Introduction à la Statistique Descriptive

Mohamed El Omari
Enseignant Chercheur,
Spécialité: Statistique et Probabilités

Ancien Inspecteur Pédagogique

Faculté Polydisciplinaire de Sidi Bennour.

December 23, 2022

Outline

La statistique descriptive a pour but d'étudier une population à partir de données exhaustives. Cette description se fait à travers

- La collecte des données;
- La présentation et l'analyse des données;
- Le résumé et l'interprétation des données.

1.1. Population, échantillon et variable statistique

La population statistique est l'ensemble des éléments sur lesquels porte l'étude.

Les éléments de la population sont appelés **individus statistiques** ou unités statistiques.

On parle de **recensement** lorsque l'on fait une étude exhaustive d'une population.

Lorsqu'on n'étudie qu'une sous-population, on parle alors d'**échantillon**.

Une variable statistique (ou caractère statistique), notée X, est une application définie sur une population statistique et à valeurs dans un ensemble M, appelé ensemble des modalités.

Les modalités correspondent aux valeurs possibles de la variable statistique.

Une variable statistique définit une partition sur une population, chaque individu appartenant à une et une seule modalité.

Exemple 1:

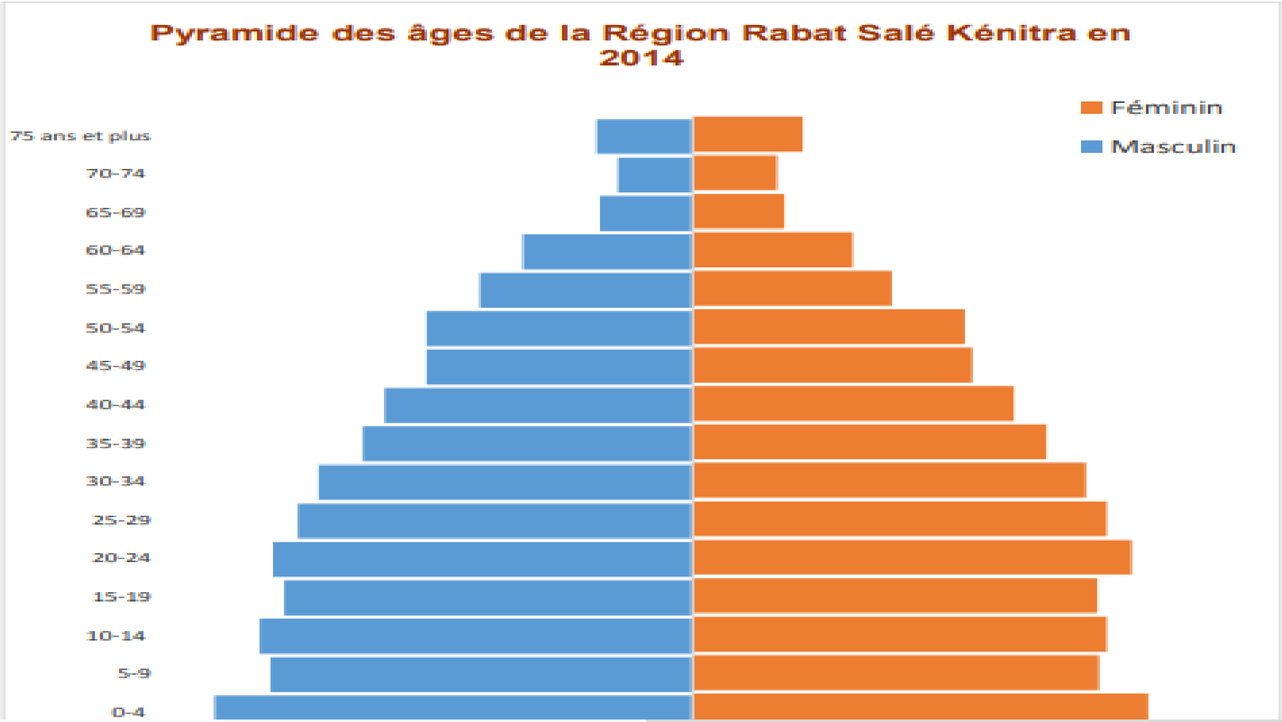
Le personnel d'une entreprise peut être décrit selon divers caractères : .. ..

Un lot de pièce mécanique peut être décrit suivant : .. ..

