LOI QUALITATIVE (C6, G11, H7, I9, J, N)
(29 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'expression « **loi qualitative** » désigne (de façon abrégée) la **loi de probabilité** d'une **variable qualitative**.

On distingue entre **variable qualitative « simple »** et **variable qualitative « multiple »**.

(i) **Loi qualitative simple**, ou **loi qualitative univariée**. Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $(\mathcal{K}, \mathcal{D})$ un **espace mesurable** et $\kappa : \Omega \mapsto \mathcal{K}$ une **variable qualitative** (simple) à valeurs dans un ensemble (fini) $\mathcal{K} = \{k_1, ..., k_M\}$ dont les éléments k_m sont appelés **modalités**.

La **loi** P^κ de κ est appelée **loi qualitative**. Elle est définie comme image de P par κ. La **suite** des **probabilités** élémentaires p_m associée aux M modalités de κ, donc à P^κ, est donc, par définition :

(1) $P([\kappa = k_m]) = P(\{\omega \in \Omega : \kappa(\omega) = k_m\}) = p_m$, $\forall\, m \in N_M^* = \{1, ..., M\}$.

La **loi de probabilité** de κ , notée P^κ ou $\mathcal{L}(\kappa)$, s'écrit sous l'une des formes :

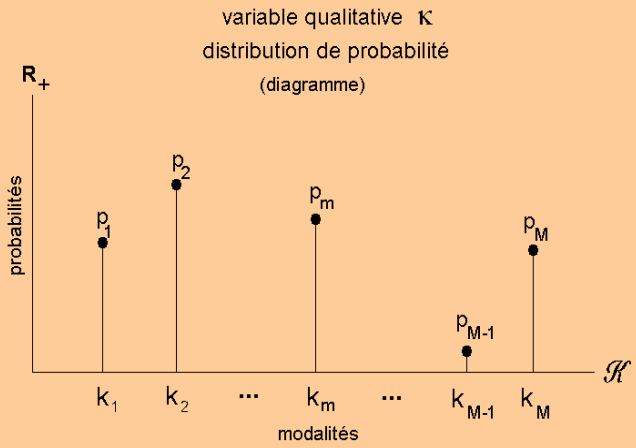
$(1)_a$ $P^\kappa = \Sigma_{m=1}^{M} p_m \,.\, \delta_{k(m)}$,

ou encore :

$(1)_b$ $P^\kappa = \Sigma_{m=1}^{M} p_m \,.\, \mathbf{1}_{[\kappa = k(m)]}$,

où $\delta_{k(m)}$ désigne la **loi de DIRAC** placée au « point » k_m (aussi noté k(m)), $\mathbf{1}_D$ désigne la **fonction indicatrice** d'une partie $D \in \mathcal{D}$, et la suite $(p_m)_{m=1,...,M}$ vérifie les propriétés simplectiques usuelles, ie : $p_m \geq 0$, $\forall\, m \in N_M^*$ et $\Sigma_{m=1}^{M} p_m = 1$.

Dans la représentation graphique suivante, les « écarts » entre modalités k_m n'ont pas de signification. Les probabilités p_m sont présentées en suivant l'ordre 1 ,..., M attribué aux modalités elles-mêmes.



Une représentation tabulaire de cette loi est la suivante (tableau en ligne).

variable qualitative κ
distribution de probabilité
(tableau en ligne)

modalités

$\mathcal{K}$	k_1	k_2	...	k_m	...	k_{M-1}	k_M
P	p_1	p_2	...	p_m	...	p_{M-1}	p_M

probabilités

(ii) **Loi qualitative multiple**, ou **loi qualitative multivariée**. Dans le cas d'une va qualitative multiple $\kappa = (\kappa_1, ..., \kappa_H)$, on considère H espaces mesurables (**espace d'observation**) $(\mathcal{K}_h, \mathcal{D}_h)_{h=1,...,H}$, avec $\mathcal{K}_h = \{k_{h,1} ..., k_{h,M(h)}\}$ (M_h modalités), $\forall\ h \in N_H^* = \{1, ..., H\}$. Chaque variable simple $\kappa_h : \Omega \mapsto \mathcal{K}_h$ ($h \in N_H^*$) prend donc ses valeurs (modalités) dans $\mathcal{K}_h$. On note donc $\mathcal{K} = \prod_{h=1}^{H} \mathcal{K}_h$, $\mathcal{D} = \otimes_{h=1}^{H} \mathcal{D}_h$ et $\kappa = (\kappa_h)_{h=1,...,H}$.

