La GREAT formation

Groupe de recherche en économie appliquée et théorique

N° 001 "Réfléchir à changer" Janvier 2015

Eléments statistiques d'économétrie

Massa Coulibaly

BP. E1255 Bamako (Mali) Tel. (+223) 66 74 22 20 Email. great@greatmali.net

Table des matières

1. Synthèse statistique 1
   1.1. Paramètres de tendance centrale 1
      1.1.1. Moyenne 1
      1.1.2. Mode 3
      1.1.3. Médiane 3
      1.1.4. Quantiles 3
   1.2. Paramètres de dispersion 4
      1.2.1. Etendue 4
      1.2.2. Variance 4
      1.2.3. Ecart-type 8
      1.2.4. Coefficient de variation 8
      1.2.5. Ecarts interquartiles 9
   1.3. Paramètres de forme 10
      1.3.1. Coefficient d'asymétrie 10
      1.3.2. Coefficient d'aplatissement 11
   1.4. Paramètres de concentration 12
   1.5. Applications sur EXCEL 14
2. Echantillonnage 21
   2.1. Types d'échantillons 21
   2.2. Tirage de l'échantillon 24
      2.2.1. Tirage élémentaire 24
      2.2.2. Tirage systématique 24
      2.2.3. Tirage par grappes 25
      2.2.4. Tirage à plusieurs degrés 26
   2.3. Taille de l'échantillon 26
   2.4. Inférence statistique 27
      2.4.1. Propriétés des estimateurs 28
      2.4.2. Estimateurs de la moyenne d'échantillon 29
      2.4.3. Estimateurs de la variance d'échantillon 31
3. Lois de répartition et tests usuels 36
   3.1. Lois de répartition 36
      3.1.1. Loi Normale (Laplace-Gauss) 36
      3.1.2. Répartition Gamma (Γ) et Beta (B) 37
      3.1.3. Loi $\chi 2$ (K. Pearson) 38
      3.1.4. Répartition Student (William Gosset) 40
      3.1.5. Fisher-Snedecor 42
   3.2. Tests usuels 44
      3.2.1. Tests du khi-deux ($\chi 2$) 45
      3.2.2. Test de Student 47
      3.2.3. Test de Fisher 48
      3.2.4. Méthode de Bayes 49
      3.2.5. Méthode Neyman-Pearson 51

1. Synthèse statistique

Pour les besoins de l'analyse statistique, il est nécessaire de résumer les séries de répartition en quelques paramètres qui les caractérisent. Lorsqu'on représente graphiquement les séries de répartition à l'aide des polygones de fréquences, on note dans la plupart des cas que les effectifs accusent une certaine concentration autour d'une valeur centrale et que de part et d'autre de cette valeur centrale le polygone présente un étalement plus ou moins important d'un coté ou de l'autre.

1.1. Paramètres de tendance centrale

Ce sont:

- la moyenne
- le mode
- la médiane
- les quantiles e.g. les quartiles.

1.1.1. Moyenne

On définit la moyenne potentielle simple par: $m_k = \sqrt[k]{\frac{\sum X^k}{n}}$

Note: la moyenne pondérée s'écrira:

$$m_k = \sqrt[k]{\sum X^k f_i} \quad où \quad f_i = \frac{n_i}{n} \quad et \quad \sum f_i = 1.$$

De la moyenne potentielle, on déduit, pour différentes valeurs de k, les différents types de moyennes de la statistique descriptive.

- k = 1 ➨ moyenne arithmétique $m_1 = \overline{m} = \frac{\sum X}{n}$
- k = 2 ➨ moyenne quadratique $m_2 = \sqrt{\frac{\sum X^2}{n}}$
- k = -1 ➨ moyenne harmonique $m_{-1} = \left(\frac{\sum X^{-1}}{n} \right)^{-1} = \frac{n}{\sum \frac{1}{X}}$

- k = 0 ➨ moyenne géométrique $m_0 = 1^{\infty}$ indéterminé $\Rightarrow$ lever l'indétermination, par la fonction log puis par le rapport de dérivés

$$\ln m_k = \frac{1}{k} ln\left(\frac{\sum X^k}{n}\right) = \frac{ln\left(\frac{\sum X^k}{n}\right)}{k} \quad \rightarrow \quad \frac{0}{0}$$

$$\lim_{k=0} \left(\ln m_k\right) = \lim_{k=0} \frac{\sum X^k \ln X}{\sum X^k} = \frac{\sum \ln X}{n} = \frac{\ln \prod X}{n} \quad \Rightarrow$$

$$m_0 = \sqrt[n]{\prod X}$$

Note: Par comparaison entre les moyennes, on a $m_{-1} \leq m_0 \leq m_1$

Ex. 1. Calculer moyenne arithmétique et moyenne quadratique des 10 premiers entiers, allant de 1 à 10.

$$\overline{X} = \frac{1}{n}\sum X = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$$

$$\overline{X^2} = \sqrt{\frac{1}{n}\sum X^2} = \sqrt{\frac{1}{n}\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{(n+1)(2n+1)}{6}} = \sqrt{\frac{(11)(21)}{6}} = \sqrt{\frac{77}{2}} = 6.2$$

Ex. 2. Dans 4 villes, on a dénombré en milliers le nombre d'habitants (n) par médecins (X) suivant le tableau ci-dessous. Calculer le nombre moyen d'habitants par médecins pour l'ensemble des 4 villes

Société d'audit	X	n
A	4	16
B	9	9
C	3	60
D	5	15

$$H = \frac{Total\ habi\tan ts}{Total\ médecins} = \frac{\sum n_i}{\sum \frac{n_i}{X_i}} = \frac{16+9+60+15}{\frac{16}{4}+\frac{9}{9}+\frac{60}{3}+\frac{15}{5}} = \frac{100}{28} = 3.57 millehats/médecin$$

$$= 3.57 millehats/médecin$$

Ex. 3. Calculer l'inflation moyenne mensuelle de la série:

Mois	Inflation	Mois	Inflation	Mois	Inflation
Janvier	1	Mai	5	Septembre	8
Février	10	Juin	7	Octobre	11
Mars	3	Juillet	2	Novembre	15
Avril	4	Août	6	Décembre	9

Si r est le taux d'inflation, alors l'indice des prix est $I = 1 + r$ et l'inflation moyenne se calcule partant de l'indice moyen lui-même calculé en tant que moyenne géométrique:

$$\bar{r} = \sqrt[12]{\prod_{i=1}^{12}(1 + r_i)} - 1 = 6.7\%$$

1.1.2. Mode

Le mode (M_0) est la valeur du caractère qui a la plus grande fréquence.

1.1.3. Médiane

La médiane (m_e) est la valeur du caractère qui correspond à l'unité statistique placée au milieu de la population, rangée par rapport aux valeurs du caractère.

1.1.4. Quantiles

Soit $\frac{P}{Q}$ un nombre rationnel compris entre 0 et 1. On appelle quantile d'ordre $\frac{P}{Q}$ le nombre $N_{\frac{P}{Q}}$ qui laisse une proportion $\frac{P}{Q}$ de valeurs inférieures et une proportion $1 - \frac{P}{Q}$ de valeurs supérieures. En particulier, les quartiles ont $\frac{P}{Q} = \frac{1}{4};\ \frac{2}{4} = \frac{1}{2};\ \frac{3}{4}$ notés $Q_1;\ Q_2 = M_e;\ Q_3$

1.2. Paramètres de dispersion

Ce sont:
1. l'étendue
2. la variance et l'écart-type
3. le coefficient de variation
4. les écarts interquartiles.

