

Chapitre 6
Tendances et corrélations
Dernière mise à jour le 3 juillet 2009

0 – [Introduction](#)

1 – [La détermination de la tendance d'une série chronologique](#)

- A – [Détermination graphique](#)
- B – [Détermination par la méthode des points extrêmes](#)
- C – [Détermination par la méthode des moindres carrés \(MCO\)](#)

2 – [L'étude de la liaison statistique entre deux variables](#)

- A – [L'exemple d'une fonction de demande](#)
- B – [L'équation de régression linéaire](#)
- C – [Le coefficient de détermination](#)

3 – [Le test d'indépendance du Khi-carré de PEARSON](#)

- A – [Introduction](#)
- B – [Exemple d'utilisation](#)
- C – [Etapas du test d'indépendance du khi-carré](#)

4 - [Résumé](#)

Exercices corrigés (PDF) :

TD5 2008 : [Sujets](#) - [Corrigés](#)

TD5 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#) (lien vers page HTML)

0 – Introduction

Ce chapitre est consacré à l'étude des outils de base permettant d'évaluer l'existence d'une liaison statistique entre deux dimensions d'une population.

Lorsque les deux dimensions sont quantitatives, la méthode de **l'ajustement linéaire**, aussi appelé « méthode des Moindres Carrés Ordinaires (MCO) » est la plus fréquemment employée. Ci-après nous l'appliquerons successivement :

- A l'étude de la tendance d'une série chronologique (section 1).
- À la mise en évidence d'une corrélation entre deux variables (section 2).

Pour l'étude de la liaison entre deux dimensions dont **l'une au moins n'est pas quantitative**, c'est le **test d'indépendance du Khi-carré**, étudié en section 3, qui remplace l'ajustement linéaire¹.

¹ Le test du Khi-carré est souvent présenté, à juste titre, comme un test d'indépendance entre deux dimensions – qu'il s'agisse de variables ou de caractères ou une combinaison des deux – **quand les données sont regroupées en catégories**. Si l'on se réfère à cette définition, les données distribuées par modalités (et à fortiori par valeurs) semblent exclues. Mais, bien entendue, elles ne le sont pas.

1 – La détermination de la tendance d'une série chronologique

Lorsque l'on veut étudier la liaison statistique entre une variable quelconque et le temps (qui est une aussi une variable) , la méthode la plus appropriée est celle des moindres carrés ordinaires. On présentera cependant, au préalable, deux méthodes plus intuitives.

A – Détermination graphique

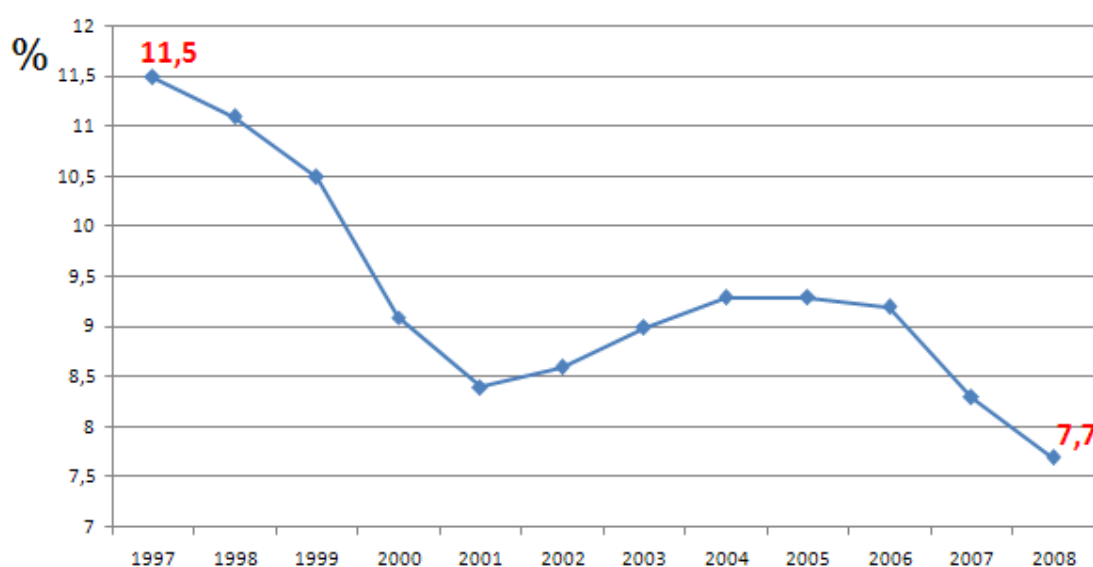
Soit le tableau suivant qui donne l'évolution du taux de chômage en France de 1994 à 2008.

Années	Taux de chômage (%) - France
1997	11,5
1998	11,1
1999	10,5
2000	9,1
2001	8,4
2002	8,6
2003	9
2004	9,3
2005	9,3
2006	9,2
2007	8,3
2008	7,7

Source : FMI

Pour étudier l'évolution de cette série chronologique, le plus simple est de la représenter à l'aide d'un graphique en ligne :

Taux de chômage en France de 1997 à 2008 (selon données FMI)



[Fichier EXCEL](#)

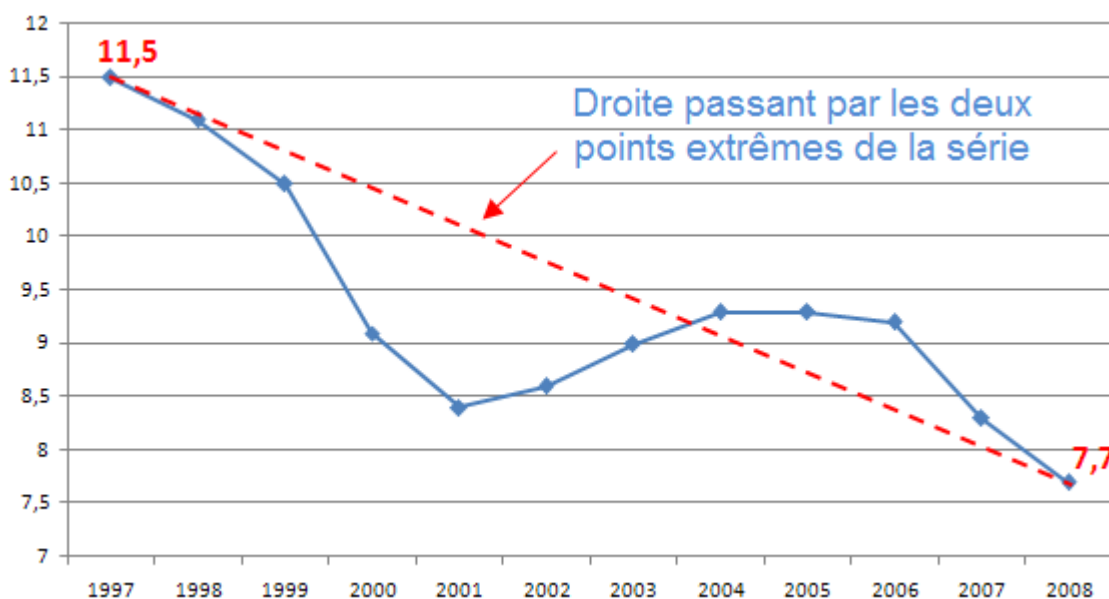
On constate alors que la « tendance » du taux de chômage sur les années 1997-2008 est à la baisse. On constate aussi qu'après avoir baissé jusqu'en 2001, le taux de chômage a connu une légère remontée avec un pic en 2004 et a ensuite encore baissé pour atteindre 7,7 % en 2008.