La loi P^κ de κ est la **loi conjointe** de cette suite de **variables qualitatives** simples. C'est donc une **loi multivariée** particulière, associée exclusivement à des **va** qualitatives.

Comme précédemment, la suite des probabilités élémentaires $p_{m(1)...m(H)}$ associée aux M_h modalités m_h de κ_h (avec $m_h \in N_{M(h)}^*$ et $h \in N_H^*$), donc à P^κ, est définie selon :

$$(2) \qquad P\left([(\kappa_h)_{h=1,...,H} = (k_{h,m(h)})_{h=1,...,H}]\right) = P\left(\{\omega \in \Omega : (\kappa_h(\omega))_{h=1,...,H} = (k_{h,m(h)})_{h=1,...,H}\}\right) = p_{m(1),...,m(H)},$$

$\forall\ m_h \in N_{M(h)}^* = \{1, ..., M_h\}$ et $\forall\ h \in N_H^* = \{1, ..., H\}$, soit $\prod_{h=1}^{H} M_h$ probabilités élémentaires.

P^{κ} s'écrit sous la forme :

$$(2)_a \quad P^{\kappa} = \Sigma_{m(1)=1}^{M(1)} \ldots \Sigma_{m(H)=1}^{M(H)} \, p_{m(1)\ldots m(H)} \, . \, \delta \, (k_{1,m(1)} \, ,\ldots, \, k_{H,m(H)}),$$

ou

$$P^{\kappa} = \Sigma_{h=1}^{H} \Sigma_{m(h)=1}^{M(h)} \, p_{m(1)\ldots m(H)} \, . \, \delta \, (k_{1,m(1)} \, ,\ldots, \, k_{H,m(H)}),$$

expressions dans lesquelles $\delta \, (k_{1,m(1)} \, ,\ldots, \, k_{H,m(H)})$ désigne la **loi de DIRAC** placée au « point » $(k_{1,m(1)} \, ,\ldots, \, k_{H,m(H)})$, les m(h) désignent, par commodité, les m_h , et $(p_{m(1)\ldots m(H)})_{m=1,\ldots,M}$ est la suite des probabilités élémentaires, qui vérifie donc les propriétés usuelles : $p_{m(1)\ldots m(H)} \geq 0$, $\forall$ m(h) $\in \{1 \, ,\ldots, \, M_h\}$ et $\forall$ h $\in N_H^*$, et $\Sigma_{h=1}^{H} \Sigma_{m(h)=1}^{M(h)} \, p_{m(1)\ldots m(H)} = 1$.

P^{κ} s'écrit aussi sous la forme :

$$(2)_b \quad P^{\kappa} = \Sigma_{m(1)=1}^{M(1)} \ldots \Sigma_{m(H)=1}^{M(H)} \, p_{m(1)\ldots m(H)} \, . \, \mathbf{1}_{D(m(1),\ldots,m(H))} \, ,$$

ou

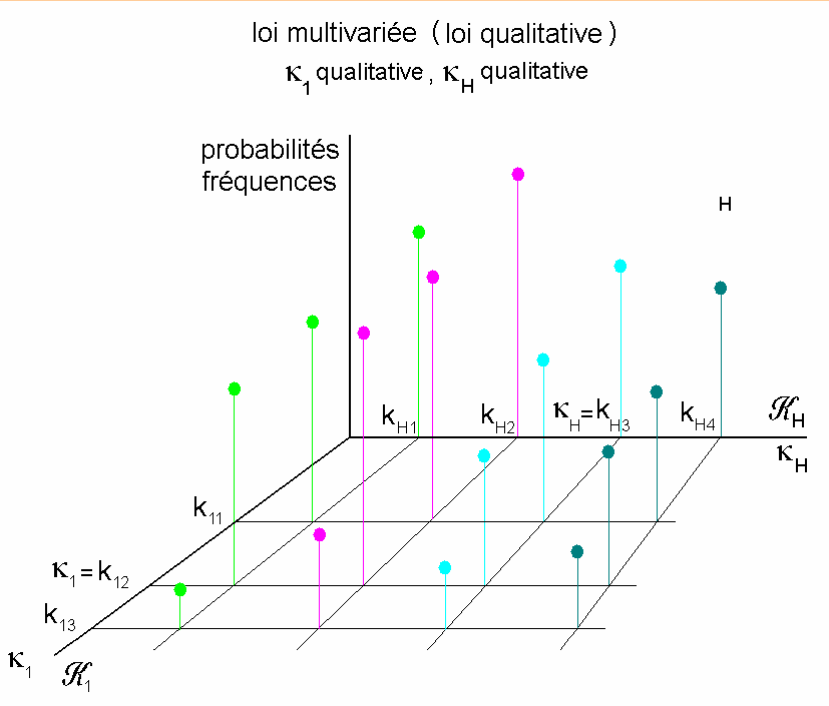
$$P^{\kappa} = \Sigma_{h=1}^{H} \Sigma_{m(h)=1}^{M(h)} \, p_{m(1)\ldots m(H)} \, . \, \mathbf{1}_{D(m(1),\ldots,m(H))} \, ,$$