1.2.1. Etendue

L'étendue est l'écart entre la plus grande valeur et la plus petite valeur du caractère. C'est une mesure de dispersion très facile à calculer avec toutefois l'inconvénient qu'elle ne dépend que des valeurs extrêmes et donc pas des autres observations.

1.2.2. Variance

La variance est le moment d'ordre 2 des écarts d'une variable par rapport à sa moyenne arithmétique:

$$V(X) = \frac{1}{n}\sum\left(X - \overline{X}\right)^2 = \sigma_X^2$$

Note 1. La covariance entre deux variables X et Y est:

$$S_{XY} = \frac{1}{n}\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$$

Note 2. Le moment centré d'ordre r est défini par: $\mu_r = \frac{1}{n}\sum\left(X - \overline{X}\right)^r$

Pour

- $r = 1 \qquad \mu_1 = \frac{1}{n}\sum\left(X - \overline{X}\right) = 0$
- $r = 2 \qquad \mu_2 = \frac{1}{n}\sum\left(X - \overline{X}\right)^2 = \frac{1}{n}\sum X^2 - \overline{X}^2 = M_2 - M_1^2$

Ainsi, les moments centrés peuvent s'exprimer en fonction des moments non centrés.

Ex. 4. Calculer moyenne arithmétique et variance des n premiers entiers, allant de 1 à n.

Soient:
$$S_1 = \sum X = 1 + 2 + \cdots + i + \cdots + n$$
$$S_2 = \sum X^2 = 1^2 + 2^2 + \cdots + i^2 + \cdots + n^2$$
$$S_3 = \sum X^3 = 1^3 + 2^3 + \cdots + i^3 + \cdots + n^3$$

$$S_1 = 1 + 2 + \cdots + i + \cdots + n$$
$$S_1 = n + (n-1) + \cdots + (n-(i-1)) + \cdots + 1$$

$$2S_1 = (n+1) + (n+1) + \cdots + (n+1) + \cdots + (n+1) = n(n+1)$$

d'ou $S_1 = \dfrac{n(n+1)}{2}$

Pour calculer S_2 on développe $(1+t)^3$ et on remplace successivement t par 1, 2, ⋯, i, ⋯, n

$$(1+t)^3 = 1 + 3t + 3t^2 + t^3$$

$$t = 1 \rightarrow (2)^3 = 1 + 3(1) + 3(1)^2 + (1)^3$$
$$t = 2 \rightarrow (3)^3 = 1 + 3(2) + 3(2)^2 + (2)^3$$
$$\vdots$$
$$t = i \rightarrow (1+i)^3 = 1 + 3(i) + 3(i)^2 + (i)^3$$
$$\vdots$$
$$t = n \rightarrow (1+n)^3 = 1 + 3(n) + 3(n)^2 + (n)^3$$

$$S_3 - 1^3 + (n+1)^3 = n + 3S_1 + 3S_2 + S_3 \Leftrightarrow 3S_2 = (n+1)^3 - (n+1) - 3S_1$$

Soit:

$$S_2 = \frac{1}{3}\left[(n+1)^3 - (n+1) - 3S_1\right] = \frac{1}{3}\left[(n+1)^3 - (n+1) - 3\frac{n(n+1)}{2}\right]$$

$$= \frac{n+1}{6}\left[2(n+1)^2 - 2 - 3n\right]$$

$$= \frac{n+1}{6}\left[2n^2 + n\right]$$

$$= \frac{n(n+1)(2n+1)}{6}$$

$$\overline{X} = \frac{S_1}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\sigma_X^2 = \frac{S_2}{n} - \overline{X}^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n+1}{12}\left[2(2n+1) - 3(n+1)\right] = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}$$

Ex. 5. Calculer moyenne et variance de la variable X qui prend les valeurs allant de 0 à n, avec des fréquences proportionnelles aux coefficients du binôme C_n^X respectivement.

$$\overline{X} = \frac{\sum_{X=0}^{n} XC_n^X}{\sum_{X=0}^{n} C_n^X} \quad \text{où} \qquad \sum_{X=0}^{n} C_n^X = C_n^0 + C_n^1 + \cdots + C_n^n = (1+1)^n = 2^n$$

$$\sum_{X=0}^{n} XC_n^X = 0C_n^0 + 1C_n^1 + 2C_n^2 \cdots + nC_n^n$$

$$= 0C_n^n + 1C_n^{n-1} + 2C_n^{n-2} + \cdots + (n-1)C_n^1 + nC_n^0$$

du fait que $C_n^X = C_n^{n-X}$

En faisant la somme des deux expressions identiques, on obtient:

$$\left[0C_n^0 + 1C_n^1 + 2C_n^2 \cdots + nC_n^n\right] + \left[0C_n^n + 1C_n^{n-1} + 2C_n^{n-2} + \cdots + (n-1)C_n^1 + nC_n^0\right]$$

$$= n\left[C_n^0 + C_n^1 + C_n^2 \cdots + C_n^n\right] = n\,2^n = 2\sum_{X=0}^{n} XC_n^X$$

d'où $\overline{X} = \dfrac{\sum_{X=0}^{n} X C_n^X}{\sum_{X=0}^{n} C_n^X} = \dfrac{n 2^{n} \diagup 2}{2^n} = \dfrac{n}{2}$

par analogie, on obtient: $V(X) = \sigma_X^2 = \dfrac{\sum X^2 C_n^X}{\sum C_n^X} - \overline{X}^2 = \dfrac{n(n+1)}{4} - \left(\dfrac{n}{2}\right)^2 = \dfrac{n}{4}$

On peut procéder autrement, sachant que:

$$C_n^X = \frac{n!}{X!(n-X)!} = \frac{n(n-1)!}{X(X-1)!(n-X)!} = \frac{n}{X} C_{n-1}^{X-1} \quad \text{d'où} \quad n C_{n-1}^{X-1} = X\, C_n^X$$

$$\sum_{X=1}^{n} X C_n^X = n \sum_{X=1}^{n} C_{n-1}^{X-1} = n(1+1)^{n-1} = n\, 2^{n-1}$$

$$\sum_{X=0}^{n} X C_n^X = n\left(0 C_n^0 + \sum_{X=1}^{n} X C_n^X \right) = n \sum_{X=1}^{n} C_{n-1}^{X-1} = n\, 2^{n-1}$$

d'où $\overline{X} = \dfrac{\sum_{X=0}^{n} X C_n^X}{\sum_{X=0}^{n} C_n^X} = \dfrac{n\, 2^{n-1}}{2^n} = \dfrac{n}{2}$

En utilisant toujours $C_n^X = \dfrac{n}{X} C_{n-1}^{X-1}$; nous allons chercher $X(X-1)C_n^X$ et déduire $X^2 C_n^X$.

