

Partie 2 : Statistique élémentaire

Plan

1. Statistique élémentaire (rappel)
2. Statistique descriptive unidimensionnelle
3. Statistique descriptive bidimensionnelle

Statistique élémentaire : vocabulaire

Population Ω (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

Individu $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Variable (statistique) : $\Omega \xrightarrow{X} \begin{cases} \mathcal{E} & \text{si qualitative} \\ \mathbb{R} & \text{si quantitative} \end{cases}$

caractéristique (âge, salaire, sexe, glycémie...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans \mathbb{R} (ou une partie de \mathbb{R} , ou un ensemble de parties de \mathbb{R}), elle est dite *quantitative* (âge, salaire, taille...); sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle...). Si les modalités d'une variables qualitatives sont ordonnées (*i.e.* tranches d'âge), elle est dite *qualitative ordinaire* et sinon *qualitative nominale*.

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

Statistique descriptive unidimensionnelle : Variable quantitative

Variable quantitative discrète

On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise ; la série statistique brute est donnée ci-dessous (il s'agit de données fictives).

43	29	57	45	50	29	37	59	46	31	46	24	33	38	49	31
62	60	52	38	38	26	41	52	60	49	52	41	38	26	37	59
57	41	29	33	33	43	46	57	46	33	46	49	57	57	46	43

Présentation des données

Le tableau statistique C'est un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X ; ces observations sont rangées par ordre croissant et non répétées ; nous les noterons $\{x_l \mid l = 1, \dots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_l , le nombre de réplications qui lui sont associées ; ces réplications sont

appelées *effectifs* et notées n_l . Les effectifs n_l sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$, appelées *fréquences* (rappelons que n désigne le nombre total d'observations, c'est-à-dire le cardinal de Ω : $n = \sum_{l=1}^r n_l$).

x_l	n_l	N_l	$f_l(\%)$	$F_l(\%)$
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,00

Les effectifs cumulés et les fréquences cumulées Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_l = \sum_{j=1}^l n_j \text{ et } F_l = \sum_{j=1}^l f_j.$$

On notera que $N_r = n$ et $F_r = 1$.

Représentations graphiques

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques, qui sont en fait complémentaires : le diagramme en bâtons et le diagramme cumulatif (en escaliers).

Le diagramme en bâtons Il permet de donner une vision d'ensemble des observations réalisées.

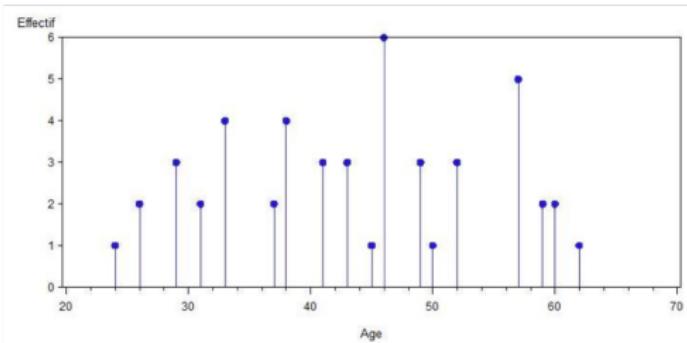


Diagramme en bâtons

Le diagramme cumulatif Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion)

d'observations inférieures ou égales à une valeur donnée de la série. Lorsqu'il est relatif aux fréquences, c'est en fait le graphe de la *fonction de répartition empirique* F_X définie de la façon suivante :

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1, \\ F_l & \text{si } x_l \leq x < x_{l+1}, \quad l = 1, \dots, r-1, \\ 1 & \text{si } x \geq x_r. \end{cases}$$

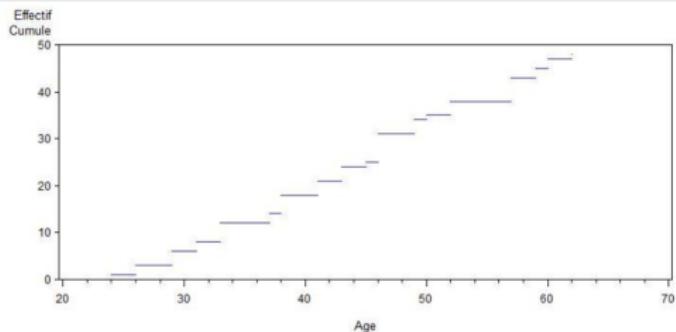


Diagramme cumulatif

Notion de quantile

Définition La fréquence cumulée F_l ($0 \leq F_l \leq 1$) donne la proportion d'observations inférieures ou égales à x_l . Une approche complémentaire consiste à se donner a priori une valeur α , comprise entre 0 et 1, et à rechercher x_α vérifiant $F_X(x_\alpha) \simeq \alpha$. La valeur x_α (qui n'est pas nécessairement unique) est appelée quantile (ou *fractile*) d'ordre α de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de α .