Exemple 2:

Considérons les données graphiques suivantes :

Source: www.hcp.ma/region-rabat/attachment/775960/



1.2. Types de variable statistique

Variable qualitative : Une variable est dite qualitative quand les modalités sont des catégories (non mesurables).
Par exemple : ....

Variable qualitative nominale : Une variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées.
Variable qualitative ordinale : Une variable est dite qualitative ordinale quand les modalités peuvent être ordonnées. Le fait de pouvoir ou non ordonner les modalités est parfois discutable.
Par exemple : ....

Variable quantitative : Une variable est dite quantitative si toute ses valeurs possibles sont numériques.
Par exemple : ....

> **Variable quantitative discrète :** Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable.
> **Variable quantitative continue :** Une variable est dite continue, si l'ensemble des valeurs possibles est continu.
> Par exemple : ....

1.3. Effectif, fréquence et pourcentage

Soit X une variable statistique dont l'ensemble des modalités est $\mathcal{M} = \{x_1, x_2, \cdots, x_n\}$. On appelle **effectif d'une modalité** x_j, le nombre de fois que cette modalité apparaît. On note n_j l'effectif de la modalité x_j. **La fréquence** f_j de cette modalité est l'effectif divisé par le nombre d'unités d'observation n(effectif total).

On écrit: $f_j = \frac{n_j}{n}$.

Le pourcentage p_j d'une cette modalité x_j est son effectif multiplié par 100. On écrit: $p_j = f_j \times 100$.

La somme des effectifs est égale à : ... .

La somme des fréquences est égale à : ... ...

La somme des pourcentage est égale à : ... ...

1.4. Représentation des données statistiques

1.4.1 Représentation numérique

1.4.2 Représentation graphique

Exercice 1: Pour évaluer les performances des élèves d'une classe collégiale en mathématiques, une épreuve a été faite. On a obtenu la série suivante:
8 - 11 - 13 - 5 - 8 - 14 - 6 - 12 - 5 - 10
16 - 7 - 12 - 13 - 8 - 13 - 8 - 7 - 13 - 13
9 - 17 - 10 - 13 - 6 - 13 - 7 - 14

- Déterminer la population étudiée.
- Quelle est la variable statistique? De quel type est-elle?
- Comment peut-on organiser les données?

Exercice 2: On interroge 50 personnes sur leur dernier diplôme obtenu. La codification a été faite selon le tableau suivant:

Dernier diplôme obtenu	**Modalité** x_j
Sans diplôme	Sd
Primaire	P
Secondaire	Se
Supérieur non-universitaire	Su
Universitaire	U

Déterminer la population, le caractère étudié ainsi que la nature de ce caractère.

compléter le tableau suivant.

Modalité x_j	**Effectif**	**Fréquence**	**Pourcentage**
Sd	.	0,08	.
P	11	.	.
Se	.	.	28
Su	9	.	.
U	.	0,24	.
Total	50	.	.

Pourquoi utiliser un graphique pour présenter des données?

Les graphiques:

- se lisent et sont compris rapidement,
- montrent les faits les plus importants,
- facilitent la compréhension des données,
- peuvent convaincre le lecteur,
- aident le lecteur à se souvenir des données.

Remarques:

- Comme les tableaux statistiques, les graphiques permettent de réduire, de synthétiser les données brutes d'une série statistique.
- Les graphiques donnent une synthèse visuelle des données.
- Les types de représentation graphiques sont toutefois différents selon la nature et le type du caractère étudié, il est donc indispensable, avant de choisir un mode de représentation, de réfléchir sur la nature des modalités.

Règles à retenir lors de la conception d'un graphique:

Un bon graphique:

comporte un titre et des étiquettes,
illustre les faits avec précision,
attire l'attention du lecteur,
complète ou démontre les arguments du texte,
est clair et simple,
montre les données sans modifier leur message,
montre clairement toute tendance ou différence dans les données,
est exact en ce qui a trait à l'aspect visuel (par exemple, pour deux valeurs, une de 15 et l'autre de 30, la deuxième devrait apparaître comme le double de la première).

Choisir le bon graphique:

Variable quantitative:

- Diagramme en bâtons et courbe cumulative (dans le cas discret).
- Histogramme (avec polygone) et courbe cumulative (dans le cas continu).
- Diagramme chronologique ou linéaire par morceaux (dans le cas d'une série temporelle).

Variable qualitative:

- Diagramme à barres (empilées).
- Digramme circulaire, parfois appelé diagramme en secteurs (préférable pour une variable nominale avec une distribution donnée en pourcentages).