$$C_n^X = \frac{n}{X} C_{n-1}^{X-1} = \frac{n}{X} \frac{(n-1)!}{(X-1)!(n-1)!} = \frac{n}{X} \frac{(n-1)(n-2)!}{(X-1)(X-2)!(n-X)!} = \frac{n(n-1)}{X(X-1)} C_{n-2}^{X-2}$$

d'où $n(n-1) C_{n-2}^{X-2} = X(X-1) C_n^X$

$$\sum_{X=2}^{n} X(X-1) C_n^X = n(n-1) \sum_{X=2}^{n} C_{n-2}^{X-2} = n(n-1)(1+1)^{n-2} = n(n-1) 2^{n-2}$$

or $\sum_{X=2}^{n} X(X-1)C_n^X = \sum_{X=2}^{n} X^2 C_n^X - \sum_{X=2}^{n} XC_n^X = n(n-1)2^{n-2}$ avec

$$\sum_{X=2}^{n} XC_n^X = n\sum_{X=2}^{n} C_{n-1}^{X-1} = n\,2^{n-1}$$

d'où

$$\sum_{X=2}^{n} X^2 C_n^X = n(n-1)2^{n-2} + \sum_{X=2}^{n} XC_n^X = n(n-1)2^{n-2} + n2^{n-1} = n2^{n-2}(n-1+2) = n(n+1)2^{n-2}$$

soit: $V(X) = \sigma_X^2 = \dfrac{\sum X^2 C_n^X}{\sum C_n^X} - \overline{X}^2 = \dfrac{n(n+1)2^{n-2}}{2^n} - \left(\dfrac{n}{2}\right)^2 = \dfrac{n}{4}$

1.2.3. Ecart-type

La mesure de dispersion la plus utilisée, l'écart-type est le radical de la variance ou la moyenne quadratique des écarts par rapport à la moyenne arithmétique:

$$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{1}{n}\sum \left(X - \overline{X}\right)^2}$$

1.2.4. Coefficient de variation

Le coefficient de variation est le rapport de l'écart-type à la moyenne arithmétique:

$$CV(X) = \frac{\sigma_X}{\overline{X}}$$

Il n'a pas d'unité de mesure et est invariant pour tout changement d'échelle.

Note. Le coefficient de corrélation entre X et Y est:

$$r = \frac{Cov(X, Y)}{\sigma_X\, \sigma_Y} = \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}}$$

1.2.5. Ecarts interquartiles

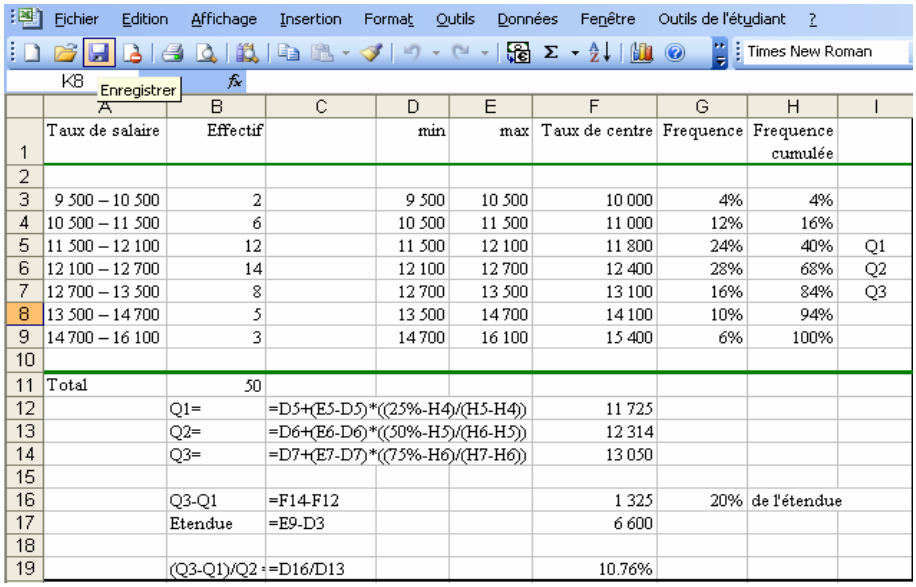
L'écart interquartile absolu est défini par: $i_{qa} = Q_3 - Q_1$

L'écart interquartile relatif est défini par: $i_{qr} = \dfrac{Q_3 - Q_1}{Q_2}$

Ex. 6. Calculer les écarts interquartiles absolu et relatif de la distribution du taux de salaire horaire de 50 travailleurs:

Taux de salaire	Effectif
9 500 – 10 500	2
10 500 – 11 500	6
11 500 – 12 100	12
12 100 – 12 700	14
12 700 – 13 500	8
13 500 – 14 700	5
14 700 – 16 100	3

En procédant sur Excel on obtient:



	A	B	C	D	E	F	G	H	I
1	Taux de salaire	Effectif		min	max	Taux de centre	Frequence	Frequence cumulée	
2									
3	9 500 – 10 500	2		9 500	10 500	10 000	4%	4%	
4	10 500 – 11 500	6		10 500	11 500	11 000	12%	16%	
5	11 500 – 12 100	12		11 500	12 100	11 800	24%	40%	Q1
6	12 100 – 12 700	14		12 100	12 700	12 400	28%	68%	Q2
7	12 700 – 13 500	8		12 700	13 500	13 100	16%	84%	Q3
8	13 500 – 14 700	5		13 500	14 700	14 100	10%	94%	
9	14 700 – 16 100	3		14 700	16 100	15 400	6%	100%	
10									
11	Total	50							
12		Q1=	=D5+(E5-D5)*((25%-H4)/(H5-H4))			11 725			
13		Q2=	=D6+(E6-D6)*((50%-H5)/(H6-H5))			12 314			
14		Q3=	=D7+(E7-D7)*((75%-H6)/(H7-H6))			13 050			
15									
16		Q3-Q1	=F14-F12			1 325	20%	de l'étendue	
17		Etendue	=E9-D3			6 600			
18									
19		(Q3-Q1)/Q2	=D16/D13			10.76%			

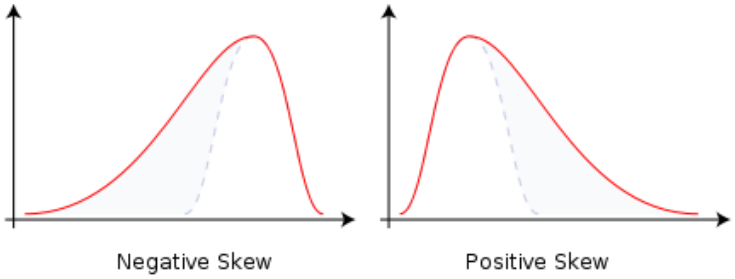
1.3. Paramètres de forme

Ce sont:
- le coefficient d'asymétrie (skewness)
- le coefficient d'aplatissement (kurtosis).

1.3.1. Coefficient d'asymétrie

Le coefficient d'asymétrie (de Fisher) – la dissymétrie – est donné par: $\gamma_1 = \frac{\mu_3}{\sigma^3}$ où μ_3 est le moment centré d'ordre 3 et σ l'écart-type.