La médiane et les quartiles La médiane est le quantile d'ordre $\frac{1}{2}$; elle partage donc la série des observations en deux ensembles d'effectifs égaux. Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane).

Le diagramme-boîte (ou “box-and-whisker plot”) Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane. La figure 3 donne le diagramme-boîte de l'exemple 2.1. Dans cet exemple, on a obtenu $x_{\frac{1}{4}} = 35$, $x_{\frac{1}{2}} = 44$ et $x_{\frac{3}{4}} = 52$; on notera que l'obtention, d'une part de $x_{\frac{1}{4}}$ et $x_{\frac{1}{2}}$, d'autre part de $x_{\frac{3}{4}}$, ne s'est pas faite de la même façon (en fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple).

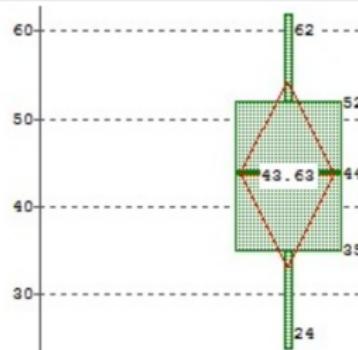


Diagramme-boîte et moyenne en rouge

Tendance centrale Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est-à-dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

- la *médiane*,
- la *moyenne* (ou moyenne arithmétique).

Formule de la moyenne pour une variable quantitative discrète :

$$\bar{x} = \frac{1}{n} \sum_{l=1}^r n_l x_l = \sum_{l=1}^r f_l x_l.$$

Dispersion Elles servent à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

- L'*étendue* ($x_r - x_1$),
- l'*intervalle inter-quartiles* ($x_{\frac{3}{4}} - x_{\frac{1}{4}}$),
- l'*écart-moyen à la médiane* ($\frac{1}{n} \sum_{l=1}^r n_l |x_l - x_{\frac{1}{2}}|$),
- l'*écart-moyen à la moyenne* ($\frac{1}{n} \sum_{l=1}^r n_l |x_l - \bar{x}|$),

sont des caractéristiques de dispersion que l'on rencontre parfois.

Mais, la caractéristique de loin la plus utilisée est l'**écart-type**, racine carrée positive de la *variance*. Formules de la variance :

$$\begin{aligned} \text{var}(X) = \sigma_X^2 &= \frac{1}{n} \sum_{l=1}^r n_l(x_l - \bar{x})^2 \\ &= \frac{1}{n} \sum_{l=1}^r n_l(x_l)^2 - (\bar{x})^2. \end{aligned}$$

L'écart-type de X sera donc noté σ_X .

Illustration En utilisant toujours l'exemple 2.1, on a calculé :

$$\bar{x} = \frac{1}{n} \sum_{l=1}^r n_l x_l = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans};$$

$$\sigma_X^2 = \frac{1}{n} \sum_{l=1}^r n_l(x_l)^2 - (\bar{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760;$$

$$\sigma_X = \sqrt{\sigma_X^2} \simeq 10,5 \text{ ans.}$$

Variable quantitative continue

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous-ensemble de \mathbb{R} des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés *classes*.

Deux exemples de variables quantitatives fréquemment considérées comme continues sont l'âge et le revenu (pour un groupe d'individus).

Nous noterons $(b_0 ; b_1), \dots, (b_{r-1} ; b_r)$ les classes considérées. Les nombres b_{l-1} et b_l sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1} + b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée a_l).

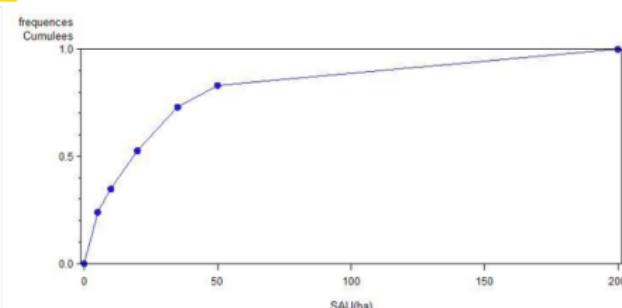
Présentation des données

Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la SAU (surface agricole utilisée) exprimée en hectares (Tableaux Économiques de Midi-Pyrénées, INSEE, 1989, p. 77) ; la SAU est ici une variable quantitative continue comportant 6 classes.

