Éléments de Statistique Descriptive

1 Présentation des Données Statistiques

Définitions :

- Une population population statistique est l'ensemble Ω des objets qu'on étudie.
- Une variable aléatoire est une fonctionnelle d'un ensemble Ω vers (en générale) l'axe réel. En statistiques descriptive on essaye de résumer l'image des variables aléatoires.

1.1 Données Brutes

A chaque individu i (i variant de 1 à n, nombre total d'individus) il lui correspond la valeur correspondante $x(i)$ de la variable X ; $y(i)$ de la variable Y ; pour chaque colonne correspondant aux variables $x, y, \ldots$ la nature de la variable statistique peut-être :

- *quantitative continue* : la variable prend des valeurs dans un intervalle de $\mathbb{R}$.
- *quantitative discrète finie* : la variable prend un ensemble fini de valeurs possibles.
- *qualitative non ordonnée* : elle possède des modalités possibles qui peuvent être codées avec une structure d'ordre.

Ce mode de présentation de données est adopté le plus souvent lors d'un traitement informatique des données.

1.2 Tableau statistique d'une Série Statistique Simple

1.2.1 Cas d'une variable discrète

Pour simplifier la présentation d'une variable, on peut regrouper les données brutes en précisant pour chaque valeur distincte de la variable x_k l'effectif correspondant n_k. La nouvelle forme de la série statistique $(x(1), x(2), \ldots, x(n))$, s'écrit quand il y a K valeurs distinctes :

$$(k, x_k, n_k)_{k=1}^{K} \text{ tel que } n = \sum_{k=1}^{K} n_k \tag{1}$$

On définit alors pour chaque valeur x_k d'effectif n_k :

- *La fréquence* : $f_k = n_k/n$
- *L'effectif cumulé croissant* : $s_k = \sum_{j=1}^{k} n_j$

ce qui suppose qu'il y a une structure d'ordre et que les valeurs x_k sont rangées par ordre croissant.

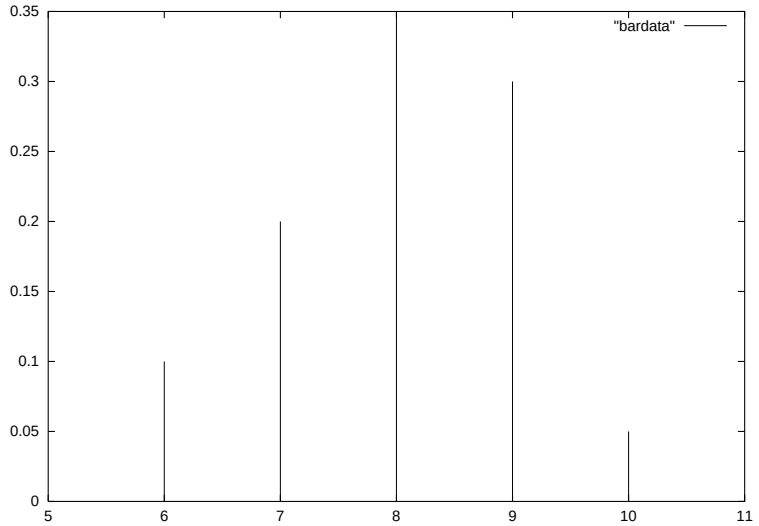
- *La fréquence cumulée croissante* : $\Phi_k = \sum_{j=1}^{k} f_j = \frac{s_k}{n}$.

Ces fréquences peuvent être exprimées en pourcentage.

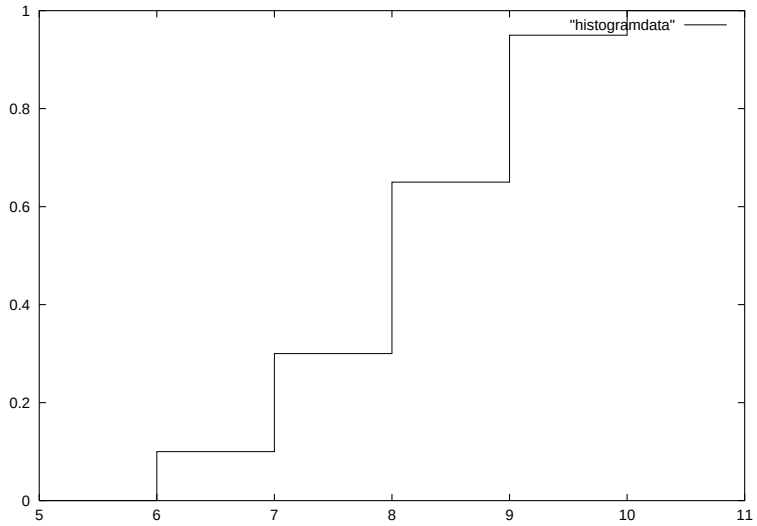
Si dessous, un tableau statistique d'une série simple (une variable quantitative discrète). L'ensemble Ω est constitué par un groupe de 40 enfants et X est l'âge exprimé à l'année près ; après regroupement de la série, on a obtenu le tableau statistique suivant :