Ainsi:
- un coefficient positif indique une distribution dont la queue est étalée vers la droite
- un coefficient négatif indique une distribution dont la queue est étalée vers la gauche
- Dans le cas d'une distribution normale, par symétrie on a: $\gamma_1 = 0..$



La définition théorique γ_1 du coefficient d'asymétrie est une mesure biaisée de l'asymétrie de la population. Un estimateur non biaisé de l'asymétrie est donné par la formule :

$$G_1 = \frac{n}{(n-1)(n-2)} \sum \left(\frac{X - \overline{X}}{\sigma} \right)^3$$

1.3.2. Coefficient d'aplatissement

On mesure l'aplatissement avec les moments centrés d'ordre pair:

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

Si

- $\beta_2 > 3$, on parle de distribution leptokurtique. La notion de leptokurticité est très utilisée dans le milieu de la finance de marché, les échantillons ayant des queues plus épaisses que la normale aux extrémités, impliquant des valeurs anormales plus fréquentes
- $\beta_2 < 3$, on parlera de distribution platikurtique
- $\beta_2 = 3$, on parle de distribution mésokurtique.

On normalise le coefficient d'aplatissement en lui soustrayant la valeur correspondant à la loi normale centrée réduite, i.e. 3:

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma^4} - 3$$

Ce coefficient d'aplatissement, γ_2 permet de comparer l'aplatissement de la distribution observée à celui d'une distribution normale de même moyenne $\overline{X}$ et de même écart-type σ.

Si

- $\gamma_2 > 0$, la distribution observée est plus pointue que la distribution normale
- $\gamma_2 < 0$, la distribution observée est plus aplatie que la distribution normale
- $\gamma_2 = 0$, la distribution observée est quasi-normale.

Un estimateur non biaisé de l'aplatissement est donné par:

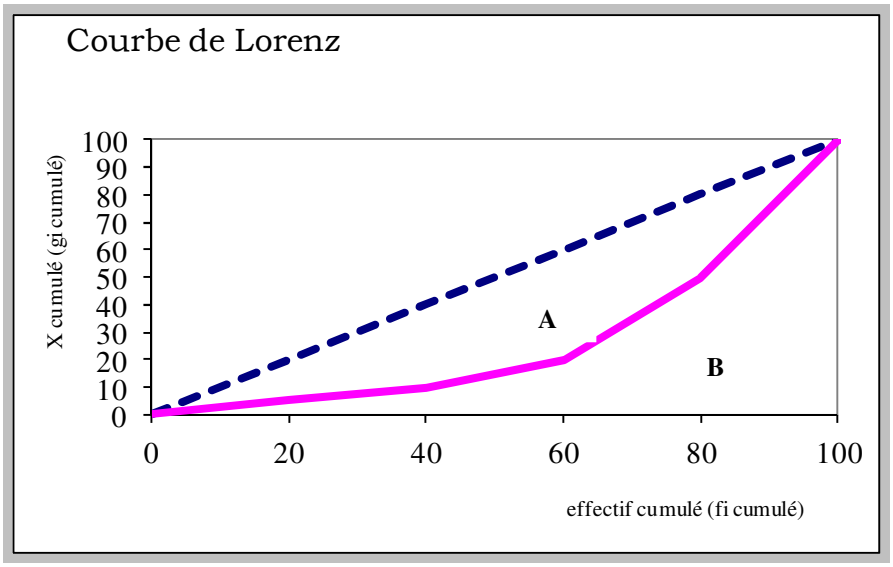
$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{X - \overline{X}}{\sigma} \right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

1.4. Paramètres de concentration

La notion de concentration s'applique uniquement à des variables continues positives pour lesquelles on dispose des effectifs et de l'importance du caractère pour chaque classe.

Soit la distribution:

Caractère (X_i)	Effectif (n_i)	Fréquence (f_i)	Fréquence (g_i)	$\sum f_i$	$\sum g_i$
X_1	n_1	$n_1 / \sum n_i$	$X_1 / \sum X_i$	f_1	g_1
X_2	n_2	$n_2 / \sum n_i$	$X_2 / \sum X_i$	$f_1 + f_2$	$g_1 + g_2$
⋮	⋮	⋮	⋮	⋮	⋮
X_i	n_i	$n_i / \sum n_i$	$X_i / \sum X_i$	$\sum_{j=1}^{i} f_j$	$\sum_{j=1}^{i} g_j$
⋮	⋮	⋮	⋮	⋮	⋮
X_n	n_n	$n_n / \sum n_i$	$X_n / \sum X_i$	$\sum_{j=1}^{n} f_j = 100$	$\sum_{j=1}^{n} g_j = 10($
$\sum X_i$	$\sum n_i$	100	100		



Sur le graphique, l'indice de Gini est le rapport de la surface A à la surface totale A+B:

$$I_G = \frac{A}{A+B} = \frac{A}{1/2} = 2A = 2\left(\tfrac{1}{2} - B\right) = 1 - 2B$$

Cet indicateur est ainsi égal, par définition, à 2 fois la surface comprise entre la courbe de Lorenz et la ligne de distribution égalitaire ou uniforme, soit: $I_G = 2\int_0^1 \left(P - L(P)\right)dP$.

Dans les applications, on ne connaît pas la forme de la fonction de Lorenz mais seulement la distribution de la variable e.g. le revenu, par "intervalles" e.g. les quintiles ou déciles de revenu. Pour n intervalles, l'indice de Gini s'obtient par la formule de Brown :

$$I_G = 1 - 2B = 1 - \sum_{j=0}^{n-1} \left(X_{j+1} - X_j\right)\left(Y_{j+1} + Y_j\right) \text{ Avec } X_0 = Y_0 = 0$$

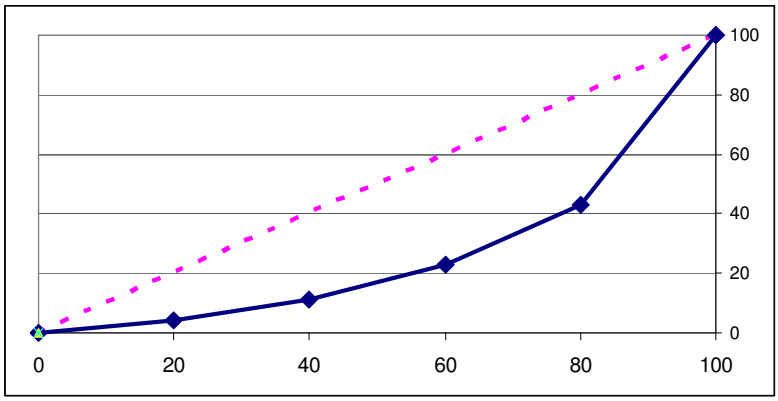
où X est le pourcentage cumulé de la population et
Y celui de la variable étudiée e.g. le revenu

Ex. 7. Soit la distribution par quintile du revenu au Mali en 2006:

Population	20%	40%	60%	80%	100%
Revenu	4%	11%	23%	43%	100%

Soit le tableau des éléments de calcul avec le graphique associé

	X_j	Y_j	$X_{j+1} - X_j$	$Y_{j+1} + Y_j$	$\left(X_{j+1} - X_j\right)\left(Y_{j+1} + Y_j\right)$
Q1	0.20	0.04	0.20	0.04	0.008
Q2	0.40	0.11	0.20	0.15	0.030
Q3	0.60	0.23	0.20	0.34	0.068
Q4	0.80	0.43	0.20	0.66	0.132
Q5	1.00	1.00	0.20	1.43	0.286
Total					0.524