SAU (en ha)	fréquences (%)
moins de 5	24,0
de 5 à 10	10,9
de 10 à 20	17,8
de 20 à 35	20,3
de 35 à 50	10,2
plus de 50	16,8

Représentations graphiques

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.



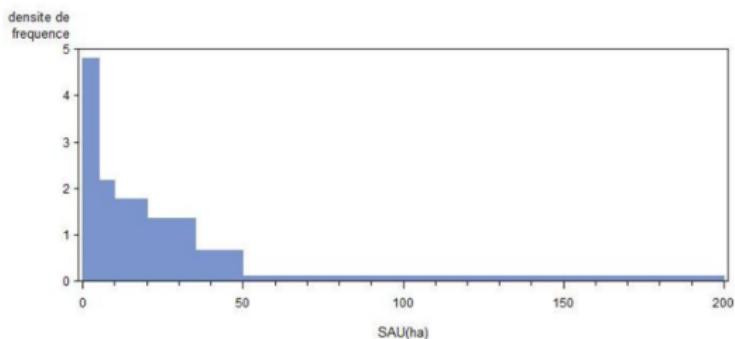
Courbe cumulative C'est encore une fois le graphe de la *fonction de répartition empirique*, cette dernière devant maintenant être précisée au moyen d'*interpolations linéaires*.

On appelle fonction de répartition empirique de la variable continue X la fonction F_X définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \dots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

(on a supposé $F_0 = 0$).

Histogramme La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses b_0, b_1, \dots, b_r . Sa fonction dérivée, éventuellement non définie en ces points, est appelée **densité empirique de X** et notée f_X . On obtient :



$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \dots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Le graphe de f_X est alors appelé histogramme de la variable X . Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$) et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées **densités de fréquence**. L'aire du $l^{\text{ième}}$ rectangle vaut donc f_l , fréquence de la classe correspondante.

Moyenne et écart-type

La moyenne, la variance et l'écart-type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour x_l les centres de classes au lieu des observations (qui ne sont pas connues). Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.

Illustration

La médiane de la variable présentée dans l'exemple 2.2 se situe dans la classe (10 ; 20), puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante (l'indice l ci-dessous désigne en fait la troisième classe) :

$$\begin{aligned}x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \\&= 10 + 10 \frac{15,1}{17,8} \\&\simeq 18,5 \text{ ha.}\end{aligned}$$

La moyenne vaut :

$$\bar{x} = \sum_{l=1}^r f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

Statistique descriptive unidimensionnelle : Variable qualitative

Variables nominales et ordinaires

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées *modalités*. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite *ordinale*. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite *nominale*.

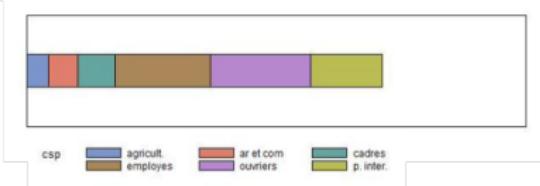
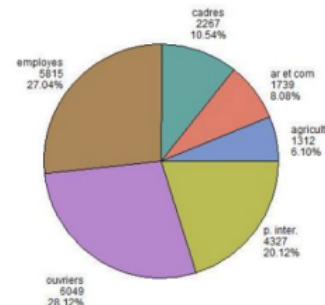
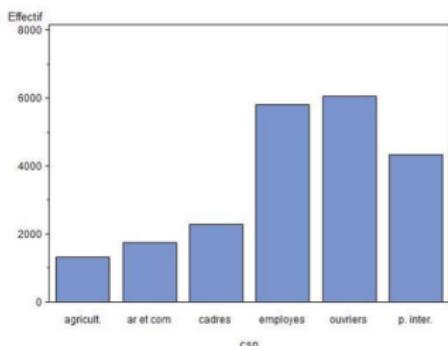
Traitements statistiques

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinaire). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques.