k	x_k	n_k	s_k	f_k	Φ_k
1	6	4	4	0,1	0,10
2	7	8	12	0,2	0,30
3	8	14	26	0,35	0,65
4	9	12	38	0,30	0,95
5	10	2	40	0,05	1

Ce tableau est résumé par un diagramme en bâton :



On dessine souvent la courbe des fréquences cumulées ou la *fonction de répartition empirique* $\Phi(x)$ où $\Phi(x) :=$ la proportion de l'effectif inférieure à x. Donc $\Phi(-\infty) = 0$, $\Phi(+\infty) = 1$, $\Phi(x_k) = \Phi_k$ et entre les valeurs de la série statistique, $\Phi(x)$ est constante.



La présentation sous forme de "Stem and Leaf" ou "tige et feuille" introduite par J.W. TUKEY (1977) pour une variable quantitative est très utilisée dans les sorties de logiciels statistiques. On verra des exemples en TP.

1.2.2 Cas d'une Variable Quantitative Continue

On utilise alors un tableau de données en constituant des classes définies par *les bornes, l'amplitude et le centre* de chaque classe. Le découpage en classe doit se faire de façon précise pour obtenir une partition. Le nombre de classes est soit fixé a priori, soit choisi arbitrairement égal environ à $\sqrt{n}$. On obtient le tableau sous la forme :

$$[k, [b_{k-1}; b_k[\,, n_k]_{k=1,K} \text{ tel que } \sum_{k=1}^{K} n_k = n$$

L'amplitude $= a_k = b_k - b_{k-1}$

Dans le cas où les *amplitudes de classes sont différentes* on ne peut pas composer les effectifs. On définit alors :

- *La densité d'effectif* de la classe k :

$$d_k := \frac{n_k}{(b_k - b_{k-1})}.$$

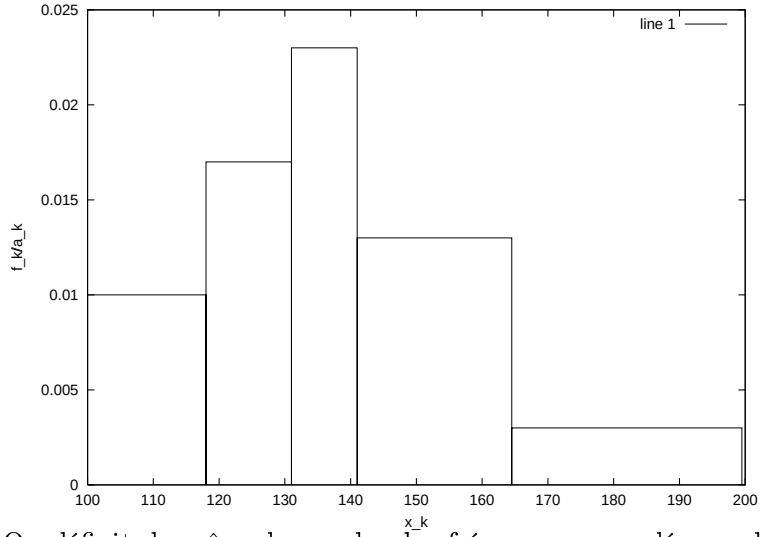
- *La densité de fréquence* :

$$\phi_k := \frac{f_k}{(b_k - b_{k-1})} = \frac{n_k}{n\,(b_k - b_{k-1})}.$$

En générale les logiciels statistiques évitent le problème de densité d'effectif en utilisant une amplitude de classe constante. Si dessous un exemple où les amplitudes de classe ne sont pas constante :

k	classes	a_k	n_k	f_k	$\frac{n_k}{a_k}$	$\frac{f_k}{a_k}$	Φ_k
1	100 ; 120	20	6	0,20	0,3	0,010	0,20
2	120 ; 130	10	5	0,17	0,5	0,017	0,37
3	130 ; 140	10	7	0,23	0,7	0,023	0,6
4	140 ; 160	20	8	0,27	0,4	0,013	0,77
5	160 ; 200	40	4	0,13	0,1	0,003	1,0