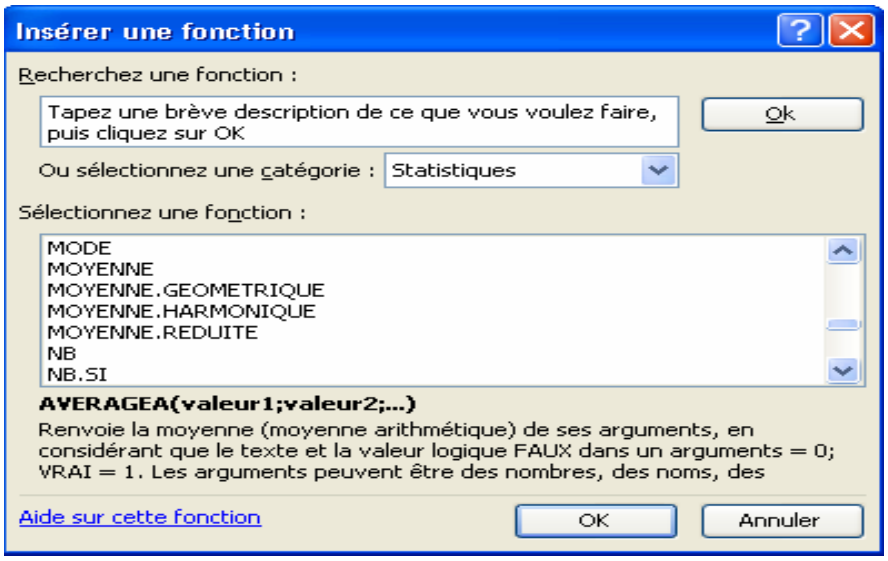


$I_G = 1 - 0.524 = 0.476 = 47.6\%$

L'indice de Gini est l'indicateur le plus utilisé, compris entre 0 et 1 quoiqu'il ne soit pas décomposable et qu'il reste sensible à tous les transferts puisqu'il pondère tous les individus de la même façon.

1.5. Applications sur EXCEL

Pour la plupart des indicateurs statistiques, il suffit d'aller dans le sous-menu "Fonction" du menu "Insertion"



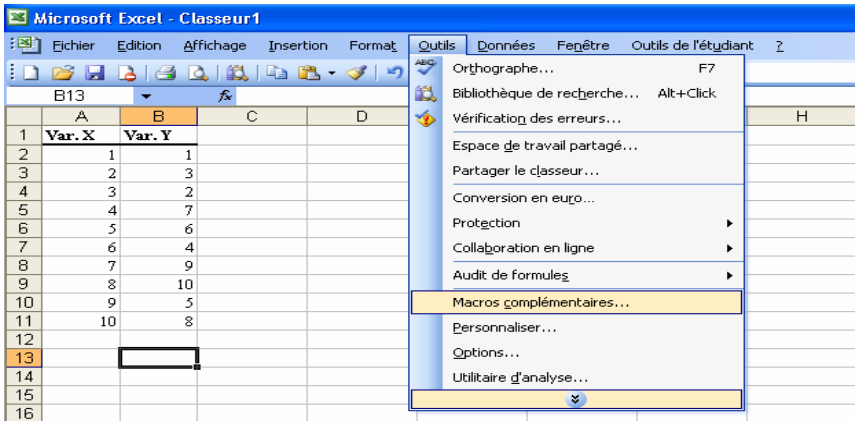
Exemple

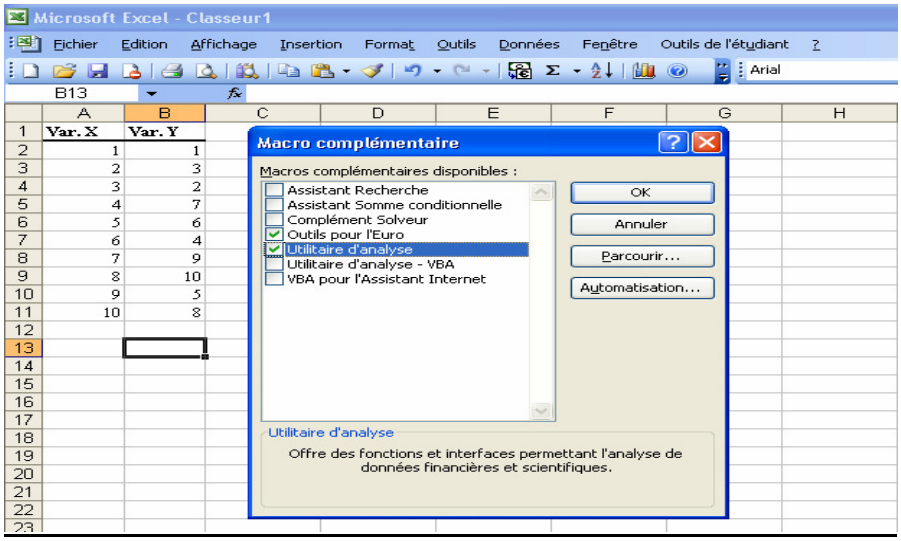
	A	B	C	D	E
1	**Var. X**		**Statitstique**	**Fonction Excel**	**Résultat**
2	1				
3	2		Moyenne arithmétique	=MOYENNE(A2:A11)	5.50
4	3				
5	4		Moyenne géométrique	=MOYENNE.GEOMETRIQUE(A2:A11)	4.53
6	5				
7	6		Moyenne harmonique	=MOYENNE.HARMONIQUE(A2:A11)	3.41
8	7				
9	8		Variance	=VAR.P(A2:A11)	8.25
10	9				
11	10		Ecart-type	=ECARTYPEP(A2:A11)	2.87
12					

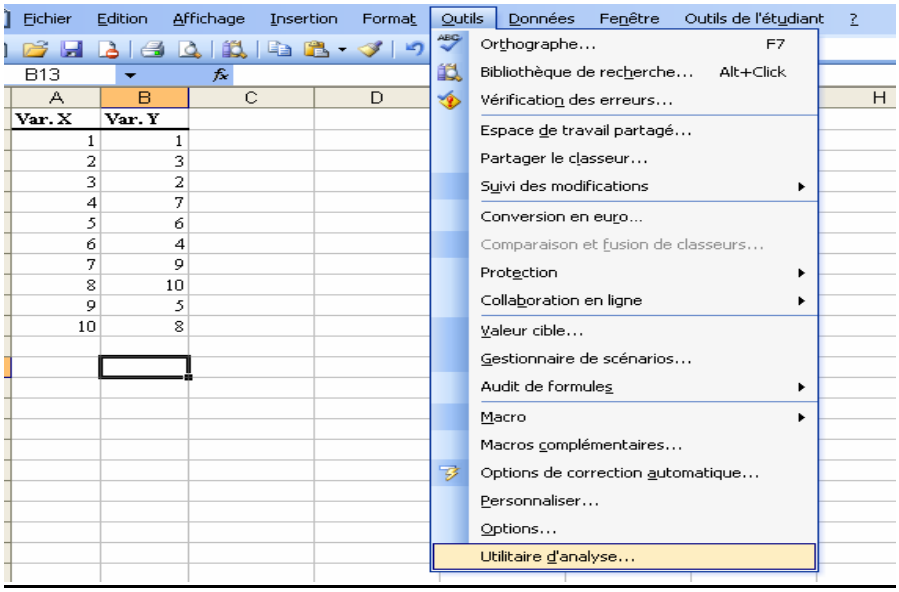
Pour certaines applications comme la régression linéaire, il faut installer des "macro complémentaires"