Bien souvent, cette analyse graphique est suffisamment éloquente pour ne pas poursuivre l'analyse. Néanmoins, il est possible de développer plus rigoureusement ce raisonnement et de déterminer mathématiquement une droite dont la pente nous donnera la « tendance ».

B – Détermination de la tendance par la méthode des points extrêmes

Puisqu'il faut deux points pour tracer une droite, une idée simple consiste à faire passer une droite par les deux points extrêmes de la série, soit {1997 ; 11,5} d'une part et {2008 ; 7,7} d'autre part. On obtient alors une droite qui nous indique une tendance négative.

Détermination de la tendance par une droite passant par les deux points extrêmes



[Fichier EXCEL](#)

Cette méthode n'est cependant pas très satisfaisante car elle ne tient compte que de des deux points extrêmes. Une meilleure méthode est celle dite des « moindres carrés ordinaires » ou MCO en abrégé.

C – Détermination de la tendance par la méthode MCO

Ce nom bizarre vient du fait que la méthode consiste à déterminer la droite d'ajustement en **minimisant la somme du carré des écarts entre cette droite et les observations**. Les détails mathématiques de cette méthode importent peu dans un cours de statistique descriptive, car **l'essentiel est de savoir calculer les coordonnées de la droite**. De plus, les machines à calculer ainsi que les logiciels comme EXCEL permettent un calcul et un tracé facile de cette droite.

Cette droite, comme toutes les droites, a pour expression l'équation :

$$y = a \cdot t + b$$

où y est un vecteur qui représente les valeurs prises par la variable y dans la population composée des n unités statistiques. Par conséquent, nous avons $y = (y_1, y_2, \dots, y_i, \dots, y_n)$ où i varie de 1 à n . Plus précisément, dans notre cas particulier, $n = 12$ et :

$$y = \{11,5 ; 11,1 ; 10,5 ; 9,1 ; 8,4 ; 8,6 ; 9 ; 9,3 ; 9,3 ; 9,2 ; 8,3 ; 7,7 \}.$$

Les chiffres ont été séparés par des points virgules afin d'éviter la confusion avec les décimales.

De la même façon, t est un vecteur qui représente les valeurs prises par la variable t dans la population composée des n unités statistiques. Par conséquent, nous avons $t = (t_1, t_2, \dots, t_i, \dots, t_n)$ où i varie de 1 à n . Plus précisément, dans notre cas particulier, nous avons 12 observations (de 1997 à 2008). La variable t va donc prendre les valeurs entières de 1 à 12 (le chiffre 1 correspond à 1997 et le chiffre 12 correspond à 2008). Nous avons ainsi :

$$t = \{1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12\}$$

Dans la terminologie du chapitre 1, nous avons ici une population composée de n unités statistiques ($n = 12$) et décrites par les deux variables (dimensions quantitatives) y et t .

Les formules de calcul des coefficients a et b sont alors données respectivement par² :

$a = \frac{\sum_i t_i y_i - n \cdot \bar{t} \cdot \bar{y}}{\sum_i t_i^2 - n(\bar{t})^2}$	$b = \bar{y} - a\bar{t}$
--	--------------------------

² Ces formules sont données ici sans démonstration, le lecteur intéressé par une démonstration rigoureuse pourra consulter avec profit le livre de PY, Bernard (2007), [Statistique descriptive : nouvelle méthode pour comprendre et bien réussir](#) 5ème édition, Economica.

Pour calculer a et b, nous pouvons utiliser une machine à calculer (la plupart des machines ont des fonctions statistiques qui calculent ces deux valeurs) ou un tableur. Ci-après, les calculs sont présentés dans EXCEL (mais ils peuvent être faits à la main en suivant la même démarche qui consiste à construire un tableau identique à celui présenté ci-après, puis à effectués les calculs dans l'ordre proposé) :

Détermination du trend (tendance) par la méthode des moindres carrés ordinaires (MCO)				
t_i	y_i	$t_i y_i$	t_i^2	
1	11,5	11,5	1	
2	11,1	22,2	4	
3	10,5	31,5	9	
4	9,1	36,4	16	
5	8,4	42	25	
6	8,6	51,6	36	
7	9	63	49	
8	9,3	74,4	64	
9	9,3	83,7	81	
10	9,2	92	100	
11	8,3	91,3	121	
12	7,7	92,4	144	
78	112	692	650	

$\sum_{i=1}^{12} t_i$	$\sum_{i=1}^{12} y_i$	$\sum_{i=1}^{12} t_i y_i$	$\sum_{i=1}^{12} t_i^2$	$\bar{t} = \frac{1}{12} \sum_{i=1}^{12} t_i$	$\bar{t}^2$	$n \times \bar{t}^2$	$\bar{y} = \frac{1}{12} \sum_{i=1}^{12} y_i$	$n \bar{t} \bar{y}$
				6,5	42,25	507	9,333333333	728

On calcule ensuite le coefficient a :

$$a = \frac{\sum_{i=1}^n t_i y_i - n \cdot \bar{t} \cdot \bar{y}}{\sum_{i=1}^n t_i^2 - n(\bar{t})^2} = \frac{692 - 728}{650 - 507} = \frac{-36}{143} = -0,251748251$$

et on en déduit le coefficient b :

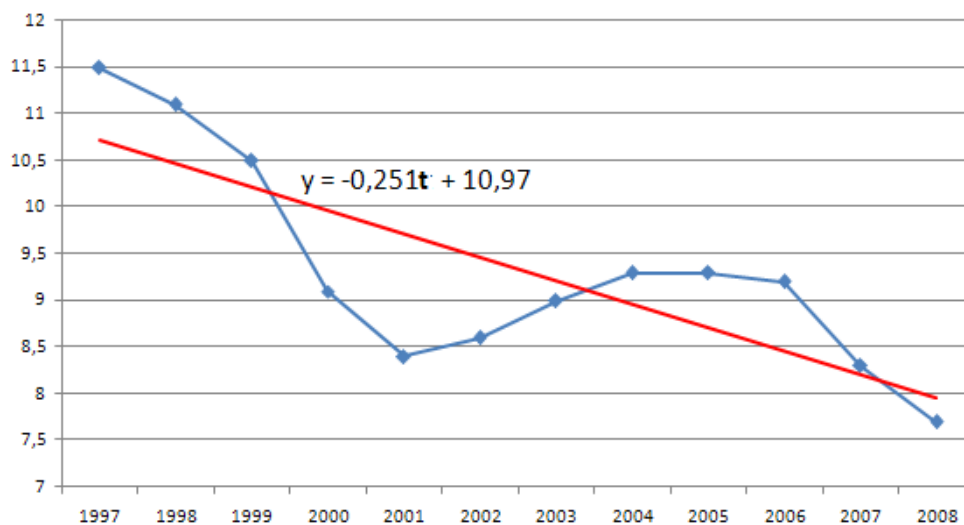
$$b = \bar{y} - a\bar{t} = 9,333 - (-0,251748251 \times 6,5) = 9,333 + 1,636363636 \approx 10,97$$

On obtient donc l'équation du trend qui est :

$$y = -0,2517t + 10,97$$

[Fichier EXCEL](#)

Nous pouvons ensuite tracer la droite MCO sur le graphique initial :



[Fichier EXCEL](#)

En résumé, la méthode MCO est plus rigoureuse que la méthode des points extrêmes car elle permet de calculer la droite de tendance en tenant compte de toutes les observations.