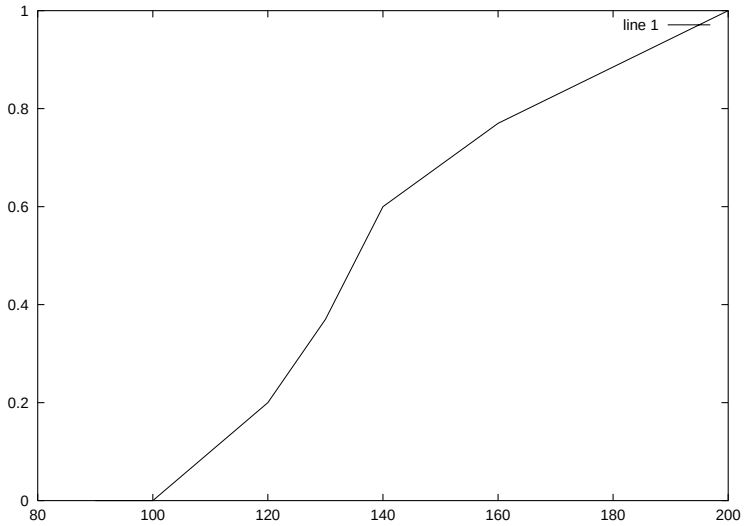
On résume ce tableau dans le cas continue avec un *histogramme* (les bars sont un peu décalés, désolé).



On définit de même la courbe des fréquences cumulées ou bien la fonction de répartition empirique. La fonction de répartition empirique peut être interprétée comme l'intégral de la fonction qui entoure l'histogramme, donc $\Phi(b_1) = 0$, $\Phi(b_2) = f_1$, $\Phi(b_3) = f_1 + f_2$, etc. Entre les b_k la fonction de répartition empirique est affine dans le cas continu.

X^Y	y_1	$\cdots$	y_j	$\cdots$	y_J	
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{l1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

TAB. 1 – Table de Contingence



1.3 Cas d'une Série Double

La série statistique $(i, x(i), y(i))_{i=1}^{n}$ se présente sous la forme d'une *table de contingence* dès que le nombre d'observations n est grand.

Supposons que $[x_1, \ldots, x_i, \ldots, x_I]$ et $[y_1, \ldots, y_j, \ldots, y_J]$ désignent les états des variables X et Y (modalités pour une variable qualitative ou centre de classe pour une variable quantitative)

On note n_{ij} *l'effectif conjoint* qui est le nombre d'individus ayant présenté l'état x_i de la variable X et de l'état y_j de la variable Y.

On note $n_{i.}$ et $n_{.j}$ les effectifs marginaux :

$$n_{i.} = \sum_{j=1}^{J} n_{ij}$$
$$n_{.j} = \sum_{i=1}^{I} n_{ij}$$

Donc

$$n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j} = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ji}.$$

On peut construire le tableau des fréquences associées en divisant chaque effectif par n : $f_{ji} = \frac{n_{ji}}{n}$. On peut calculer les *profils* qui sont des fréquences conditionnelles et qui permettent de juger de l'indépendance des variables X et Y.

- *profil ligne* pour la ligne i :$\frac{n_{ij}}{n_{\cdot j}}$
- *profil colonne* pour la colonne j : $\frac{n_{ji}}{n_{i\cdot}}$.

2 Représentations Graphiques

Les méthodes de représentations graphiques des données différent selon les modes de présentation des données et selon la nature de la variable statistique (quantitative continue ou discrète, qualitative). Un graphique n'a d'intérêt que s'il est bien renseigné :

- Titre clairement défini.
- Explication des sigles et des données utilisées.
- Mention des sources des données.
- Définition claire des grandeurs représentés sur les axes où les échelles seront précisées.

Vous verrez des possibilités proposées par SPSS en TP.