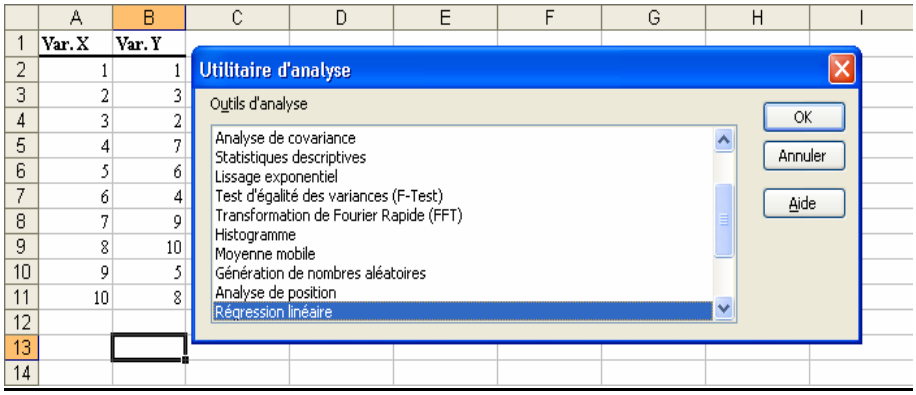
1. aller dans le menu "OUTILS" puis le sous menu "MACROS COMPLEMENTAIRES"
2. cocher "UTILITAIRE D'ANALYSE"

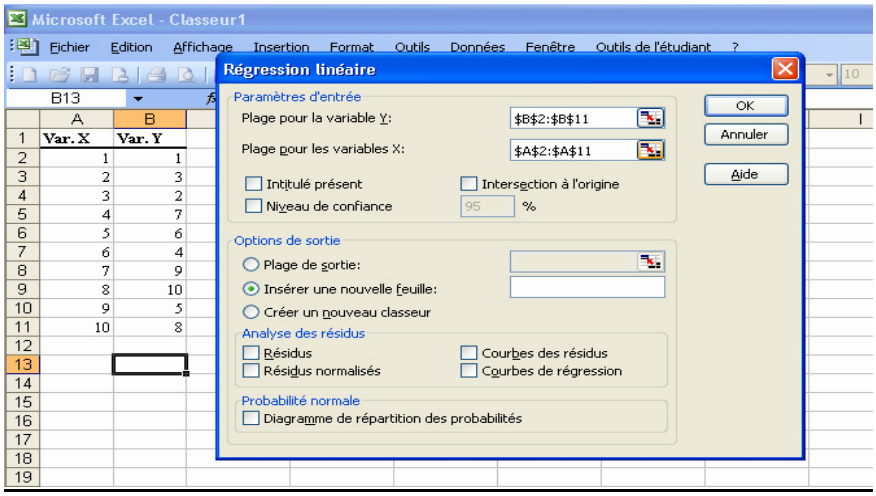
Une fois la macro complémentaire installée, seules les manipulations suivantes sont nécessaires pour réaliser une régression. Dans le menu "outils", il faut cliquer sur "utilitaire d'analyse" et cliquer sur "régression linéaire". La boîte de dialogue qui s'affiche ne nécessite pas de commentaire particulier, sauf que le modèle estimé est de type Y=aX+b. Y est donc la variable expliquée et X la variable explicative (qui peut être une matrice de variables explicatives). Il est possible d'imposer la nullité de la constante (b) en cochant "intersection à l'origine".




Microsoft Excel - Classeur1
Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant ?
Arial
B13
Var. X
Var. Y
Macro complémentaire
Macros complémentaires disponibles :
Assistant Recherche
Assistant Somme conditionnelle
Complément Solveur
Outils pour l'Euro
Utilitaire d'analyse
Utilitaire d'analyse - VBA
VBA pour l'Assistant Internet
OK
Annuler
Parcourir...
Automatisation...
Utilitaire d'analyse
Offre des fonctions et interfaces permettant l'analyse de données financières et scientifiques.


Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant ?
B13
Var. X
Var. Y
Orthographe... F7
Bibliothèque de recherche... Alt+Click
Vérification des erreurs...
Espace de travail partagé...
Partager le classeur...
Suivi des modifications
Conversion en euro...
Comparaison et fusion de classeurs...
Protection
Collaboration en ligne
Valeur cible...
Gestionnaire de scénarios...
Audit de formules
Macro
Macros complémentaires...
Options de correction automatique...
Personnaliser...
Options...
Utilitaire d'analyse...

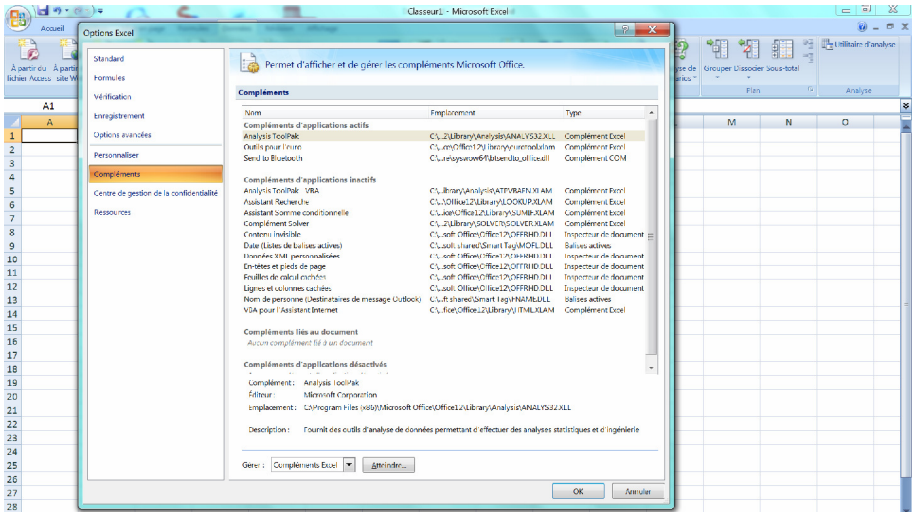

Var. X
Var. Y
Utilitaire d'analyse
Outils d'analyse
Analyse de covariance
Statistiques descriptives
Lissage exponentiel
Test d'égalité des variances (F-Test)
Transformation de Fourier Rapide (FFT)
Histogramme
Moyenne mobile
Génération de nombres aléatoires
Analyse de position
Régression linéaire
OK
Annuler
Aide


Microsoft Excel - Classeur1
Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant ?
Régression linéaire
Paramètres d'entrée
Plage pour la variable Y:
$B$2:$B$11
Plage pour les variables X:
$A$2:$A$11
Intitulé présent
Intersection à l'origine
Niveau de confiance
95
%
Options de sortie
Plage de sortie:
Insérer une nouvelle feuille:
Créer un nouveau classeur
Analyse des résidus
Résidus
Courbes des résidus
Résidus normalisés
Courbes de régression
Probabilité normale
Diagramme de répartition des probabilités
OK
Annuler
Aide