2 – L'étude de la liaison statistique entre deux variables

Nous allons maintenant nous intéresser à la mise en évidence d'une relation statistique entre deux variables à partir de la méthode MCO. Nous prendrons comme exemple l'étude de la relation entre prix et quantité d'un produit.

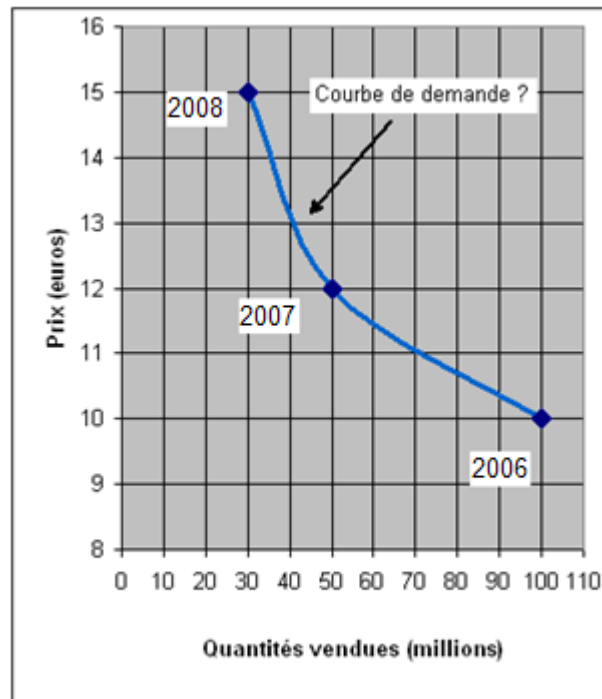
A – L'exemple d'une fonction de demande

Il est important pour une entreprise d'avoir une bonne idée de la demande qui s'adresse à son produit. Comment faire pour connaître la fonction de demande pour un produit ?

La première idée qui vient à l'esprit consiste à tracer un repère quantité/prix, avec la quantité en abscisse et le prix en ordonnée, comme ci-dessous. Supposons que l'on dispose pour cela des informations suivantes :

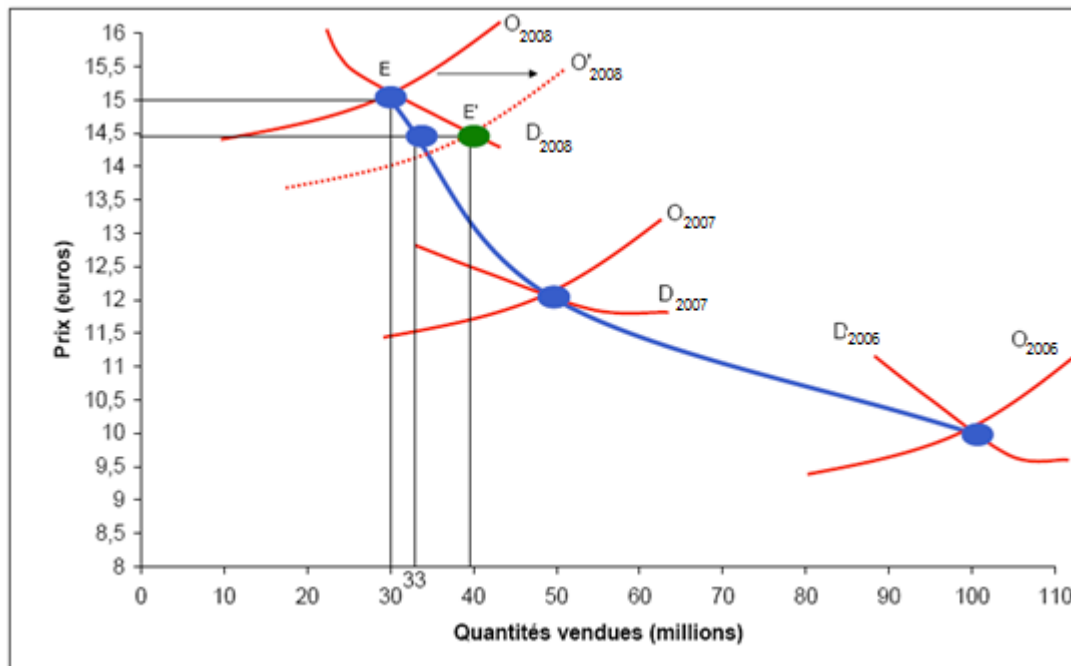
	Prix du produit (euros)	quantités vendues (millions d'unités)
2006	10	100
2007	12	50
2008	15	30

Un graphique basé sur ces informations donnerait le résultat suivant :



La courbe obtenue en joignant les 3 points est bien décroissante et suggère que plus le prix augmente, plus la quantité demandée diminue. S'agit-il pour autant d'une fonction de demande ? Pas nécessairement. Le prix et la quantité d'un bien sont normalement déterminés à la fois par l'offre et la demande, du moins lorsque le marché est concurrentiel (si le marché n'est pas concurrentiel, les choses n'en sont que plus compliquées). Mais de toute façon le prix et la quantité du produit s'établissent à l'intersection de l'offre et de la demande.

Les 3 points du graphique précédent sont généralement interprétés comme trois points d'équilibre, ainsi qu'illustré ci-dessous :