Représentations graphiques

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :

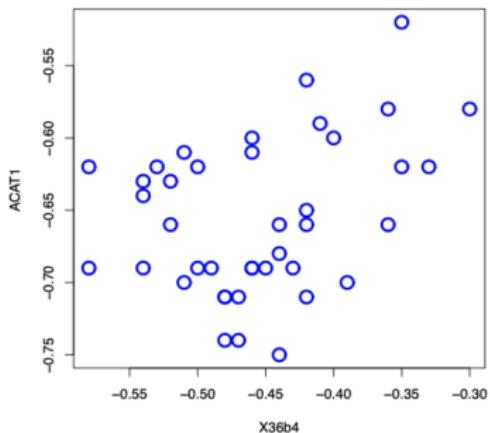
- le *diagramme en colonnes*,
- le *diagramme en barre*,
- le *diagramme en secteurs*.



Statistique descriptive bidimensionnelle : Deux variables quantitatives

NUAGE DE POINTS

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y , puis à représenter chaque individu observé par les coordonnées des valeurs observées. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction plus fidèle de l'anglais *scatter-plot*.



Variables centrées et réduites

Si X est une variable quantitative de moyenne \bar{x} et d'écart-type σ_X , on appelle variable centrée associée à X la variable $X - \bar{x}$ (elle est de moyenne nulle et d'écart-type σ_X), et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\frac{X - \bar{x}}{\sigma_X}$ (elle est de moyenne nulle et d'écart-type égal à un). Une variable centrée et réduite s'exprime sans unité.

Indice de liaison

Le coefficient de corrélation linéaire est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Il est défini à partir de la covariance qui généralise à deux variables la notion de variance :

$$\begin{aligned}\text{cov}(X, Y) &= \sum_{i=1}^n w_i [x_i - \bar{x}] [y_i - \bar{y}] \\ &\equiv \sum_{i=1}^n w_i x_i y_i - \bar{x} \bar{y}.\end{aligned}$$

la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque".

C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (appelé coefficient de Pearson ou de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types :

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $\text{corr}(X, Y) = \text{cov}\left(\frac{X-\bar{x}}{\sigma_X}, \frac{Y-\bar{y}}{\sigma_Y}\right)$.

Par conséquent, $\text{corr}(X, Y)$ est indépendant des unités de mesure de X et de Y . Le coefficient de corrélation est *symétrique* et prend ses valeurs entre -1 et +1. Les valeurs -1 et +1 correspondent à une liaison linéaire parfaite entre X et Y (existence de réels a , b et c tels que : $aX + bY + c = 0$).

Statistique descriptive bidimensionnelle : Une variable quantitative et une qualitative

Soit X la variable qualitative considérée, supposée à m modalités notées

$$x_1, \dots, x_\ell, \dots, x_m$$

et soit Y la variable quantitative de moyenne \bar{y} et de variance σ_Y^2 . Désignant par Ω l'échantillon considéré, chaque modalité x_ℓ de X définit une sous-population (un sous-ensemble) Ω_ℓ de Ω : c'est l'ensemble des individus, supposés pour simplifier de poids $w_i = 1/n$ et sur lesquels on a observé x_ℓ ; on obtient ainsi une *partition* de Ω en m classes dont nous noterons n_1, \dots, n_m les cardinaux (avec toujours $\sum_{\ell=1}^m n_\ell = n$, où $n = \text{card}(\Omega)$).

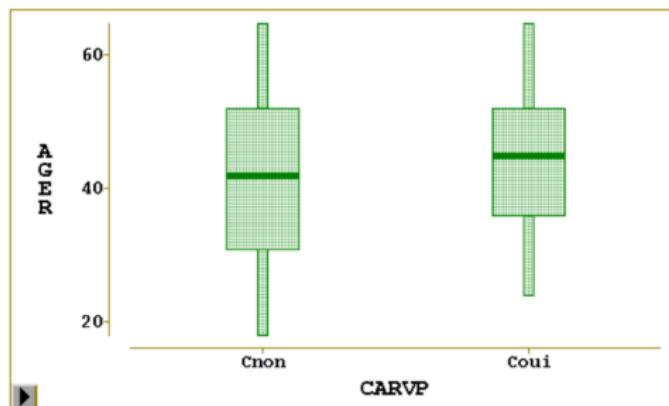
Considérant alors la restriction de Y à Ω_ℓ ($\ell = 1, \dots, m$), on peut définir la moyenne et la variance partielles de Y sur cette sous-population ; nous les noterons respectivement \bar{y}_ℓ et σ_ℓ^2 :

$$\bar{y}_\ell = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i);$$

$$\sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \bar{y}_\ell]^2.$$

Boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des diagrammes-boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y un diagramme-boîte pour chacune des sous-populations définies par X . La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y , c'est-à-dire de la liaison entre les deux variables.