	A	B	C	D	E	F	G
1	RAPPORT DÉTAILLÉ						
2							
3	*Statistiques de la régression*						
4	Coefficient de détermination multiple	0.733333333					
5	Coefficient de détermination R^2	0.537777778					
6	Coefficient de détermination R^2	0.48					
7	Erreur-type	2.183269719					
8	Observations	10					
9							
10	ANALYSE DE VARIANCE						
11		Degré de liberté	Somme des carrés	Moyenne des carrés	*F*	Valeur critique de F	
12	Régression	1	44.36666667	44.36666667	9.307692308	0.015800596	
13	Résidus	8	38.13333333	4.766666667			
14	Total	9	82.5				
15							
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*
17	Constante	1.466666667	1.491457155	0.983378344	0.354222895	-1.972639697	4.90597303
18	Variable X 1	0.733333333	0.240370085	3.050851079	0.015800596	0.179038924	1.287627743

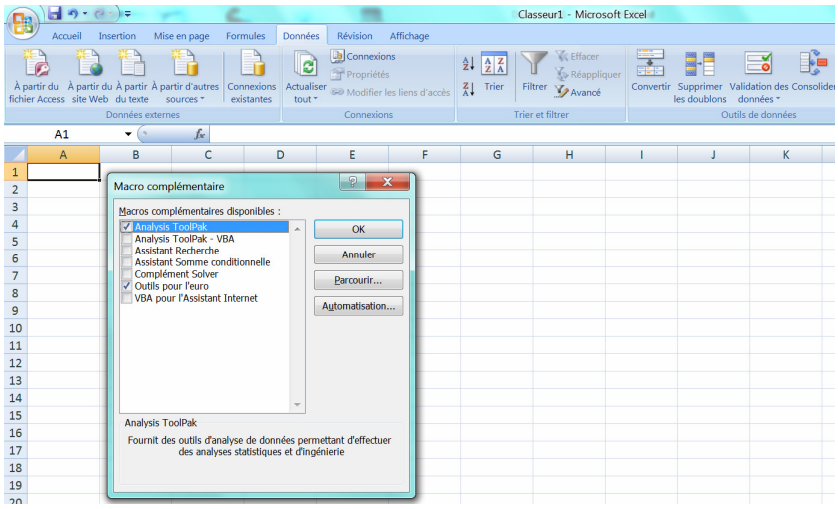
Pour utiliser Excel 2007, quelques étapes:

- aller dans "**Options d'excel**" en cliquant sur la boule en haut à gauche
- activer "**Compléments**"

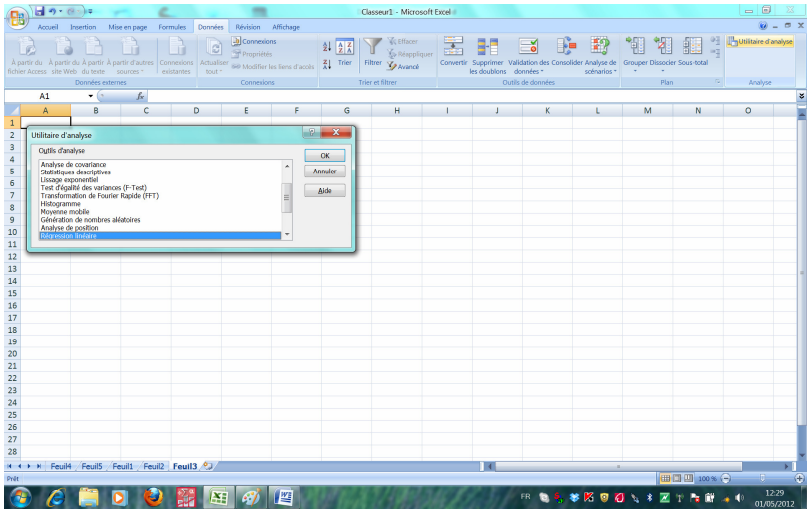


- cliquer sur "**Atteindre**"

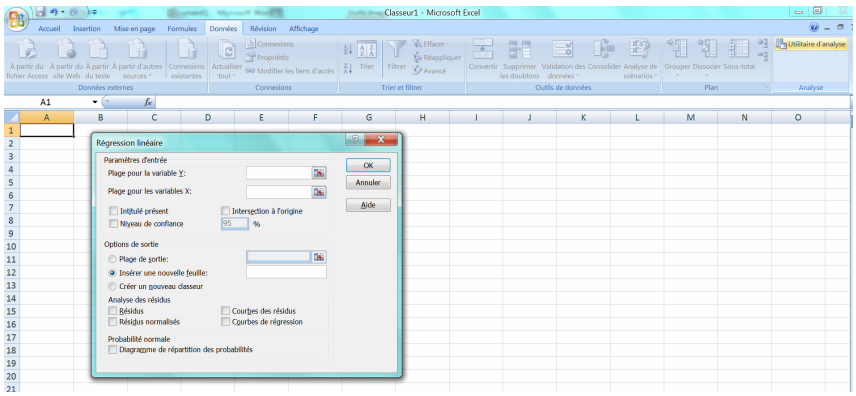
- cocher "**analysis toolpack**" puis cliquer sur "OK" pour installer l'utilitaire d'analyse et bien d'autres macro complémentaires



- aller chercher dans l'onglet "**Données**" l'utilitaire d'analyse et sélectionner "***régression linéaire***".



- utiliser la "**Régression linéaire**" à votre guise


Classeur1 - Microsoft Excel
Accueil
Insertion
Mise en page
Formules
Données
Révision
Affichage
Régression linéaire
Paramètres d'entrée
Plage pour la variable Y:
Plage pour les variables X:
Intitulé présent
Intersection à l'origine
Niveau de confiance
95
%
OK
Annuler
Aide
Options de sortie
Plage de sortie:
Insérer une nouvelle feuille:
Créer un nouveau classeur
Analyse des résidus
Résidus
Résidus normalisés
Courbes des résidus
Courbes de régression
Probabilité normale
Diagramme de répartition des probabilités

2. Echantillonnage

En mathématiques,
on ne comprend pas les choses,
on s'y habitue

John von Neumann

Les modalités de choix des unités statistiques permettent de distinguer deux types d'observation: le recensement (exhaustif) et l'enquête (partielle). Dans le premier cas, toutes les unités statistiques, sans omission ni répétition, sont observées e.g. RGPH 2009. Dans le second (celui de l'enquête par sondage), les observations ne portent que sur une partie de la population, appelée échantillon.

En règle générale, l'enquête par questionnaire est la méthode de recherche, de loin, la plus utilisée. Elle sied bien à la recherche descriptive et même interprétative lorsqu'on cherche à découvrir les conditions d'occurrence d'un phénomène. Pourtant, son choix exige du chercheur une connaissance des techniques d'échantillonnage, une formation en statistique descriptive et analyse statistique de données, une initiation aux programmes informatiques de gestion et d'analyse de données d'enquête e.g. SPSS.