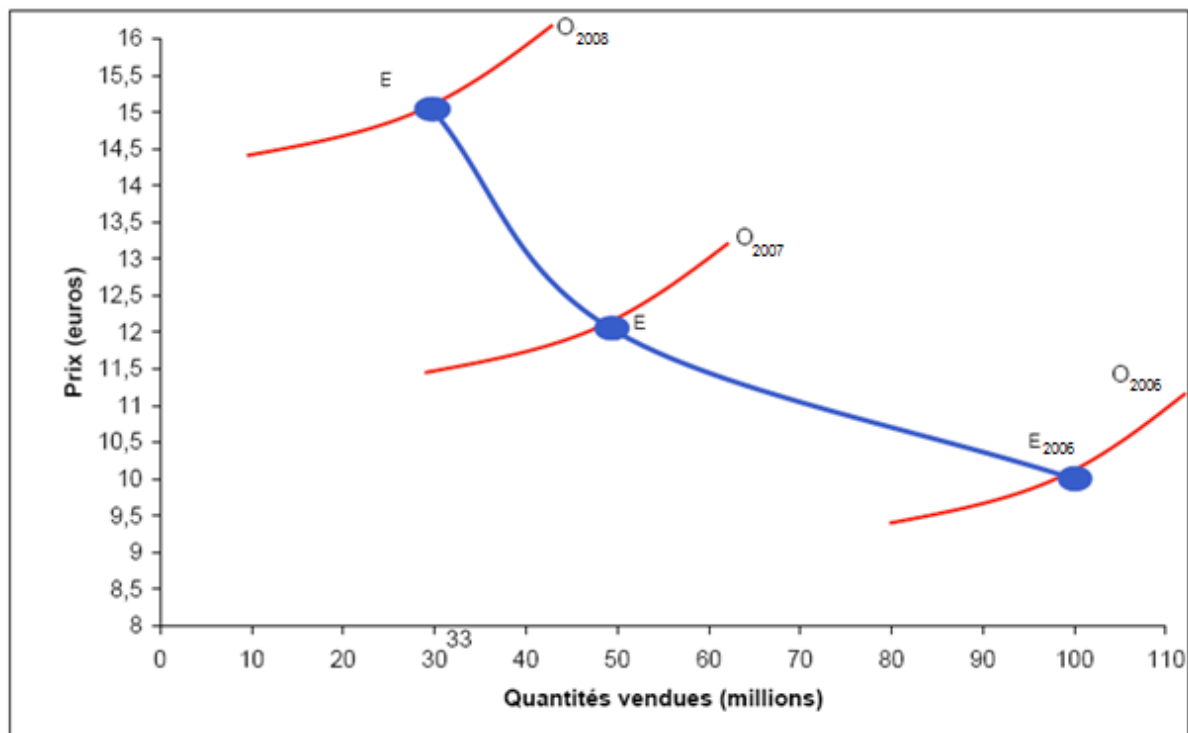


Sur le graphique précédent, nous voyons que les 3 points sont trois points d'équilibre qui résultent de l'intersection de courbes d'offre et de demande. Par exemple, en 2008, il s'est vendu 30 millions d'unités au prix unitaire de 15 euros, ce qui correspond au point d'équilibre E, qui est à l'intersection des courbes d'offre et de demande de l'année 2008.

En reliant les 3 observations temporelles pour en faire une courbe de demande, on commet sans doute une erreur. On est conduit à sous-estimer l'élasticité de la demande par rapport au prix. Par exemple, on pourrait penser qu'une baisse du prix de 15 à 14,5 euros a pour effet d'augmenter la demande de 30 à 33 millions d'unités. Or, ainsi qu'on peut le voir sur le graphique, une baisse du prix de 15 à 14,5 entraîne une augmentation bien plus importante de la demande (de 30 à 40 millions). Pour que la quantité vendue soit effectivement égale à 40 millions, il ne faut pas seulement qu'il y ait un déplacement le long de D_{2008} , il faut aussi que la courbe d'offre O_{2008} se déplace de façon à ce que le nouveau point d'intersection soit en E' (ce que nous supposons ici).

Cependant, comme illustré sur le graphique ci-dessous, on ne peut pas exclure que les 3 observations temporelles correspondent à 3 points sur la fonction de demande. Mais cela signifie en fait que la courbe de demande n'a pas changé, alors que la courbe d'offre s'est déplacée vers la gauche (en supposant que maintenant on commence en 2006, puis on continue avec 2007 et ensuite 2008).

Trois courbes d'offre successives, face à une courbe de demande supposée inchangée, constituent ainsi une justification simple de l'estimation d'une courbe de demande par un nuage de points constitués de couples prix/quantité observés à différents points du temps (de préférence en un même lieu). Il existe des analyses bien plus subtiles et le lecteur intéressé peut se rapporter pour plus de détails à un ouvrage d'économétrie.



B - L'équation de régression linéaire

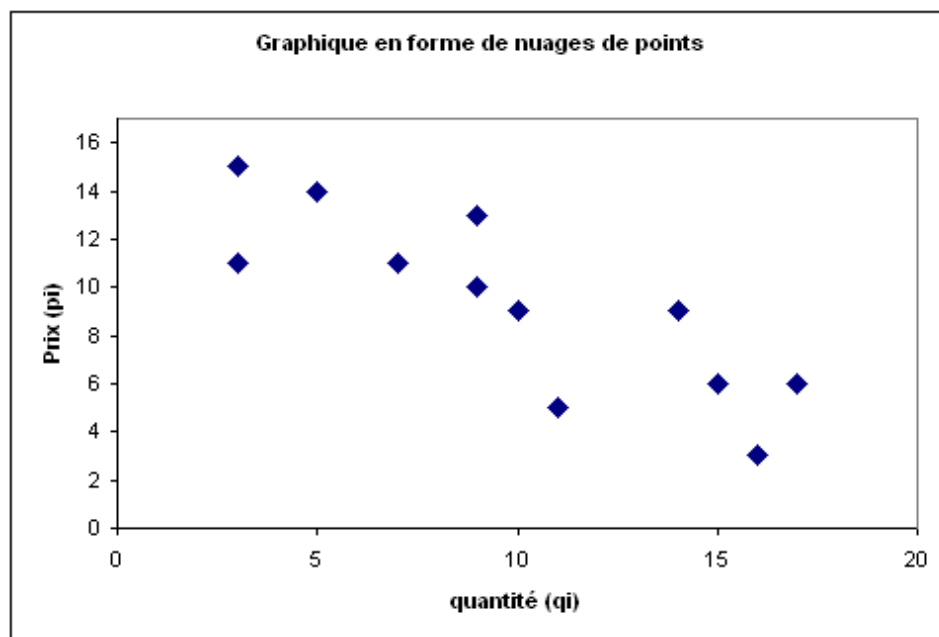
Passons maintenant à la procédure d'estimation proprement dite, en supposant que le modèle approprié soit celui décrit par le schéma ci-dessus. Cependant, trois observations ne suffisent pas pour faire une estimation par la méthode des moindres carrés. Nous allons donc :

- ajouter des observations et pour ce faire
- changer de cadre temporel (nous allons passer de l'année au mois).

Le tableau ci-après donne les ventes mensuelles et le prix unitaire correspondant. Ces valeurs représentent les observations qui vont servir à l'estimation de la fonction de demande. Ces chiffres, sont inventés pour l'occasion et il serait sans doute plus difficile d'obtenir une relation aussi évidente avec des chiffres réels. Cependant, ce qui nous intéresse principalement ici, c'est d'illustrer la méthode.

Dates	Quantité q_i (milliers d'unités)	Prix p_i (euros)
Janvier	5	14
Février	15	6
Mars	9	10
Avril	14	9
Mai	3	11
Juin	9	13
Juillet	10	9
Août	17	6
Septembre	11	5
Octobre	16	3
Novembre	7	11
Décembre	3	15

Le graphique en nuage de points (graphique de dispersion) des observations laisse présupposer l'existence d'une relation décroissante. Pour que notre graphique soit conforme à la présentation traditionnelle du graphique de la fonction de demande, nous avons mis le prix en ordonnée et la quantité en abscisse. *Toutefois, il faut garder présent à l'esprit le fait qu'au niveau de la causalité économique, c'est le prix qui détermine la quantité et non l'inverse.*



[Fichier EXCEL](#)

Appliquons maintenant la méthode des moindres carrés ordinaires introduite pour l'analyse du trend. Cette fois, nous devons estimer les coefficients a et b d'une équation de la forme $y = a * x + b$. Par rapport à la formule du trend temporel, il n'y a que le symbole t qui change.