2. Caractéristiques numériques d'une série quantitative

Un paramètre est un nombre qui décrit une caractéristique de la population étudiée. Citons, à titre d'exemples, le mode et la moyenne. Dans le cas d'un recensement, tous les paramètres peuvent être calculés.

2.1. Paramètres de position

- Le mode
- La moyenne
- La médiane
- Le quantile

Le mode:

Le mode est la valeur distincte correspondant à l'effectif le plus élevé; il est noté x_M.

Remarques:

- Le mode peut être calculé pour tous les types de variable, quantitative et qualitative.
- Le mode n'est pas nécessairement unique.
- Quand une variable continue est découpée en classes, on peut définir une classe modale.

La moyenne:

La moyenne ne peut être définie que sur une variable quantitative. La moyenne est la somme des valeurs observées divisée par leur nombre; elle est noté $\overline{x}$:

$$\begin{aligned}\overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n}, &(1)\\ &= \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{n_1 + n_2 + \cdots + n_p}, &(2)\end{aligned}$$

où $x_1, x_2, \cdots, x_p$ sont des valeurs distinctes ayant pour effectifs $n_1, n_2, \cdots, n_p$, respectivement. $n = n_1 + n_2 + \cdots + n_p$ est l'effectif total.

La moyenne n'est pas nécessairement égale à la valeur d'une des données.

La médiane:

La médiane d'une série statistique, dénotée $x_{1/2}$, est une valeur telle que 50% des observations lui sont supérieures et 50% lui sont inférieures. Si $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ sont les données en ordre croissant alors

$$\begin{cases} x_{1/2} = x_{\left(\frac{n+1}{2}\right)}, & \text{si } n \text{ est impair,} \\ x_{1/2} = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{si } n \text{ est pair.} \end{cases}$$

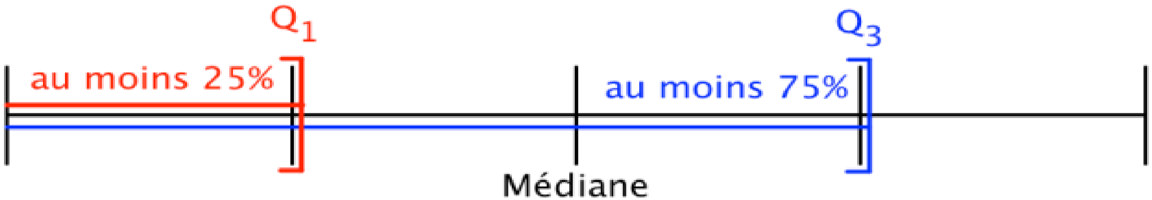
Remarques:

- La médiane peut être calculée sur des variables quantitatives et sur des variables qualitatives ordinales.
- Formellement, la médiane est la valeur pour laquelle la fréquence relative cumulée est égale à $1/2$.
- Dans le cas où les données sont regroupées sous forme de classes, on parle de classe médiane. Pour obtenir une valeur plus précise de la médiane, on procède à une interpolation linéaire.

Le quantile:

La notion de quantile d'ordre p (où $0 < p < 1$) généralise la médiane. Formellement, un quantile est la valeur x_p pour laquelle la fréquence relative cumulée est égale à p.

Quand $p = k/4$, $k = 1, 2, 3$ on parle de premier, deuxième et troisième quartiles Q_1, Q_2 et Q_3. La médiane coïncide avec le deuxième quartile.



Méthode pour le calcul des quartiles:

Utiliser la médiane pour diviser les données en deux parties égales. Ne pas inclure la médiane dans les deux sous-ensembles obtenus, si elle est théorique.

Poser : $Q_2 =$ médiane de la population.

Poser

$Q_1 =$ médiane du sous-ensemble des valeurs inférieures à Q_2.

$Q3 =$ médiane du sous-ensemble des valeurs supérieures à Q_2.

2.2. Paramètres de dispersion

- L'étendue
- L'intervalle interquartile
- La variance et l'écart-type
- Le coefficient de variation

L'étendue:

L'étendue est simplement la différence entre la plus grande et la plus petite valeur observée.