Banque : Diagrammes-boîtes illustrant les différences de distribution des âges en fonction de la possession d'une carte Visa Premier.

Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-à-dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables.

$$\bar{y} = \frac{1}{n} \sum_{\ell=1}^m n_\ell \bar{y}_\ell;$$
$$\sigma_Y^2 = \frac{1}{n} \sum_{\ell=1}^m n_\ell (\bar{y}_\ell - \bar{y})^2 + \frac{1}{n} \sum_{\ell=1}^m n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2.$$

Le premier terme de la décomposition de σ_Y^2 , noté σ_E^2 , est appelé *variance expliquée* (par la partition, c'est-à-dire par X) ou *variance inter* (between) ; le second terme, noté σ_R^2 , est appelé *variance résiduelle* ou *variance intra* (within).

Rapport de corrélation

Il s'agit d'un indice de liaison entre les deux variables X et Y qui est défini par :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}};$$

X et Y n'étant pas de même nature, $s_{Y/X}$ n'est pas symétrique et vérifie $0 \leq s_{Y/X} \leq 1$.

Statistique descriptive bidimensionnelle : Deux variables qualitatives

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X , possède r modalités notées $x_1, \dots, x_\ell, \dots, x_r$, et que la seconde, notée Y , possède c modalités notées $y_1, \dots, y_h, \dots, y_c$.

Ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

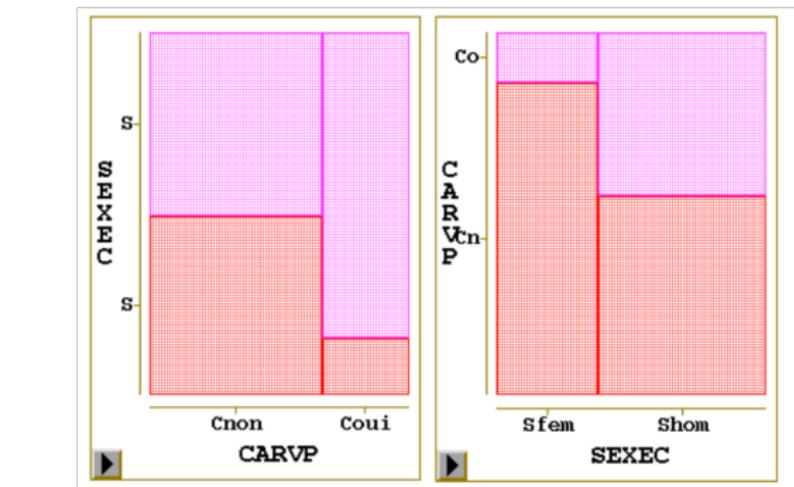
Une table de contingence se présente donc sous la forme suivante :

	y_1	\dots	y_h	\dots	y_c	sommes
x_1	n_{11}	\dots	n_{1h}	\dots	n_{1c}	n_{1+}
	\vdots		\vdots		\vdots	\vdots
x_ℓ	$n_{\ell 1}$	\dots	$n_{\ell h}$	\dots	$n_{\ell c}$	$n_{\ell +}$
	\vdots		\vdots		\vdots	\vdots
x_r	n_{r1}	\dots	n_{rh}	\dots	n_{rc}	n_{r+}
sommes	n_{+1}	\dots	n_{+h}	\dots	n_{+c}	n

Les quantités $n_{\ell+} (\ell = 1, \dots, r)$ et $n_{+h} (h = 1, \dots, c)$ sont appelées les *effectifs marginaux*; ils sont définis par $n_{\ell+} = \sum_{h=1}^c n_{\ell h}$ et $n_{+h} = \sum_{\ell=1}^r n_{\ell h}$, et ils vérifient $\sum_{\ell=1}^r n_{\ell+} = \sum_{h=1}^c n_{+h} = n$. De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

Représentations graphiques des profils

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.



Banque : Diagrammes en barres des profils lignes et colonnes (mosaïque plot) de la table de contingence croisant le sexe et la possession de la carte Visa Premier. La superficie de chaque case est en plus proportionnelle à l'effectif de la cellule associée.