En effet, dans l'équation du trend, nous avons :

$$y = a * t + b$$

Maintenant, nous avons :

$$y = a * x + b$$

Par conséquent au lieu d'avoir :

$a = \frac{\sum_i t_i y_i - n \cdot \bar{t} \cdot \bar{y}}{\sum_i t_i^2 - n(\bar{t})^2}$	$b = \bar{y} - a\bar{t}$
--	--------------------------

Nous avons maintenant :

$a = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}$	$b = \bar{y} - a\bar{x}$
--	--------------------------

Dans notre exemple, nous avons une équation de la forme $p = a * q + b$ où p remplace y et où q remplace x . Nous nous attendons ici à ce que le coefficient a soit négatif.

Les principaux calculs nécessaires sont donnés ci-contre.

Quantité q_i	Prix p_i	$p_i q_i$	q_i^2						
5	14	70	25						
15	6	90	225						
9	10	90	81						
14	9	126	196						
3	11	33	9						
9	13	117	81						
10	9	90	100						
17	6	102	289						
11	5	55	121						
16	3	48	256						
7	11	77	49						
3	15	45	9						
119	112	943	1441						

$\sum_{i=1}^{12} q_i$
↑

$\sum_{i=1}^{12} p_i$
↑

$\sum_{i=1}^{12} q_i p_i$
↑

$\sum_{i=1}^{12} q_i^2$
↑

$\bar{q} = \frac{1}{12} \sum_{i=1}^{12} q_i$
9,91666667

$\bar{q}^2$
98,3402778

$n \times \bar{q}^2$
1180,08333

$\bar{p} = \frac{1}{12} \sum_{i=1}^{12} p_i$
9,33333333

$n \bar{q} \bar{p}$
1110,66667

On calcule ensuite le coefficient a :

$$a = \frac{\sum_{i=1}^n q_i p_i - n \cdot \bar{q} \cdot \bar{p}}{\sum_{i=1}^n q_i^2 - n(\bar{q})^2} = \frac{943 - 1110,66667}{1441 - 1180,08333} = \frac{-167,66667}{260,91667} = -0,6426$$

et on en déduit le coefficient b :

$$b = \bar{p} - a\bar{q} = 9,33333333 - (-0,642606 \times 9,91666667) = 15,706$$

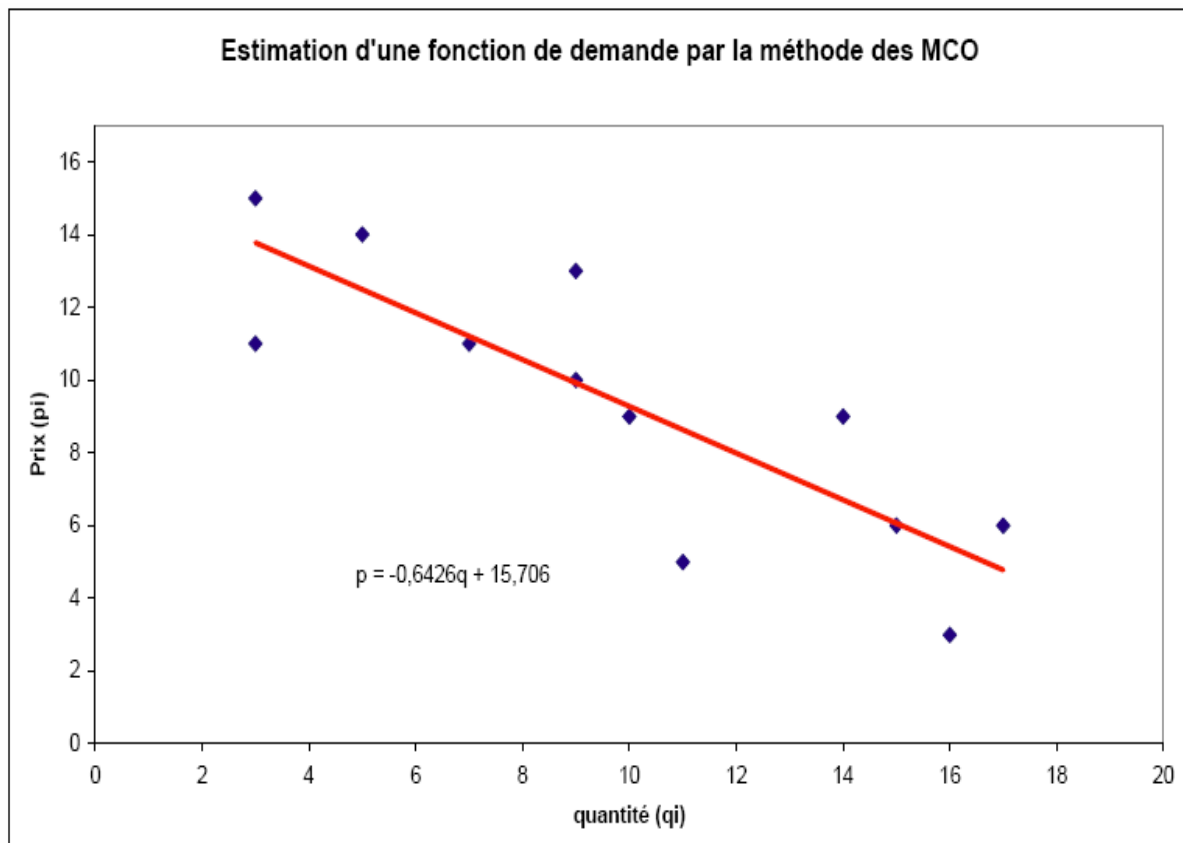
On obtient donc l'équation du trend qui est :

$$p = -0,6426q + 15,706$$

Fichier EXCEL

Un didacticiel Microsoft est disponible pour l'utilisation directe de la fonction MCO sans faire soi-même les calculs : <http://office.microsoft.com/fr-fr/excel/HA010877851036.aspx>

On a ainsi la droite de demande décroissante comme illustré sur la figure :



On peut maintenant exprimer p en fonction de q si l'on préfère. Sachant que :

$$p = aq + b \Leftrightarrow q = \frac{1}{a}p - \frac{b}{a}$$

On a :

$$q = - (1,55618 \times p) + 24,4413$$

On peut ensuite se servir de la fonction de demande ainsi obtenue pour évaluer les conséquences d'une baisse du prix sur la quantité demandée et donc sur la recette totale.

C - Le coefficient de détermination

Lorsque l'on a estimé la droite de régression, on doit se demander si cette estimation est de bonne qualité. On dispose d'un premier outil pour répondre à cette question : c'est le **coefficient de détermination** dont la formule est donnée par :

$$r^2 = \frac{\left[n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]^2}{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}$$

Ce coefficient est compris entre 0 et 1. Plus il est proche de 1 et plus la qualité globale de la régression est bonne.