L'écart (ou la distance) interquartile:

La distance interquartile est la différence entre le troisième et le premier quartile. On écrit

$$IQ = x_{3/4} - x_{1/4}.$$

La variance:

La variance est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations:

$$s_x^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}, \quad (3)$$

$$= \frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+\cdots+n_p(x_p-\overline{x})^2}{n_1+n_2+\cdots+n_p}, \quad (4)$$

où $x_1, x_2, \cdots, x_p$ sont des valeurs distinctes ayant pour effectifs $n_1, n_2, \cdots, n_p$, respectivement. La variance peut aussi s'écrire

$$s_x^2 = \frac{x_1^2+x_2^2+\cdots+x_n^2}{n} - \overline{x}^2. \quad (5)$$

La variance corrigée:

La variance corrigée est définie par:

$$S_x^2 = \frac{n}{n-1} s_x^2. \tag{6}$$

L'écart-type est la racine carrée de la variance: $s_x = \sqrt{s_x^2}$.

3. Distribution statistique à deux variables

3.1. Représentation graphique de deux séries statistiques

3.2. Comparaison de deux séries statistiques

3.3. Étude de la corrélation entre deux variables statistiques

3.3. Théorie des indices

3.1. Représentation graphique de deux séries statistiques

Les graphiques:

- se lisent et sont compris rapidement,
- **montrent les faits les plus importants,**
- facilitent la compréhension des données,
- peuvent convaincre le lecteur,
- **aident le statisticien à comparer les séries statistiques,**
- aident le lecteur à se souvenir des données.

Les graphiques à discuter:

- Diagramme chronologique (ou linéaire par morceaux); Diagramme en bâtons.
- Histogramme.
- Diagramme à barres; Diagramme à barres empilées.
- Diagramme circulaire (ou diagramme en secteurs).
- Boîte moustaches (ou diagramme en boîte).
- Nuage de points.

3.2. Comparaison de deux séries statistiques

Pour comparer deux séries statistiques on compare leur paramètres de position et de dispersion.

L'écart-type, malgré sa pertinence dans la mesure de la dispersion d'une série statistique, possède un inconvénient majeur:

(a) Il est exprimé dans l'unité de la variable à laquelle il se rapporte.

(b) Il est alors impossible de comparer les dispersions de deux séries statistiques ayant un lien entre elles et dont les valeurs s'expriment dans des unités différentes.

Pour comparer la dispersion de deux séries qui ne sont pas exprimées dans les mêmes unités, on utilise le coefficient de variation noté **CV**.

Le **CV** d'une série statistique $\{x_1, \cdots, x_n\}$ de moyenne $\overline{x}$ et d'écart type s_x est défini par

$$\mathbf{CV} = \frac{s_x}{\overline{x}}.$$

Une pratique empirique courante est de considérer que la série possède une variabilité significative si $\mathbf{CV} > 0.15$.

Si $\mathbf{CV} \leq 0.15$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de toute la série.

La statistique bivariée a pour but d'étudier les relations éventuelles entre deux variables statistiques X et Y observées sur une même population. Les données statistiques forment une série des couples de valeurs prises par les deux variables X et Y. On écrit

$$(x_1, y_1), \cdots, (x_m, y_m).$$

Chacune des variables X et Y peut être, soit quantitative, soit qualitative. On examine les deux cas suivants.

1er Cas: Les deux variables sont quantitatives.

2ème Cas: Les deux variables sont qualitatives.

3.3. Étude de la corrélation entre deux variables statistiques

3.3.1. Le coefficient de corrélation de deux variables quantitatives:

La statistique la plus utilisée dans le contexte de deux séries numériques est la corrélation. Pour la définir, la notion de **covariance** doit être introduite.

On appelle covariance des séries numériques $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$ la valeur

$$
\begin{aligned}
s_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}, \\
&= \frac{\sum_{i=1}^{n} x_i y_i}{n} - (\bar{x} \cdot \bar{y}) .
\end{aligned}
$$

Le coefficient corrélation (linéaire) des séries numériques $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$ est défini par

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Interprétation du coefficient de corrélation

On a: $-1 \leq \rho_{xy} \leq 1$.

Si $\rho_{xy} = +1$ alors il y a corrélation parfaite (positive) entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont alignés sur une droite de pente positive.

Si $\rho_{xy} = -1$ alors il y a corrélation parfaite (négative) entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont alignés sur une droite de pente négative.

Si $\rho_{xy} = 0$ alors il n'y a pas de corrélation entre les x_i et les y_i. Les points $M_i(x_i, y_i)$ sont distribués “au hasard”dans le plan.

