Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

Cours 9 – Une variable numérique : distribution et répartition

Lorsqu'une variable est qualitative et l'autre numérique, il est courant que la première identifie des sous-populations (sexe, catégories socio-économiques, sous-populations géographiques, groupes de référence, ...) sur lesquelles on mesure l'autre ; l'étude de la liaison des deux variables revient alors à expliquer par la première, `la variable explicative`, les variations de la seconde, `la variable expliquée`, en comparant les distributions sur les sous-populations. Par la suite X sera la variable qualitative explicative et Y la variable quantitative expliquée.

Exemples

1 - Nombre d'enfants de 0 à 6 ans et structures familiales en 2005 (source Insee) :

Nbre enfants [0-6]	1	2	3	4	5 et plus	Total
F. monoparentales	518	413	92	11	1	1035
Couples	3269	2174	963	120	10	6536
Total	3787	2587	1055	131	11	7571

2 - 400 familles américaines sont classées par revenu en milliers de dollars et par région (adapté de T. et R. Winnicott, 1991).

Région / Revenu	0-4	5-9	10-14	15 et plus	Total
Sud	28	42	30	24	124
Nord	44	78	78	76	276
Total	72	120	108	100	400

3 - Données cliniques mesurées sur 3 groupes de patients :

Groupe 1	65	72	85	93	104	110
Groupe 2	45	51	58	67	79	
Groupe 3	93	98	116	121	123	

Variable numérique

1 **Variable discrète.** Les valeurs observables sont généralement des nombres entiers obtenus par dénombrement ; la variable "nombre d'enfants" de l'exemple 1 est discrète.

Les modalités de la distribution sont les valeurs observables ou éventuellement un regroupement en classes de valeurs observables successives : dans la 5ème modalité de la variable "nombre d'enfants" on a regroupé toutes les valeurs observables au moins égales à 5 (supérieures à 4).

La représentation graphique de la distribution est un `diagramme en bâtons` :

1. On trace un axe gradué horizontal sur lequel on place les valeurs observables.
2. au-dessus de chaque valeur observée on place un bâton vertical de hauteur égale à son effectif ou à sa fréquence.

2 **Variable continue.** Les valeurs observables sont des nombres décimaux lus sur un instrument de mesure réel ou imaginaire ; la variable "revenu" de l'exemple 2 est continue.

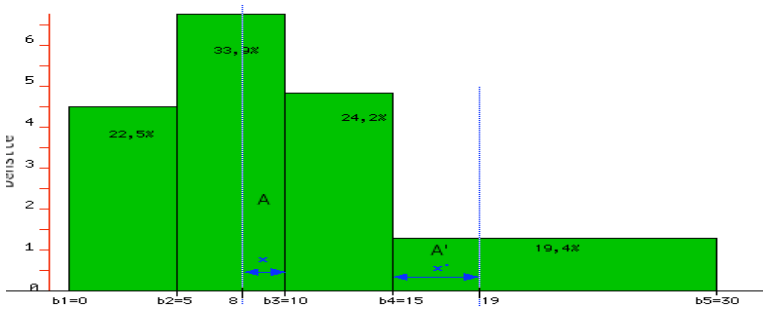
Les modalités de la distribution sont des intervalles contigus de valeurs, dont les amplitudes peuvent être inégales.

La représentation graphique est un `histogramme` :

1. On trace un axe gradué horizontal sur lequel on place les bornes des modalités-intervalles.

2. au-dessus de chaque modalité on trace un rectangle de hauteur égale à sa densité de fréquence (fréquence par unité d'amplitude=$\frac{f_j}{a_j}$).

Exemple : histogramme de la distribution du revenu dans le sud (elle est détaillée au paragraphe 7)



3 Surface et proportion. La surface de chaque rectangle est égale à la fréquence de la modalité : surface = largeur*hauteur = amplitude*densité = (amplitude*fréquence)/amplitude = fréquence ; la surface totale de l'histogramme est donc égale à 1, ou 100 si on exprime les fréquences en pourcentage. Dans ce cas, et avec un peu d'imagination, on peut voir chaque rectangle occupé par les f_j individus appartenant à la modalité, chacun disposant de la même unité de surface.