À titre d'exemple, calculons le coefficient de détermination de l'équation de régression de demande précédent. Remplaçons x par q et y par p dans la formule de r^2 , afin d'avoir :

$$r^2 = \frac{\left[n \sum_{i=1}^n q_i p_i - \left(\sum_{i=1}^n q_i \right) \left(\sum_{i=1}^n p_i \right) \right]^2}{\left[n \sum_{i=1}^n q_i^2 - \left(\sum_{i=1}^n q_i \right)^2 \right] \left[n \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i \right)^2 \right]}$$

Un didacticiel Microsoft est disponible ici pour l'obtention directe du coefficient de détermination sans faire soi-même les calculs : <http://office.microsoft.com/fr-fr/excel/HA010877851036.aspx>

Pour faire les calculs, voici comment procéder :

	A	B	C	D	E	F	G	H	I	J	K										
	Quantité	Prix																			
1	q_i	p_i	La formule générale du coefficient de détermination est donnée par :																		
2	5	14	$r^2 = \frac{\left[n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]^2}{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}$																		
3	15	6																			
4	9	10																			
5	14	9																			
6	3	11																			
7	9	13																			
8	10	9																			
9	17	6																			
10	11	5	En remplaçant x_i par q_i et y_i par p_i , on obtient :																		
11	16	3	$r^2 = \frac{\left[n \sum_{i=1}^n q_i p_i - \left(\sum_{i=1}^n q_i \right) \left(\sum_{i=1}^n p_i \right) \right]^2}{\left[n \sum_{i=1}^n q_i^2 - \left(\sum_{i=1}^n q_i \right)^2 \right] \left[n \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i \right)^2 \right]}$																		
12	7	11																			
13	3	15																			
14																					
15																					
16																					
17																					
18																					
19																					
20																					
21			Pour trouver r^2 , il faut d'abord calculer ses différentes composantes à l'aide du tableau ci-dessous :																		
22																					
23	Quantité	Prix	$q_i \times p_i$	q_i^2	p_i^2																
24	q_i	p_i																			
25	5	14	70	25	196																
26	15	6	90	225	36																
27	9	10	90	81	100																
28	14	9	126	196	81																
29	3	11	33	9	121																
30	9	13	117	81	169																
31	10	9	90	100	81																
32	17	6	102	289	36																
33	11	5	55	121	25																
34	16	3	48	256	9																
35	7	11	77	49	121																
36	3	15	45	9	225																
37	119	112	943	1441	1200																
38	↑	↑	↑	↑	↑																
39																					
40	$\sum_{i=1}^n q_i$	$\sum_{i=1}^n p_i$	$\sum_{i=1}^n q_i p_i$	$\sum_{i=1}^n q_i^2$	$\sum_{i=1}^n p_i^2$																
41																					
42																					
43	$\left(\sum_{i=1}^n q_i \right)^2$	$\left(\sum_{i=1}^n p_i \right)^2$	$n \times \sum_{i=1}^n q_i p_i$	$n \times \sum_{i=1}^n q_i^2$	$n \times \sum_{i=1}^n p_i^2$																
44																					
45																					
46																					
47																					
48	14161	12544	11316	17292	14400																
49						$r^2 = \frac{\left[n \sum_{i=1}^n q_i p_i - \left(\sum_{i=1}^n q_i \right) \left(\sum_{i=1}^n p_i \right) \right]^2}{\left[n \sum_{i=1}^n q_i^2 - \left(\sum_{i=1}^n q_i \right)^2 \right] \left[n \sum_{i=1}^n p_i^2 - \left(\sum_{i=1}^n p_i \right)^2 \right]}$ $r^2 = \frac{[11316 - 13328]^2}{[17292 - 14161][14400 - 12544]}$ $r^2 = \frac{4048144}{3131 \times 1856} = \frac{4048144}{5811136} = 0,6966184$															
50	$\left(\sum_{i=1}^n q_i \right) \left(\sum_{i=1}^n p_i \right)$																				
51																					
52																					
53																					
54																					
55	13328																				
On peut éviter de faire les calculs ci-dessus en appliquant la fonction EXCEL :																					
$=COEFFICIENT.DETERMINATION(A28:A39;B28:B39) \rightarrow 0,6966184$																					

[Fichier EXCEL](#)

Interprétation du résultat : Plus le coefficient r^2 tend vers 1, plus la qualité globale de la régression, est bonne. Ici, le r^2 est proche de 0,7. **On peut juger que c'est insuffisant.** Il faut de toute manière compléter ce premier diagnostic par le calcul d'autres statistiques, mais ceci est l'objet d'un cours **d'économétrie** et non plus de statistique descriptive.

3 – Le test d'indépendance du Khi-carré de PEARSON

A - Introduction

Le test d'indépendance du khi-carré (l'écriture anglaise est « chi-square ») a été développé par [Karl PEARSON](#) (1857-1936), l'un des fondateurs de la **statistique mathématique**.

L'expression test du khi-carré recouvre plusieurs tests statistiques³, trois tests principalement :

- le **test d'ajustement ou d'adéquation**, qui compare globalement la distribution observée dans un échantillon statistique à une distribution théorique, celle du khi-carré.
- Le **test d'indépendance** du khi-carré qui permet de contrôler l'indépendance de deux caractères dans une population donnée.
- le **test d'homogénéité** du khi-carré qui teste si des échantillons sont issus d'une même population.

Le test qui nous intéresse ici est uniquement le test d'indépendance du khi-carré. Ce test sert à apprécier l'existence ou non d'une relation entre **deux dimensions** au sein d'une population, lorsque ces dimensions sont mesurées sur des échelles qualitatives et/ou que les modalités de ces échelles de mesure ont été regroupées en catégories. Cette définition inclut l'utilisation du test du khi-carré pour apprécier l'existence d'une relation entre deux dimensions mesurées sur des échelles quantitatives groupées en catégories, mais le test est moins approprié que la régression (sauf si l'on ne peut pas remonter aux données brutes et que l'on ne dispose que des données regroupées en catégories). On peut enfin l'utiliser pour étudier la relation entre une dimension quantitative et une dimension qualitative.

À noter que ce test permet de contrôler l'existence d'une dépendance mais en aucun cas le sens de cette dépendance (sauf dans certains cas particuliers où l'existence d'une relation implique une causalité univoque comme dans l'exemple ci-après où le sexe peut avoir une influence sur le choix d'une certaine matière mais où il est impossible que le choix d'une certaine matière ait une influence sur le sexe).

À noter enfin que les différents tests du khi-carré ne doivent pas être confondus avec la **distribution théorique du khi-carré**, dont les valeurs tabulées servent seulement à valider ces différents tests.

³ Une présentation synthétique des différents tests est donnée dans Wikipédia (voir l'article « [Test du khi-2](#) »)

B - Exemple d'utilisation⁴

Nous allons maintenant montrer comment ce test peut-être utilisé dans le cas d'une distribution à deux dimensions.