3 Résumés Statistiques d'une Variable Quantitative X

3.1 Caractéristiques de Tendance Centrale

- *Le Mode* : valeur de la variable qui correspond à un effectif ou densité d'effectif maximum.
- *Un mode relatif* : valeur de la variable qui correspond à un effectif ou densité d'effectif maximum local.
- *La Médiane* : valeur de la variable qui partage la série statistique en 2 groupes d'effectifs égaux (50%).
- *La Moyenne (arithmétique)* : elle se calcule suivant le mode de présentation de données :

Pour des données brutes

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dans le cas où il y a des poids ρ_i sur chaque individu on calcule la moyenne pondérée :

$$\overline{x} = \sum \rho_i x_i$$

avec $\sum_{i=1}^{n} \rho_i = 1$.

Pour des données regroupées en K classes :

$$\overline{x} = \sum_{k=1}^{K} f_k x_k.$$

Propriété de la moyenne : Si on translate les données x_i d'une valeur a, c'est à dire : $y_i = x_i + a$, alors $\overline{y} = \overline{x} + a$.

Remarque 1
La moyenne $\overline{x}$ peut ne pas être un paramètre significatif, par exemple, quand la série est bimodale ou dissymétrique (valeurs extrêmes).

3.2 Caractéristiques de Dispersion

- *L'étendue* : $e = x_{\max} - x_{\min}$.
- *L'interquartile :* Q_1,Me, Q_3 sont les 3 quartiles qui partagent la série statistique en 4 groupes d'effectifs égaux. On définit *l'intervalle interquartile* : $[Q_1, Q_3[$ (et sa longuer : $Q_3 - Q_1$) qui contient les 50% des valeurs qui se trouvent au "centre" de la série.
- *La variance :* elle mesure la dispersion des valeurs x_i de la variable X par rapport à la moyenne $\overline{x}$. Elle se calcule suivant le mode de présentation de données :

Pour les données brutes :

$$\text{Var } X = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

ou encore par une formule équivalente :

$$\text{Var } X = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^2\right] - \overline{x}^2.$$

Dans le cas où il y a des poids p_i, elle se calcule par :

$$\text{Var } X = \sum_{i=1}^{n} p_i(x_i - \overline{x})^2 = \sum_{i=1}^{n} p_i(x_i)^2 - \overline{x}^2.$$

Pour des données regroupées :

$$\text{Var } X = \sum_{k=1}^{K} f_k(x_k - \overline{x})^2 \text{ où } f_k = \frac{n_k}{n},$$

ou encore par la formule équivalente :

$$\text{Var } X = \left[\frac{1}{n}\sum_{k=1}^{K} n_k(x_k)^2\right] - (\overline{x})^2.$$

Propriété de la variance : Var $aX = a^2$Var X.
Soit $\mu_k := \sum_{i=1}^{n}(x_i - \bar{x})^k$.

– *L'écart type :* il a la même unité que la variable X et est défini par :

$$\sigma_x = \sqrt{\text{Var } X} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

On utilise aussi *la déviation standard* :

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

– *Le coefficient de variation* : $c_X = \frac{\sigma_X}{x}$.
– *Le coefficient de skewness* : $(\mu_3)^2/(\mu_2)^3$ (mesure d'asymétrie).
– *Le coefficient de kurtosis* : $\mu_4/(\mu_2)^2$ (mesure d'aplatissement).

4 Liaison Entre 2 Variables

4.1 Cas de 2 Variables Quantitatives X et Y

A partir de la Série double $(i, x(i), y(i))_{i=1}^n$ on calcule :
– *la covariance :*

$$\begin{aligned}\text{Cov } (X,Y) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i y_i\right] - \overline{xy}\end{aligned}$$

Propriété : Var $(x+y)$ = Var x+Var $y + 2\text{Cov}(X,Y)$.
La covariance n'a pas de signification concrète car c'est le produit de 2 unités différentes (celles de X et Y). Elle est une caractéristique conjointe de dispersion de X et Y.
– *le coefficient de corrélation linéaire de Bravais Pearson*