Si on trace deux verticales coupant l'axe des valeurs en v et v' (par exemple 8 et 19 sur la figure précédente), la surface de l'histogramme située entre les deux verticales représente la proportion des individus de l'échantillon dont la mesure est comprise entre v et v' ; ou plus exactement une approximation de cette proportion : en effet, représenter une modalité par un rectangle équivaut à considérer la densité constante dans cette modalité, ou encore à supposer que les individus se répartissent uniformément dans l'intervalle ; cette hypothèse est très commode, mais elle n'est généralement pas réaliste, la seule information sûre étant la surface et non la forme de la représentation.

Si les deux valeurs sont des bornes, la proportion des individus dont la mesure est comprise entre v et v' (notons la $Prop(v \leq Y \leq v')$) se calcule à partir de la distribution en fréquence ; par exemple $Prop(5 \leq Y_{sud} \leq 15) = 33{,}9 + 24{,}2 = 58{,}1\%$

Si l'une au moins de ces valeurs n'est pas une borne, le tableau de contingence ne permet pas de déterminer la proportion avec exactitude, et on peut seulement en avoir une approximation ; évaluons par exemple $Prop(8 \leq Y_{sud} \leq 10)$, la proportion des individus du sud dont la mesure est comprise entre 8 et 10, représentée par la surface A de la figure précédente : A est la surface dont la largeur x est égale à $10 - 8 = 2$ et dont la hauteur est la densité de la modalité, égale au rapport $\frac{33{,}9}{5}$: $A \approx 2 * \frac{33{,}9}{5} = 13{,}6$. D'une manière analogue, $Prop(8 \leq Y_{sud} \leq 19)$ est la surface comprise entre les verticales passant par 8 et 19 : $Prop(8 \leq Y_{sud} \leq 19) \approx A + 24{,}2 + A'$, et comme $A' = x' * \frac{19{,}4}{15} = 5{,}2\%$, où $x' = 19 - 15$, alors $Prop(8 \leq Y_{sud} \leq 19) \approx 43\%$.

4 Comparaison des distributions. Lorsque la variable numérique est discrète, la comparaison des distributions conditionnelles par leur représentation simultanée en bâtons sur un même graphique peut être suggestive ; si elle est continue, la représentation simultanée des histogrammes est difficile à lire, et il est préférable de les comparer par d'autres moyens, par exemple en comparant les fonctions de répartition.

Fonction de répartition

5 déf La fonction de répartition F d'une distribution de X est une fonction mathématique qui prend en entrée n'importe quel nombre q et retourne comme valeur notée $F(q)$ la proportion des individus de l'échantillon dont la mesure par X est inférieure ou égale à q ; $F(q)$ est donc un

nombre compris entre 0 et 1, ou entre 0 et 100 si on l'exprime en pourcentage ; par "n'importe quel nombre" il faut entendre un nombre entier ou décimal, positif ou négatif.

On distingue trois cas :

q est inférieur à la première modalité : comme aucun individu n'a de mesure inférieure à q, F retourne 0 ;

q est supérieur à la dernière modalité : comme tous les individus ont une mesure inférieure à q, F retourne 1

pour les variables intermédiaires : le mode d'évaluation de $F(q)$ diffère selon qu'il s'agit d'une variable discrète ou continue.

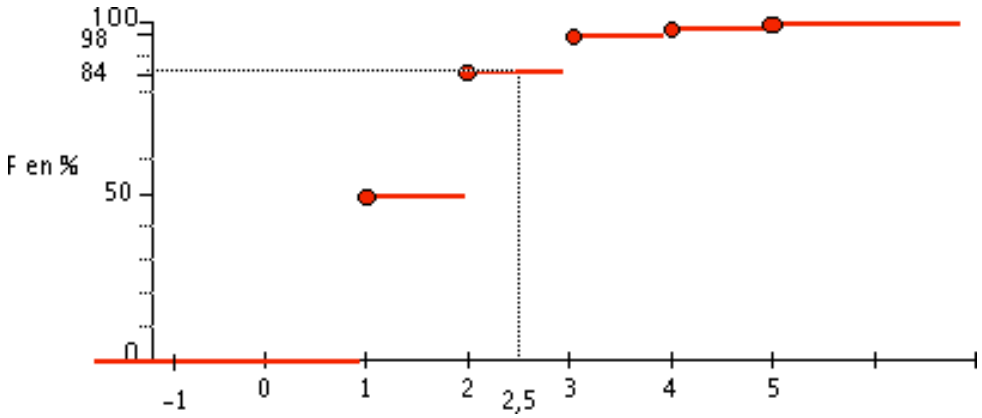
6 Variable discrète. Pour une valeur entière égale à une modalité m'_l, F retourne la proportion des individus ayant au plus cette modalité comme mesure, qui est la somme des fréquences des l premières modalités (la lème fréquence cumulée) : $F(m'_l) = \sum_{j=1}^{l} f_j$; pour une valeur q située entre deux modalités m'_l et m'_{l+1}, F retourne ce qu'elle retourne pour la modalité de gauche, $F(m'_l)$, puisque qu'aucun individu ne peut prendre comme mesure une valeur comprise entre les modalités m'_l et m'_{l+1}.

