrique

. f = inverse : moyenne harmonique

. f = carré : moyenne quadratique

Variance = var (x) = $\sigma^2 = \frac{\sum_i n_i (x_i - \bar{x})^2}{n} = \sum_i f_i (x_i - \bar{x})^2 = \sum_i f_i x_i^2 - \bar{x}^2$

Propriétés : Var(ax) = a² var(x)

Var (x + y) = var(x) + var(y) + 2 cov (x,y)

Coefficient de variation : σ_x / moy(x)

Statistique descriptive monovariée

	Variable nominale	Variable ordinale	Variable quantitative discrète	Variable quantitative continue
Table des fréquences absolues (effectifs) n_i et relatives f_i pour chaque x_i	xi ni fi A 15 0,23 B 24 0,37 C 14 0,22 D 7 0,11 E 5 0,08 Total 65 1,00	xi ni fi Fi A 15 0,23 0,23 B 24 0,37 0,60 C 14 0,22 0,82 D 7 0,11 0,92 E 5 0,08 1,00 Total 65 1,00	xi ni fi Fi 1 15 0,23 0,23 2 24 0,37 0,60 3 14 0,22 0,82 4 7 0,11 0,92 5 5 0,08 1,00 Total 65 1,00	ai bi xi ni fi Fi 0 15 7,5 15 0,23 0,23 15 25 20 24 0,37 0,60 25 35 30 14 0,22 0,82 25 45 35 7 0,11 0,92 45 55 50 5 0,08 1,00 Total 65 1,00
Graphique des fréquences absolues ou relatives (x_i , n_i) ou $(x_i ; f_i)$	Diagramme en secteurs circulaires	Diagramme en tuyaux d'orgue	Diagramme en bâtons	Histogramme
Graphique des fréquences cumulées $(x_i ; F_i)$	Non défini	Diagramme en tuyaux d'orgue	Courbe en escalier	Polygone des fréquences cumulées
Mode valeur la plus fréquente	Mode = B	Mode = B	Mode = 2	Classe modale : 15-25
Quantiles c_α : plus petit x_i tel que $F_i \geq \alpha$	Non définis	$Q_1 = c_{25\%}$ = B Me= $c_{50\%}$ = B $Q_3 = c_{75\%}$ = C	Q_1 = 2 Me = 2 Q_3 = 3	Q_1 = 15,5 Me = 22,3 Q_3 = 31,8
Moyenne $\bar{x} = \sum f_i x_i$	Non définie	Non définie	$\bar{x}$ = 2,43	$\bar{x}$ = 23,7
Ecart-type s $\sqrt{\sum f_i (x_i - \bar{x})^2}$ $\sqrt{\sum f_i x_i^2 - \bar{x}^2}$	Non défini	Non défini	s = 1,18	s = 12,5