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

$r(X,Y)$ n'a pas d'unité et est compris entre -1 et +1. Si $r = +1$, il existe une liaison linéaire entre X et Y et on peut écrire : $Y = aX + b$ avec $a > 0$.
Si $r = -1$, il existe une liaison *linéaire* entre X et Y et on peut écrire : $Y = aX + b$ avec $a < 0$.
Si r et voisin de 0, il n'y a pas de liaison linéaire entre X et Y. On dit que X et Y ne sont pas corrélées linéairement.
– *La régression linéaire simple.* On suppose qu'il existe une relation causale entre X et Y, par exemple X est cause de Y ou bien Y est une fonction de X, $y_i = f(x_i)$. On suppose de plus que f est de la forme : $y_i = f(x_i) = ax_i + b + \epsilon_i$. On peut alors réaliser la régression linéaire de Y en X en estimant les paramètres de la droite de régression a et b de $Y = aX + b$.

Le problème qu'on se pose est donc d'estimer les paramètres de la droite de régression a et b de $Y = aX + b$ avec les hypothèses :

$$y_i = ax_i + b + \epsilon_i \text{pour } i = 1, ..., n$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - ax_i - b]^2 \text{ est minimal}$$

On obtient les estimations $\widehat{a}$ et $\widehat{b}$ des paramètres a et b par la méthode des moindres carrés :

$$\widehat{a} = \frac{\text{Cov } (X,Y)}{\text{Var } X}$$
$$\widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

Remarque 2
La droite ajustée a pour équation :

$$Y = \widehat{a}(x - \overline{x}) + \overline{y.}$$

Elle passe par le centre de gravité du nuage de points : $(\overline{x}, \overline{y})$. L'estimation de la pente de la droite de régression est égale aussi à : $\widehat{a} = r(x, y)\frac{\sigma_y}{\sigma_x}$.

À partir d'une table de contingence, on peut calculer :
- *La moyenne marginale* :

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{I} n_{i.} x_i.$$

- *La variance marginale* :

$$\sigma^2(X) := \frac{1}{n} \sum_{i=1}^{I} n_{i.} (x_i - \bar{x})^2.$$

- *Les moyennes conditionnelles* :

$$\bar{x}_j := \frac{1}{n_{.j}} \sum_{i=1}^{I} n_{ij} x_i.$$

- *Les variances conditionnelles* :

$$\sigma_j^2(X) := \frac{1}{n.j} \sum_{i=1}^{I} n_{ij} (x_i - \bar{x}_j)^2.$$

- *La variance des moyennes conditionnelles de X* :

$$\sigma^2(\bar{x}_j) := \frac{1}{n} \sum_{j=1}^{J} n_{.j} (\bar{x}_j - \bar{x})^2.$$

– *La moyenne des variances conditionnelles de* X :

$$\overline{\sigma_j^2(x)} := \frac{1}{n}\sum_{j=1}^{J} n_{.j}\sigma_j^2(X).$$

– *Les quotients (ou rapports) de corrélation* $\eta^2_{X/Y}$ et $\eta_{Y/X}$:

$$\begin{aligned}\eta^2_{X/Y} : &= \frac{\sigma^2(\bar{x}_j)}{\sigma^2(X)}\\ \eta^2_{Y/X} : &= \frac{\sigma^2(\bar{y}_i)}{\sigma^2(Y)}\end{aligned}$$

Les quotients de corrélations sont utilisés pour chercher une relation fonctionnelle (pas forcément linéaire) entre X et Y.

4.2 Cas d'une Série Chronologique Simple

Les indices :
– Indices élémentaires : $I_{\frac{n}{0}} = P_n/P_0$
– L'indice de Laspeyre :

$$L_{\frac{n}{0}} := \frac{\sum_{i=1}^{N} p_n^i q_0^i}{\sum_{i=1}^{N} p_0^i q_0^i} = \sum_i \rho_i \frac{p_n^i}{p_0^i}.$$

– L'indice de Paasche :

$$P_{\frac{n}{0}} := \frac{\sum_{i=1}^{N} p_n^i q_n^i}{\sum_{i=1}^{N} p_0^i q_n^i} = \frac{1}{\sum_i \rho_i \frac{p_0^i}{p_n^i}}.$$

C'est le cas d'une série double où les données présentées sous la forme brute s'écrivent : $(t, x_t)_{t=1}^{T}$. On définit

$$\begin{aligned}T_t &:= \text{trend ou tendance général} &(2)\\ s_t &:= \text{variations saisonières} &(3)\\ c_t &:= \text{variations cycliques} &(4)\\ \epsilon_t &:= \text{variations aléatoires.} &(5)\end{aligned}$$