La première dimension, désignée par X , pourra être une variable ou un caractère, comprenant des **catégories** (ou des classes) (issues généralement d'un regroupement des valeurs de la variable ou des modalités du caractère). On aura ainsi les classes $A_1, \dots, A_L$

La seconde dimension, désignée par Y , pourra être une variable ou un caractère, comprenant des **catégories** (ou des classes) (issues généralement d'un regroupement des valeurs de la variable ou des modalités du caractère). On aura ainsi les classes $B_1, \dots, B_C$.

Dans ces conditions, l'effectif n de la population se distribue dans un tableau croisé⁵ :

		Catégories de la dimension Y						
		B_1	B_2	$\dots$	B_j	$\dots$	B_C	Total
Catégories de la dimension X	A_1	n_{11}			n_{1j}		n_{1C}	
	A_2	n_{21}						
	$\vdots$							
	A_i	n_{i1}			n_{ij}		n_{iC}	L_i
	$\vdots$							
	A_L	n_{L1}			n_{Lj}			
Total					C_j			n

Où $n_{i,j}$ représente l'effectif qui appartient simultanément à la catégorie A_i de la dimension X et à la catégorie B_j de la dimension Y .

⁴ Pour une très bonne explication de la façon d'effectuer un test d'indépendance du khi-2 (ou chi-2), voir Charles McCREERY « The CHI-SQUARE test : A test of Association Between Categorical Variables ». Sur internet : <http://www.celiagreen.com/charlesmccreery/statistics/chisquare.pdf>. Voir aussi les explications très claires données sur BibMath dont nous nous sommes inspirés ci-après : <http://www.bibmath.net/dico/index.php3?action=affiche&quoi=.c/chideux/test.html>.

⁵ Appelé « Contingency table » en anglais et abusivement traduit en français par l'expression « Tableau de contingence »

L_i représente la somme des effectifs appartenant à la catégorie A_i de la dimension X. C'est donc une distribution conditionnelle (voir le chapitre 2). C_j représente la somme des effectifs de la catégorie B_j . C'est donc aussi une distribution conditionnelle.

Exemple : Sexe et préférence pour un cours au sein d'une filière « économie »

Soit le tableau ci-dessous, qui donne les résultats d'une enquête hypothétique effectuée auprès de 400 étudiants, sur leurs préférences en matière de cours. On leur a demandé : « Parmi ces 4 matières : HPE, Droit, Micro et Macro, laquelle préférez-vous ? » (Il était interdit de répondre : « aucune »).

	H	F	Total
HPE	50	50	100
Droit	110	25	135
Micro	40	25	65
Macro	50	50	100
Total	250	150	400

Dans cet exemple, la dimension Y est le sexe et comprend deux modalités (« H » et « F »). La dimension X est la matière, qui comprend 4 modalités (« HPE », « Droit », « Micro » et « Macro »). On remarquera que les « catégories » des dimensions X et Y ne sont pas issues d'un regroupement, mais qu'il s'agit simplement des modalités brutes de chacun des deux caractères étudiés.

Pour savoir si le sexe a une influence significative sur le choix des matières, nous allons faire un test du khi-carré. On remarque que le droit et la micro sont davantage préférés par la population masculine tandis que HPE et macro semblent ne pas être préférés plus par la population masculine que par la population féminine.

Le test du khi carré va apporter une information supplémentaire. Il va permettre de dire si les différences de préférences pour les diverses matières qui sont attribuées au sexe sont le fait du hasard du tirage ou si elles sont réelles. Elles peuvent en effet être dues au hasard de l'échantillon. **Ce que le test va nous dire c'est dans quelle mesure la différence est indépendante de l'échantillon choisi** (et donc se retrouverait en général si l'on prenait n'importe quel autre échantillon).

Pour cela on doit calculer l'expression suivante, que nous appellerons, faute d'une expression plus appropriée, le « khi-carré calculé »⁶ :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^l \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad \text{Avec :} \quad e_{ij} = \frac{L_i C_j}{n} = \frac{C_j L_i}{n}$$

⁶ Pour obtenir directement le khi-2 calculé, voir le calculateur en ligne (très pratique pour vérifier ses résultats) : <http://www.seuret.com/biostat/chi.php>

	H	F	Total
HPE	50	50	100
Droit	110	25	135
Micro	40	25	65
Macro	50	50	100
Total	250	150	400

Et en effectuant les multiplications
comme indiqué par la formule

$$e_{ij} = \frac{C_{i.} L_{.j}}{n}$$

$$\begin{aligned} (250 \times 100)/400 &= 62,5 & (150 \times 100)/400 &= 37,5 \\ (250 \times 135)/400 &= 84,375 & (150 \times 135)/400 &= 50,625 \\ (250 \times 65)/400 &= 40,625 & (150 \times 65)/400 &= 24,375 \\ (250 \times 100)/400 &= 62,5 & (150 \times 100)/400 &= 37,5 \end{aligned}$$

	Hommes	Femmes	Total
HPE	62,5	37,5	100
Droit	84,375	50,625	135
Micro	40,625	24,375	65
Macro	62,5	37,5	100
Total	250	150	400

Ensuite calculons les $n_{ij} - e_{ij}$:

n_{ij}

	Hommes	Femmes
HPE	50	50
Droit	110	25
Micro	40	25
Macro	50	50
Total	250	150

e_{ij}

	Hommes	Femmes
HPE	62,5	37,5
Droit	84,375	50,625
Micro	40,625	24,375
Macro	62,5	37,5
Total	250	150

$n_{ij} - e_{ij}$

	Hommes	Femmes
HPE	-12,5	12,5
Droit	25,625	-25,625
Micro	-0,625	0,625
Macro	-12,5	12,5

Puis calculons $(n_{ij} - e_{ij})^2$:

	Hommes	Femmes
HPE	156,25	156,25
Droit	656,640625	656,640625
Micro	0,390625	0,390625
Macro	156,25	156,25

Les chiffres de ce
tableau au carré

Ensuite calculons $(n_{ij} - e_{ij})^2 / e_{ij}$:

	Hommes	Femmes
HPE	2,5	4,16666667
Droit	7,78240741	12,970679
Micro	0,00961538	0,01602564
Macro	2,5	4,16666667

Ensuite effectuons la somme des huit chiffres obtenus :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^b \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 2,5 + 7,782 + 0,00961 + 2,5 + 4,166 + 12,97 + 0,016 + 4,1666 = 34,11 \text{ environ}$$

Une fois que l'on connaît le khi-carré calculé, il reste à le comparer avec la valeur khi-carré issue de la distribution du khi-carré (voir le tableau ci-après).

Pour trouver cette valeur dans le tableau, nous devons prendre en compte deux informations supplémentaires :

- Le nombre de « degrés de liberté » qui se calcule ainsi :

$$\begin{aligned} &\text{Degrés de liberté} \\ &= \\ &\{(\text{Nb de catégories [ou modalités ou valeurs] de X}) - 1\} \\ &\quad \times \\ &\{(\text{Nb de catégories [ou modalités ou valeurs] de Y}) - 1\} \end{aligned}$$

Ici, il y a 4 modalités pour X (les 4 matières) et 2 modalités pour Y (les deux sexes). Donc, le nombre de degrés de liberté est égal à :

$$(4 - 1) \times (2 - 1) = 3 \times 1 = 3.$$

- Ensuite, nous devons choisir la probabilité de fiabilité du test : 5% de chances de se tromper, 1% ou 1 pour 1000. Nous allons choisir 5%, soit $P = 0,05$.