3.3.2 La droite de régression linéaire

Le but de la régression simple est d'expliquer une variable Y à l'aide d'une variable X. La variable Y est appelée variable dépendante, ou variable à expliquer et la variable X est appelée variable indépendante, ou variable explicative.

Remarques:

- Il est indispensable de commencer par l'étude corrélation entre X et Y avant de chercher une ligne de régression entre X et Y.
- La régression diffère de l'analyse de la corrélation où toutes les variables jouent un rôle symétrique (pas de variable dépendante versus indépendante).
- Toutefois, tout comme dans le contexte de l'analyse de la corrélation, il faut être prudent lorsqu'on formule des relations de causalité! L'existence d'une relation entre X et Y n'implique pas nécessairement une relation de causalité entre elles.

Un modèle de régression linéaire simple est défini par l'équation:

$$\boxed{Y = aX + b + \text{Erreur.}}$$

Les paramètres a et b sont inconnus et estimés par la méthode des moindres carrés.

$$\begin{aligned} \widehat{a} &= \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \\ \widehat{b} &= \overline{y} - \widehat{a}\,\overline{x}, \end{aligned}$$

où $\overline{x}$ et $\overline{y}$ désignent les moyennes des séries d'observations de obtenues à partir des variables X et Y, respectivement.

La droite (Δ) d'équation $y = \widehat{a}x + \widehat{b}$ est appelée la droite de régression linéaire .

Le point $M(\overline{x}, \overline{y})$ est appelé point moyen et appartient à la droite (Δ).

Le couple $\left(\widehat{a}, \widehat{b}\right)$ minimise la somme des carrés des résidus e_i:

$$\sum_{i=1}^{n} e_i^2 \quad = \quad \sum_{i=1}^{n} (y_i - ax_i - b)^2, \quad a, b \in \mathbb{R}.$$

Un des buts de la régression est de proposer des **prédictions** pour la variable à expliquer Y lorsque nous avons de nouvelles valeurs de X.

Pour une nouvelle valeur x_0, on prévoit la valeur y_0 associée

$$y_0 = \widehat{a}x_0 + \widehat{b}.$$

Exemple 1: Le tableau suivant donne l'évolution du nombre de spectateurs (en millions) dans les salles de cinéma en France sur une période de 7 ans.

Année	1989	1993	1994	1995	1996
Rang (X_i)	0	4	5	6	7
Nombre de spectateurs (Y_i)	120,9	132,7	124,5	130,2	136,3

Représenter la série statistique (X_i, Y_i) par un nuage de points, et préciser le point moyen M.

Donner une équation de la droite de régression (Δ) de Y en X par la méthode des moindres carrés.

Si l'évolution constatée s'est poursuivie jusqu'à la fin du XX^e siècle, donner une estimation du nombre de spectateurs dans les salles de cinéma en France en l'an 2000.

Exemple 2: On a relevé pour chacune des années t de 1920 à 1929, numérotées de 1 à 10, la température moyenne X des mois d'été (en degrés centigrades) et la mortalité infantile Y (nombre de décès d'enfants de moins d'un an pour 1000 naissances vivantes).

t	1	2	3	4	5	6	7	8	9	10
X	15,9	18,8	15,4	18	14,6	16,2	17,9	16,5	18,1	19,8
Y	98	116	87	96	85	89	97	83	91	95

Étudier la corrélation entre X et Y.

Déterminer la droite de régression linéaire de Y en X.

3.3.3 Tableau de contingence et tableau des fréquences

On considère deux variables statistiques X et Y. Les données observées obtenues à partir de X et Y peuvent être regroupées sous la forme d'un tableau croisé (appelé tableau de contingence).

X \ **Y**	y_1	$\cdots$	y_j	$\cdots$	y_q	**Effectif marginal**
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$						
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.} = \sum_{j=1}^{q} n_{ij}$
$\vdots$						
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
Effectif marginal	$n_{.1}$	$\cdots$	$n_{.j} = \sum_{i=1}^{p} n_{ij}$	$\cdots$	$n_{.q}$	$n = n_{..}$

3.3.4 Profils lignes et profils colonnes

Le profil ligne "i " est la ligne des fréquences conditionnelles

$$\mathbf{L}_i \qquad \left(f_{1/i} = \frac{n_{i1}}{n_{i.}} \quad \cdots \quad f_{j/i} = \frac{n_{ij}}{n_{i.}} \quad \cdots \quad f_{q/i} = \frac{n_{iq}}{n_{i.}} \right)$$

Le profil marginal correspondant est:

$$\mathbf{M} \qquad \left(f_{.1} = \frac{n_{.1}}{n} \quad \cdots \quad f_{.j} = \frac{n_{.j}}{n} \quad \cdots \quad f_{.q} = \frac{n_{.q}}{n} \right)$$

De même on définit le profil colonne **"j "** et son profil marginal correspondant.