expressions dans lesquelles $\mathbf{1}_{D(m(1),\ldots,m(H))}$ désigne la **fonction indicatrice** d'une partie $D_{m(1),\ldots,m(H)} \in \mathcal{D}$, définie par :

$$(3) \quad D_{m(1),\ldots,m(H)} = \cap_{h=1}^{H} ([\kappa_1 = k_{1,m(1)}], \cap \ldots \cap [\kappa_H = k_{H,m(H)}]),$$

où les m(h) désignent resp les m_h et $(p_{m(1)\ldots m(H)})$ est la suite des probabilités élémentaires précédente, indexée par $\Pi_{h=1}^{H} N_{M(h)}^* = \{1 \ldots M_1\} \times \ldots \times \{1 \ldots M_H\}$.

Exemple de représentation graphique d'une loi qualitative à 2 variables κ_1 et κ_H



Le graphe d'une loi de ce type, dont les variables sont entièrement qualitatives, doit être distingué de la notion de **stéréogramme**, qui généralise (aux variables multidimensionnelles quantitatives) la notion d'**histogramme** (relative à une variable numérique scalaire).

(iii) Une loi qualitative (multiple), qui est donc la loi jointe d'une suite finie constituée exclusivement de va qualitatives, peut aussi se représenter sous la forme d'un **tableau de contingence** à H dimensions. Ce tableau est ainsi caractérisé par deux informations :

(a) la suite $\kappa = (\kappa_h)_{h=1,...,H}$ des H variables qualitatives et celle de leurs modalités $\mathcal{K}_h = \{k_{h,1} ..., k_{h,M(h)}\}$;

(b) la suite (multiple) des probabilités élémentaires $(p_{m(1)...m(H)})$ précédente, avec $m(h) \in \{1 ,..., M_h\}$, $\forall\ h \in \{1 ,..., H\}$.

En pratique, un tel tableau est défini par le « croisement » de plusieurs critères (souvent deux). Une représentation tabulaire de cette loi multiple est la suivante (tableau « croisé » en lignes et colonnes).

Tableau de contingence (cas de deux qualitatives κ_1 et κ_H)

tableau de contingence $M_1 \times M_H$ (distribution de probabilité)

variables qualitatives κ_1 et κ_M

modalités	$\mathcal{K}_2$ / $\mathcal{K}_1$					
	k_{H1}	...	k_{Hm_H}	...	k_{HM_H}
k_{11}	$p_{11,H1}$	...	p_{11,Hm_H}	...	p_{11,HM_H}
...	...	...	...	...	...
k_{1m_1}	$p_{1m_1,H1}$	...	p_{1m_1,Hm_H}	...	p_{1m_1,HM_H}
...	...	...	...	...	...
k_{1M_1}	$p_{1M_1,H1}$	...	p_{1M_1,Hm_H}	...	p_{1M_1,HM_H}

modalités (vertical) ; probabilités

Comme toute loi de probabilité, la loi $\mathcal{L}(\kappa)$ peut posséder certaines **caractéristiques légales**. Ces caractéristiques peuvent se calculer à partir des suites $p_{m(1)...m(H)} \geq 0$, $\forall$ m(h) $\in \{1, ..., M_h\}$ et $\forall$ h $\in N_H^*$.

(iv) Ainsi, dans le cas d'une variable simple, et quel que soit son type (nominal ou ordinal) de κ, P^κ possède toujours une **caractéristique de centralité** simple, son **mode**, usuellement défini par les valeurs de κ tq :

$(4)_a$ $\quad k_{max} = \{k_m : p_m = max_{s=1,...,H}\ p_s\}$.

Si les p_m sont toutes distinctes deux à deux, k_{max} est unique, et peut être noté k_{mod} .