Nous avons donc 3 degrés de liberté et une probabilité de fiabilité du test de $P=0,05$. Par conséquent, nous voyons dans la table que le khi-carré théorique est égal à :

$$\chi^2_{0,05} = 7,82$$

Il nous reste maintenant à comparer le khi carré théorique issu de la table (7,82) avec le khi-carré calculé (34,11 environ) :

$$\chi^2_{0,05} = 7,82 < \chi^2_{\text{calculé}} = 34,11$$

La règle est la suivante :

- **Si le khi-carré calculé est inférieur au khi-carré théorique : indépendance**
- **Si le khi-carré calculé est supérieur au khi-carré théorique : dépendance**

Etant donné que le chi-carré calculé est supérieur au khi carré théorique, nous pouvons conclure que le sexe a une influence sur le choix de la matière. Notre observation initiale sur la base de l'échantillon est donc probablement vraie à l'extérieur de l'échantillon (avec cependant 5% de chances de nous tromper).

Degrés de liberté	P=0,05	P=0,01	P=0,001	Degrés de liberté	P=0,05	P=0,01	P=0,001
1	3.84	6.64	10.83	50	67.51	76.15	86.66
2	5.99	9.21	13.82	51	68.67	77.39	87.97
3	7.82	11.35	16.27	52	69.83	78.62	89.27
4	9.49	13.28	18.47	53	70.99	79.84	90.57
5	11.07	15.09	20.52	54	72.15	81.07	91.88
6	12.59	16.81	22.46	55	73.31	82.29	93.17
7	14.07	18.48	24.32	56	74.47	83.52	94.47
8	15.51	20.09	26.13	57	75.62	84.73	95.75
9	16.92	21.67	27.88	58	76.78	85.95	97.03
10	18.31	23.21	29.59	59	77.93	87.17	98.34
11	19.68	24.73	31.26	60	79.08	88.38	99.62
12	21.03	26.22	32.91	61	80.23	89.59	100.88
13	22.36	27.69	34.53	62	81.38	90.80	102.15
14	23.69	29.14	36.12	63	82.53	92.01	103.46
15	25.00	30.58	37.70	64	83.68	93.22	104.72
16	26.30	32.00	39.25	65	84.82	94.42	105.97
17	27.59	33.41	40.79	66	85.97	95.63	107.26
18	28.87	34.81	42.31	67	87.11	96.83	108.54
19	30.14	36.19	43.82	68	88.25	98.03	109.79
20	31.41	37.57	45.32	69	89.39	99.23	111.06
21	32.67	38.93	46.80	70	90.53	100.42	112.31
22	33.92	40.29	48.27	71	91.67	101.62	113.56
23	35.17	41.64	49.73	72	92.81	102.82	114.84
24	36.42	42.98	51.18	73	93.95	104.01	116.08
25	37.65	44.31	52.62	74	95.08	105.20	117.35
26	38.89	45.64	54.05	75	96.22	106.39	118.60
27	40.11	46.96	55.48	76	97.35	107.58	119.85
28	41.34	48.28	56.89	77	98.49	108.77	121.11
29	42.56	49.59	58.30	78	99.62	109.96	122.36
30	43.77	50.89	59.70	79	100.75	111.15	123.60
31	44.99	52.19	61.10	80	101.88	112.33	124.84
32	46.19	53.49	62.49	81	103.01	113.51	126.09
33	47.40	54.78	63.87	82	104.14	114.70	127.33
34	48.60	56.06	65.25	83	105.27	115.88	128.57
35	49.80	57.34	66.62	84	106.40	117.06	129.80
36	51.00	58.62	67.99	85	107.52	118.24	131.04
37	52.19	59.89	69.35	86	108.65	119.41	132.28
38	53.38	61.16	70.71	87	109.77	120.59	133.51
39	54.57	62.43	72.06	88	110.90	121.77	134.74
40	55.76	63.69	73.41	89	112.02	122.94	135.96
41	56.94	64.95	74.75	90	113.15	124.12	137.19
42	58.12	66.21	76.09	91	114.27	125.29	138.45
43	59.30	67.46	77.42	92	115.39	126.46	139.66
44	60.48	68.71	78.75	93	116.51	127.63	140.90
45	61.66	69.96	80.08	94	117.63	128.80	142.12
46	62.83	71.20	81.40	95	118.75	129.97	143.32
47	64.00	72.44	82.72	96	119.87	131.14	144.55
48	65.17	73.68	84.03	97	120.99	132.31	145.78
49	66.34	74.92	85.35	98	122.11	133.47	146.99
50	67.51	76.15	86.66	99	123.23	134.64	148.21
				100	124.34	135.81	149.48

Source de la table : <http://www.apprendre-en-ligne.net/random/tablekhi2.html>

C – Etapes du test d'indépendance du khi-carré

Pour résumer, les principales étapes du test d'indépendance du Khi-carré sont :

1) Si ce n'est pas déjà fait, distribuer la population statistique dans un tableau à deux dimensions où les modalités des caractères et, le cas échéant, les valeurs de la variable, sont regroupés par catégories.

2) Calculer le khi-carré dans l'hypothèse d'indépendance des deux dimensions :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^l \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

3) Calculer le nombre de degrés de liberté par la formule :

$$(\text{Nombre de lignes} - 1) \times (\text{Nombre de colonnes} - 1)$$

4) Définir une probabilité d'erreur (en pratique 5%, 1% ou 1 pour mille)

5) Utiliser le nombre de degrés de liberté et la probabilité d'erreur pour déterminer le khi-carré théorique à partir de la table fournie.

6) comparer la valeur khi-carré calculée avec la valeur khi-carrée théorique (issue de la table) et appliquer la règle suivante :

- Si le khi-carré calculé est inférieur au khi-carré théorique : indépendance
- Si le khi-carré calculé est supérieur au khi-carré théorique : dépendance

4 – Résumé

Ce chapitre a montré l'utilisation d'un même outil statistique, l'**ajustement linéaire**, appelé aussi « **méthode des Moindres Carrés Ordinaires** (MCO) pour la mise en évidence de la **tendance d'une série chronologique**, puis pour la mise en évidence d'une **liaison statistique entre deux variables**. L'étude de la liaison statistique entre deux variables peut ensuite être affinée par le calcul du **coefficient de détermination**.

Lorsqu'on s'intéresse à la liaison statistique entre deux **caractères** ou entre une variable et un caractère, ou même à la limite quand on a deux variables **mais que les données sont regroupées en catégories**, l'un des tests les plus facile et répandu est le **test d'indépendance du khi-carré** de Karl PEARSON.

Il existe ensuite d'autres outils plus élaborés pour étudier les corrélations entre variables et/ou caractères, mais ceux-ci relèvent alors d'un cours d'économétrie et/ou de statistique mathématique.

Exercices corrigés (PDF) :

TD5 2008 : [Sujets](#) - [Corrigés](#)

TD5 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#) (lien vers page HTML)