On considère le modèle :

$$x_t = T_t + s_t + c_t + \epsilon_t.$$

Dans certains cas,

$$x_t = T_t s_t c_t \epsilon_t$$

Dans lequel cas on considère Log x_t pour retrouver le premier modèle. On suppose s_t périodique de moyenne nulle, c'est á dire qu'il existe $k = 2p + 1$ (pour simplifier les formules) tel que

$$s_t = s_{t+ik},\ i = 1, 2, \ldots$$
$$\sum_{i=-p}^{p} s_{t+i} = 0\ \forall t.$$

On défini alors les *moyennes mobiles* sur k périodes :

$$y_t := \frac{x_{t-p} + x_{t-p+1} + \cdots + x_t + \cdots + x_{t+p}}{2p+1} \tag{6}$$

pour $k = 2p + 1$

$$y_t := \frac{x_{t-p}/2 + x_{t-p+1} + \cdots + x_{t+p-1} + x_{t+p}/2}{2p} \tag{7}$$

pour $k = 2p$.

Pour faire des *prévisions* :

- Première étape : on fait un ajustement sur y_t,

$$\hat{y}_t = \hat{a}t + \hat{b}.$$

- Deuxième étape : on peux estimer les effets saisoniers

$$\hat{s}_t = x_t - \hat{T}_t = x_t - \hat{y}_t$$

- Troisième étape : on a estimer le même parametre plusieurs fois, donc on fait une moyenne :

$$\bar{\hat{s}}_t := \frac{1}{I} \sum_{i=1}^{I} \hat{s}_{t+ik}.$$

Donc :

$$\hat{x}_t = \hat{y}_t + \bar{\hat{s}}_t.$$

4.3 Cas de 2 Variables Qualitatives

Les données se présentent sous forme d'un tableau de contingence (cf 1.4). On mesure la liaison entre les 2 variables qualitatives X et Y par le χ^2 de contingence :

$$\chi^2 = \sum_{l=1}^{L} \sum_{c=1}^{C} \frac{(n_{lc} - \widehat{n}_{lc})^2}{\widehat{n}_{lc}}$$

où

$$\begin{array}{lcl} n_{lc} & := & \text{effectif conjoint observé} \\ \widehat{n}_{lc} & := & \frac{n_{l.}n_{.c}}{n} \end{array}$$

On remarque que $\hat{n}_{lc}$ est l'effectif théorique vrai si les 2 variables sont indépendantes. Le χ^2 de contingence se calcule de manière équivalente par la formule :

$$\chi^2 = n[(\Sigma_{l=1}^{L}\Sigma_{c=1}^{C}\frac{n_{lc}^2}{n_{l.}n_{.c}}) - 1].$$

Remarque 3

- Si $\chi^2 = 0$, les variables X et Y ne sont pas liées.
- χ^2 dépend de la taille n de l'échantillon ainsi que de L et C.

Si $C + L \geq 2$, on obtient un encadrement de χ^2 :

$$0 \leq \chi^2 \leq n[\min(L, C) - 1].$$

Cet encadrement permet de conclure, en fonction de la valeur χ^2 si les variables X et Y sont plus ou moins liées.

4.4 Paramètres de Concentration

Soit $(x_k, e_k)_{k=1}^{K}$ une série statistique discrète, telle que pour tout indice k, x_k soit positif et notons n son effectif total. Soit

$$M = \sum_{k=1}^{K} n_k x_k. \tag{8}$$

Soit P_k, $k \in [1, K]$ le point de coordonnées (Φ_k, m_k) dans un repère cartésien plan, défini par :

$$\Phi_k := \frac{\sum_{j=1}^{k} n_j}{n}, m_k := \frac{\sum_{j=1}^{k} n_j x_j}{M} \tag{9}$$

Définition 1

On appelle courbe de Lorentz (ou de Gini) ou courbe de concentration de cette série, la ligne polygonale $OM_1M_2 \ldots M_I$ où O désigne l'origine du repère.

Pour tout indice $i \in [1, I]$ soit

$$\phi_i := \frac{n_i x_i}{M} \tag{10}$$

Définition 2

On appelle *médiale* de la série $(x_i, n_i)_{i=1}^{I}$ la médiane de la série (x_i, ϕ_i).