Ce **mode** est donc toujours défini, mais (comme pour une variable numérique) il n'est pas nécessairement unique. En effet, une **partie modale**, ou **zone modale**, de P^κ (ou de κ) est une partie mesurable S $\kappa \in \mathcal{D}$ tq, par définition (en supposant fixée la « mesure » de D) :

$(4)_b$ $\quad P^\kappa(S\ \kappa) \geq P^\kappa(D), \quad \forall\ D \in \mathcal{D}$,

ou encore :

$(4)_c$ $\quad P^\kappa(S\ \kappa) = sup_{D \in \mathcal{D}}\ P^\kappa(D)$.

En particulier, si Card (S κ) = 1, on parle de **mode** de P^κ (ou de κ) ;

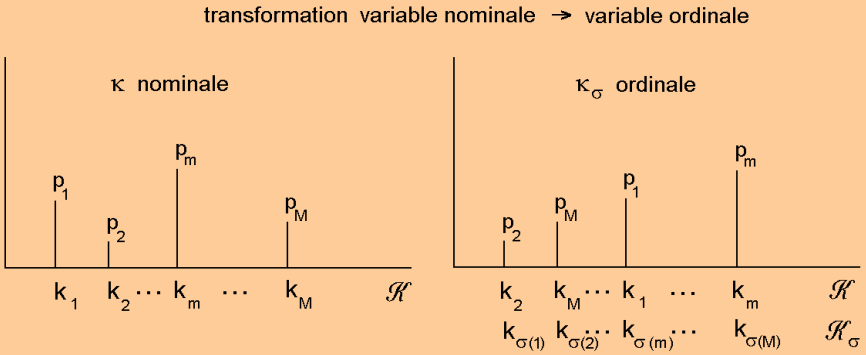
On peut définir, de façon parallèle, la notion d'**antimode** comme l'ensemble des valeurs de κ tq leurs probabilités soient minimum.

(v) Si κ est de type ordinal (et ses valeurs k_m ordonnées par $\leq$), on peut définir une notion de **fonction de répartition** G basée sur l'ordre (ou préordre) $\leq$ selon (cf aussi **fonction de répartition**) :

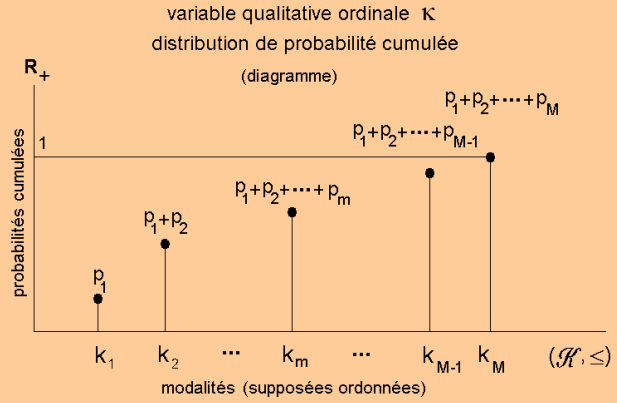
(5) $\quad G(k) = P(\kappa \leq k) = \Sigma_{m=1}^{M} \mathbf{1}_{[\kappa \leq k]} \cdot p_m = \Sigma_{m=1}^{M} u(\kappa \leq k) \cdot p_m$,

où u désigne la **fonction de HEAVYSIDE** et où les p_m sont ordonnées comme les k_m .

Par ailleurs, on peut transformer une nominale simple en ordinale simple à partir d'un ordre entre modalités défini par les **fréquences** (cf aussi **fréquence relative**).



Cumul des probabilités élémentaires d'une variable ordinale



G s'incrémente donc d'une valeur p_m entre une modalité k_{m-1} et la suivante k_m . Les « écarts apparents $k_m - k_{m-1}$ n'ont pas de sens : ce sont les masses p_m qui définissent l'ordre.

(vi) Toujours lorsque κ est ordinale (et ses valeurs ordonnées par $\leq$), d'autres **caractéristiques légales** peuvent se déduire de G : leurs définitions sont analogues à celles d'une variable numérique : cf **médiane**, **quantile**. En effet :

(a) il est généralement possible de définir des **quantiles** à partir de l'ordonnancement des modalités de κ dans $\mathscr{K}$, puisque la notion de « cumul » (donc de « fréquence cumulée ») peut être définie pour une variable ordinale.

Si $p \in]0, 1[$, on appelle **quantile d'ordre p**, ou **p-quantile**, ou encore **p-fractile**, théorique, de P^{κ} (de κ, ou encore de G), s'il existe, toute modalité k_p , souvent notée $Q_p\ \kappa \in \mathscr{K}$, de κ tq, à la fois :

(6) $P([\kappa \leq Q_p\ \kappa]) \geq p$ et $P([\kappa \geq Q_p\ \kappa]) \geq 1 - p$.