Ainsi, F est une fonction en escalier dont les contremarches sont à l'aplomb des modalités.

Exemple 1 portant sur le nombre d'enfants (on identifie "plus de 5" et "5") :

Nbre enfants [0-6]	1	2	3	4	5	Total
Fréq. f_{fm}	50	39,9	8,9	1,1	0,1	100
F_{fm}	50	90	98,8	99,9	100	
Fréq. f_c	50	33,3	14,7	1,8	0,2	100
F_c	50	83,3	98	99,8	100	
Fréq. f_Y	50	34,2	13,9	1,7	0,2	100
F_Y	50	84,2	98,2	99,9	100	

La fonction de répartition globale F_Y en pourcentage retourne 0 pour toutes les valeurs inférieures à 1, et 100 pour toutes les valeurs supérieures ou égales à 5 ; F change sur les entiers 1 à 5, et reste constante entre ces entiers ($F(2{,}5) = 84{,}2$ par exemple) :



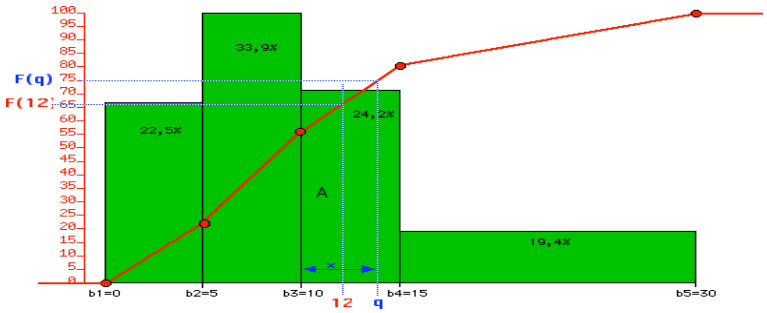
7 Variable continue. La surface de l'histogramme située à gauche de la verticale passant par q est égal à $F(q)$ ou à une approximation de $F(q)$, puisque cette surface représente la proportion de l'échantillon dont la valeur est inférieure ou égale à q, ou une approximation de cette proportion.

Si q est la borne gauche b_l d'une modalité-intervalle m'_l, $F(b_l)$ est exactement la somme des fréquences des l-1 premières modalités : $F(b_l) = \sum_{j=1}^{l-1} f_j$, la valeur retournée pour la dernière borne droite étant 1 (ou 100 en pourcentage).

Dans le cas du revenu des familles américaines, on peut évaluer exactement les fonctions de répartition sur les bornes 0 5 10 15 et 30 :

Modalités		0-4		5-9		10-14		15-		Total
Bornes	0		5		10		15		30	
Fréq. f_{Sud} (%)		*22,5*		*33,9*		*24,2*		*19,4*		100
F_{Sud} (%)	0		22,5		56,4		80,6		100	
Fréq. f_{Nord} (%)		*15,9*		*28,3*		*28,3*		*27,5*		100
F_{Nord} (%)	0		15,9		44,2		72,5		100	
Fréq. f_Y (%)		*18*		*30*		*27*		*25*		100
F_Y (%)	0		18		48		75		100	

Si q n'est pas une borne, la surface à gauche de la verticale passant par q est seulement une approximation de $F(q)$. On la calcule par la méthode décrite dans le paragraphe 3 ; évaluons par exemple $F_{Sud}(12)$, la proportion des individus du sud dont la mesure (donc ici le revenu) est inférieure ou égale à 12, approximée par la surface à gauche de la verticale passant par 12.



$F_{Sud}(12) = Prop(0 \leq Y_{sud} \leq 12) = 22{,}5 + 33{,}9 + A$; comme la largeur x de A vaut $12 - 10 = 2$ et comme sa hauteur est la densité $\frac{24{,}2}{5}$, $A = 2 * \frac{24{,}2}{5} \approx 9{,}7$ si bien que $F_{Sud}(12) \approx 66{,}1\%$.