Remarques:

On obtient le **tableau des profils** lignes (resp. colonnes) par la division des lignes (resp. colonnes) du tableau contingence par leur effectifs marginaux.

On obtient le tableau des **fréquences conjointes** et les **fréquences marginales** par la division des lignes du tableau contingence par l'effectif total.

Si les deux variables qualitatives X et Y étaient indépendantes, les profils lignes seraient tous identiques, et donc identiques au profil marginal correspondant. On trouve

$$n_{ij} = \frac{n_{i.}n_{.j}}{n}$$

L'indépendance est un cas extrême que l'on rencontre rarement à l'état pur dans la pratique.

On peut cependant mesurer l'intensité de la dépendance entre deux caractères qualitatifs X et Y.

On définit le coefficient d'association χ^2 par

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}.$$

Il est d'autant plus grand que la liaison entre X et Y est forte.

Le problème du χ^2 est qu'il dépend de la taille de la population n et des nombres de modalités p et q.

Que peut alors signifier grand dans ce cas?

Exercice: Prouver les assertions suivantes:

(a) Si les deux caractères qualitatifs X et Y sont indépendants, alors $f_{j/i} = f_{.j}$ et $f_{ij} = f_{i.}f_{.j}$.

(b) Les caractères qualitatifs X et Y sont indépendants si et seulement si $\chi^2 = 0$.

(c)

$$\chi^2 = n\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right] \quad \text{et} \quad \chi^2 \leq n\,(p-1\ ,\ q-1)\,.$$

Qu'est-ce qu'un indice?

Un indice est la valeur d'une grandeur par rapport à une valeur de référence. Prenons l'exemple du tableau ci-dessous contenant le prix p_t (exprimé en MAD) d'un bien de consommation de 2000 à 2006.

Année	2000	2001	2002	2003	2004	2005	2006
t	0	1	2	3	4	5	6
p_t	200	230	240	280	300	350	400

On définit l'**indice simple** par

$$I(t/t') \quad = \quad 100 \times \frac{p_t}{p_{t'}}, \quad t, t' = 0, 1, \cdots, 6.$$

L'indice $I(t/t')$ est sans dimension et indépendant du choix des unités.

Considérons un indice quelconque $I(t/0)$ (avec 0 est le temps le temps de référence). On dit que cet indice possède les propriétés de

- réversibilité si $I(t/0) \times I(0/t) = 100^2$.
- identité si $I(t/t) = 100$.
- transitivité si $I(t/u) \times I(u/s) = 100 \times I(t/s)$.

Il est facile de montrer que ces trois propriétés sont satisfaites pour un indice simple.

Un indice synthétique est une grandeur d'un ensemble de biens par rapport à une année de référence.

On ne peut pas construire un indice synthétique en additionnant simplement des indices simples.

Il faut, en effet, tenir compte des quantités achetées.

Il existe deux méthodes fondamentales pour calculer les indices de prix, l'indice de **Paasche** et l'indice de **Laspeyres**.

L'indice de **Laspeyres** est défini par

$$L(t/0) = 100 \times \frac{\sum_{i=1}^{n} q_0^i p_t^i}{\sum_{i=1}^{n} q_0^i p_0^i},$$

où

- p_t^i représente le prix du bien de consommation i au temps t.
- q_t^i représente la quantité de bien i consommée au temps t.

L'indice de **Paasche** est défini par

$$P(t/0) \quad = \quad 100 \times \frac{\sum_{i=1}^{n} q_t^i p_t^i}{\sum_{i=1}^{n} q_t^i p_0^i},$$

- L'indice de Laspeyres et lindice de Paasche ne possèdent ni la propriété de transitivité ni de réversibilité.
- L'indice de Laspeyres est en général plus grand que l'indice de Paasche.

L'indice de Fisher

Fisher a proposé d'utiliser un compromis entre l'indice de Paasche et de Laspeyres en calculant simplement la moyenne géométrique de ces deux indices.

$$F(t/0) \quad = \quad \sqrt{L(t/0) \times P(t/0)}.$$