Lorqu'il ne se réduit pas à un seul élément de $\mathscr{K}$, l'ensemble Q_p des éléments $Q_p\ \kappa$ vérifiant (4) est appelé **ensemble des quantiles** (ou parfois « **intervalle » quantilaire**) d'ordre p de P^{κ}.

On peut exprimer (6) à l'aide de la fr G associée à P^{κ}, selon :

(7) $G(Q_p\ \kappa) \geq p$ et $1 - G(Q_p\ \kappa) \geq 1 - p$;

(b) on peut déduire de (a) une seconde **caractéristique de centralité** $Q_{1/2}\ \kappa$ (parfois notée k_{med}) associée à P^{κ} (ou à G), appelée **médiane**, définie par les valeurs de κ tq, à la fois :

(8) $P([\kappa \leq Q_{1/2}\ \kappa]) \geq 1 / 2$ et $P([\kappa \geq Q_{1/2}\ \kappa]) \geq 1 / 2$,

ie tout quantile d'ordre $p = 1 / 2$ de P^{κ}.

L'ensemble $\mathscr{E}$ des valeurs $Q_{1/2}\ \kappa \in \mathscr{K}$ vérifiant (8), ie l'ensemble des valeurs médianes, est appelé **intervalle** (ou **ensemble**) **médian** de P^{κ} (ou de κ). La définition (8) s'écrit aussi :

(9) $G(Q_{1/2}\ \eta) \geq 1 / 2$ et $1 - G(Q_{1/2}\ \eta) \geq 1 / 2$.

(vii) si P^{κ} est dominée par une **mesure positive** ν définie sur $\mathscr{D}$, on peut (formellement) parler de « **densité de probabilité** » de P^{κ} pr à ν, ce qui se note généralement $g = dP^{\kappa} / d\nu$ (**dérivée de NIKODYM-RADON**).

(viii) L'estimation d'une loi qualitative (eg exprimée sous forme d'un tableau de contingence) revient à estimer les « **probabilités des cases** », ie les $(p_m)_{m=1,...,M}$, dans le cas univarié, ou les $p_{m(1)...m(H)}$ (avec $m(h) \in \{1 ,..., M_h\} = N_{M(h)}^*$ et $h \in N_H^*$), dans le cas multivarié.

L'**échantillon aléatoire** observé est constitué de N **unités statistiques** $\{u_1 ,..., u_N\}$ susceptibles de posséder certains des « caractères » défini par les variables qualitatives considérées. On calcule alors les **fréquences** de ces événements, fréquences absolues puis relatives (cf **fréquences absolues**) :

(a) dans le cas univarié, le nombre d'unités u_n possédant la modalité k_m ($m = 1, ..., M$) de κ est noté n_m , avec $\Sigma_{m=1}^{M}\ n_m = N$ (parfois noté n. ou n_o). L'estimateur de p_m est alors simplement $\tilde{p_m} = n_m / N$;

(b) dans le cas multiple, la démarche est analogue. Le nombre d'unités u_n possédant, à la fois, la modalité $k_{1,m(1)}$ de κ_1 ,..., et $k_{H,m(H)}$ de κ_H est notée $n_{m(1)...m(H)}$, $\forall$ h $\in$ {1 ,..., H}, avec $\Sigma_{h=1}^{H} \Sigma_{m(h)=1}^{M(h)} n_{m(1)...m(H)} = N$ (parfois noté n. ... . ou $n_{o...o}$). L'estimateur de $p_{m(1)...m(H)}$ est alors simplement $\tilde{p}_{m(1)...m(H)} = n_{m(1)...m(H)} / N$.

(ix) Par « **tableau de contingence** » on entend, selon le **contexte statistique** :

(a) soit la loi de probabilité qualitative elle-même (concept abstrait), constituée de M (cas simple) ou Card $\mathscr{K} = \Pi_{h=1}^{H} M_h$ (cas multiple) probabilités élémentaires ;

(b) soit le tableau des fréquences (absolues ou relatives) n_m ou $n_{m(1)...m(H)}$ précédent.

(x) Il est possible de définir, à partir d'une loi qualitative P^{κ} (théorique ou estimée), les notions usuelles de :

(a) **loi marginale** ;

(b) **loi conditionnelle**, dont on déduit la notion de **relation fonctionnelle** (relation fonctionnelle numérique reliant des probabilités ou fréquences d'attributs).