8 **Calcul inverse.** Il consiste à déterminer la valeur q sachant que $F(q)$ est une proportion donnée p ; c'est un calcul utilisé dans les cours suivants, notamment à propos des quantiles.

Si p est la valeur de F sur une borne (c'est-à-dire p est une fréquence cumulée), la valeur recherchée est évidemment cette borne ; par exemple, si on doit déterminer la valeur q pour laquelle $F_{Sud} = 56{,}4\%$, on prendra $q = 10$.

Si ce n'est pas le cas, on obtient une approximation de q de la manière suivante qui s'apparente à la méthode décrite dans le paragraphe 3 :

1. on détermine d'abord l'intervalle $[b_l\ ; b_{l+1}[$ dans lequel se trouve q : on parcourt les intervalles de gauche à droite en cumulant leur fréquence, et on s'arrête dès qu'on dépasse p ; on a alors $F(b_l) < p$ et $F(b_{l+1}) > p$ (la surface à gauche de b_l est inférieure à p, et la surface à gauche de b_{l+1} est supérieure à p) ;
2. on détermine la distance x qui sépare b_l de q : pour que la surface à gauche de q soit égale à p il faut que la surface A dont x est la largeur soit égale à $p - F(b_l)$; comme la hauteur de A est la densité de la modalité $d_l = \frac{f_l}{a_l}$, on a $x * d_l = p - F(b_l)$, ou encore $x = \frac{p - F(b_l)}{d_l} = a_l * \frac{p - F(b_l)}{f_l}$ et $q = b_l + a_l * \frac{p - F(b_l)}{f_l}$

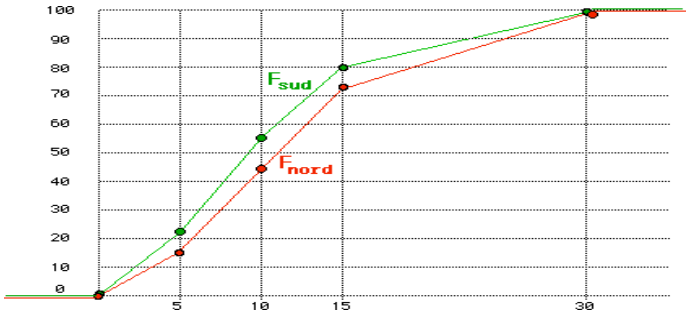
Déterminons par exemple la valeur q pour laquelle F_{Sud} vaut 75% :

1. q se trouve dans l'intervalle [10 ; 15[puisque $F_{Sud}(10) = 56{,}4 < 75\%$ et que $F_{Sud}(15) = 80{,}6 > 75\%$;
2. x doit être la largeur d'une surface de $75 - 56{,}4 = 18{,}6\%$; comme la densité de la modalité est $\frac{24{,}2}{5} = 4{,}84$, on a $x * 4{,}84 = 18{,}6$, ou encore $x = \frac{18{,}6}{4{,}84} = 3{,}8$, et donc $q = 10 + 3{,}8 = 13{,}8$.

Classement par fonction de répartition

9 Comme dans le cas des variables ordonnées, un ordre partiel ou total des fonctions de répartition conditionnelles induit un classement des distributions conditionnelles associées et par conséquent des sous-populations.

Appliquons cette procédure à l'exemple 1 ; la représentation graphique simultanée des fonctions de répartition conditionnelles, F_{Sud} en vert et F_{Nord} en rouge donne :



Comme $F_{Nord} < F_{Sud}$, la distribution conditionnelle Y_{Nord} est globalement supérieure à la distribution conditionnelle Y_{Sud} (attention à l'inversion de l'ordre), ce qui peut s'interpréter comme le fait que le revenu des familles est globalement supérieur dans la sous-population du nord des États-Unis que dans celle du sud, ou encore que la sous-population du nord est globalement plus riche que la sous-population du sud.

Programme de travail

Savoir définir :

- une fonction de répartition ;
- la notation $F(q)$.

Savoir expliquer :

- la différence entre variable discrète et variable continue ;
- la construction d'un diagramme en bâtons ;
- la construction d'un histogramme .

Savoir faire :

- construire un diagramme en bâtons ;
- construire un histogramme ;
- calculer la surface d'une portion d'histogramme ;
- déterminer la valeur q connaissant $F(q)$;
- dessiner une fonction de répartition dans le cas discret et continu ;
- classer les sous-populations par comparaison des fonctions de répartition conditionnelles .