2.1. Types d'échantillons

Un échantillon est un sous-ensemble d'unités statistiques prélevé sur une population dont on veut appréhender certaines caractéristiques. Sa représentativité sous-entend qu'il donne une image simplifiée mais fidèle de la population dont il est issu. Ainsi, suppose-t-on que les pauvres sont bien représentés dans les enquêtes nonobstant les difficultés que l'enquête peut rencontrer à les toucher.

L'enquête par sondage suppose un choix judicieux des individus de l'échantillon. Il existe une multitude de techniques d'échantillonnage réparties en deux classes selon qu'on adopte :

- le principe dit de la maquette i.e. reproduire le plus fidèlement possible la population mère en tenant compte de ses caractéristiques connues
- le principe du hasard i.e. laisser au hasard le choix des individus faisant partie de l'échantillon.

On distingue ainsi deux catégories d'échantillons:

- les échantillons non probabilistes
- les échantillons probabilistes.

i). Echantillons non probabilistes

(1). Echantillons accidentels, construits au "petit bonheur" e.g. les 100 premiers passants rencontrés au coin d'une rue

(2). Echantillons systématiques, où les individus sont pris à intervalle fixe dans une liste. Seul le premier individu est générateur d'aléa, les autres étant cet individu + le chiffre qui indique l'intervalle de tirage, e.g. le premier individu + 7, ce résultat + 7, et ainsi de suite.

(3). Echantillons de volontaires, souvent justifiés par la nature des questions, par exemple le cas des questions indiscrètes

(4). Echantillons par choix raisonné, où on ajoute à un noyau d'individus tous ceux qui sont en relation (d'affaires, de travail, d'amitié, etc.) avec eux, e.g. l'échantillon en boule de neige (snowball sampling)

(5). Echantillons par quota, où le principe est de reproduire le plus fidèlement possible la population à étudier. Appliqué pour la première fois, il y a plus de 50 ans, par Georges GALLUP (GALLUP est devenu aujourd'hui synonyme de sondage représentatif), ce type d'échantillon suppose :

- ☞ qu'on dégage un certain nombre de caractéristiques de la population
- ☞ qu'à l'aide d'un recensement récent ou d'un sondage probabiliste antérieur précis, on détermine comment la population se répartit suivant ces caractéristiques ou variables de contrôle

- qu'on construise l'échantillon en respectant cette répartition.

ii). Echantillons probabilistes

(6). Echantillons aléatoires simples, tous les individus ont une chance égale, connue et non nulle

(7). Echantillons en grappes ou par groupes/faisceaux, où l'échantillon n'est plus constitué d'individus mais d'ensembles, le plus homogènes possible, d'individus appelés grappes e.g. un échantillon de 20 écoles de 10 élèves chacune

(8). Echantillons aréolaires (area sampling), un cas particulier d'échantillon par grappes lorsqu'il n'existe ni liste pour base de sondage ni recensement récent pouvant conduire à la méthode des quotas. Le principe est celui de la division géographique du secteur en îlots pour retenir un échantillon d'îlots dont tous les individus seront interrogés e. g. tous les locataires d'une concession

(9). Echantillons stratifiés, la technique la plus raffinée qui comprend :
- la division de la population à étudier en sous-populations appelées strates
- le tirage au hasard d'un échantillon dans chaque strate
- l'ensemble des échantillons choisis constitue l'échantillon final qui sera soumis à l'analyse.

L'échantillon aléatoire, lorsque la taille est suffisante, est jugé le plus représentatif. Il a l'avantage de permettre au chercheur, grâce aux lois du calcul des probabilités, de préciser les risques qu'il prend. Toutefois, l'échantillon probabiliste n'est pas toujours possible et peut-être même pas nécessaire. D'ailleurs dans bien de cas, c'est un mythe car n'oublions pas qu'il suppose :

- une liste précise, récente et sûre de la population mère
- une base de sondage pertinente
- la possibilité de contacter tous les individus tirés dans l'échantillon
- suffisamment de moyens et de temps disponibles.

Pour toutes ces raisons, il n'est pas toujours possible d'obtenir un échantillon aléatoire. On peut privilégier celui par choix raisonné qui en est le plus

proche de tous les échantillons non aléatoires. Les enquêtes INSTAT utilisent des tirages à plusieurs degrés dont:

- la section d'énumération (SE), aire géographique suite à un redécoupage fictif du territoire, renfermant 800 – 1000 personnes en milieu rural et 1000 – 1500 en milieu urbain
- le ménage, groupe d'individus, apparentés ou non vivant généralement dans la même concession ou dans le même bâtiment et partageant leurs repas et mettant en commun les éléments essentiels à leur niveau de vie sous la responsabilité d'un chef dont l'autorité est reconnue par tous les membres.

2.2. Tirage de l'échantillon

Il existe plusieurs procédés de tirage d'un échantillon dont:

- le tirage élémentaire
- le tirage systématique
- le tirage par grappes
- etc.

2.2.1. Tirage élémentaire

Ce procédé de tirage consiste à tirer l'échantillon en donnant à chaque individu de la population la même probabilité non nulle d'être tiré. Il suppose:

- disposer d'une base de sondage
- numéroter les individus de 1 à N
- donner la taille n de l'échantillon
- tirer n numéros, entre 1 et N, avec une probabilité non nulle égale pour tous les N individus.

Le tirage pouvant être avec ou sans remise, on privilégie dans la pratique le tirage sans remise ou échantillon exhaustif dont les estimations sont plus précises, la variance étant toujours inférieure à celle d'un échantillon indépendant (tirage avec remise).

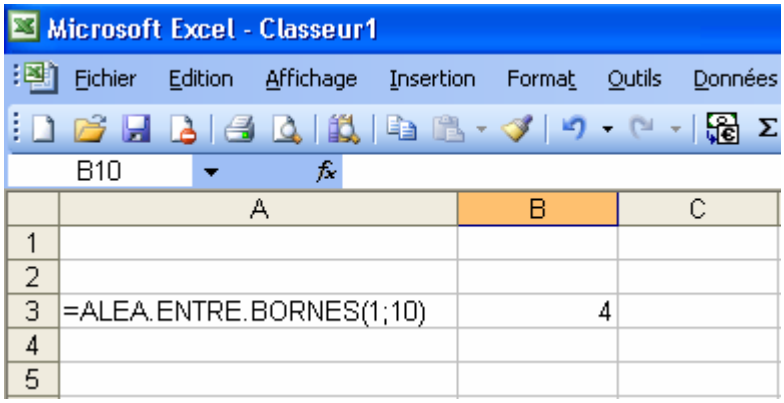
2.2.2. Tirage systématique

Pour tirer l'échantillon:

- on divise la population totale par la taille de l'échantillon $\left(\frac{N}{n}\right)$

- on tire au hasard le premier individu dans le premier lot de quotient e.g. à l'aide de la fonction Excel "=ALEA.ENTRE.BORNES$\left(1; \frac{N}{n}\right)$"
- on obtient les individus suivants par addition successive du quotient à ce premier tirage.

Par exemple, si on veut tirer 3 entiers dans les 30 premiers, allant de 1 à 30, on dénombre au total 10 lots (soit 30 divisé par 3) de 3 nombres chacun. On tire au hasard sur Excel un nombre entre 1 et 10, par exemple le numéro 4 est tiré, ce numéro correspond au premier entier inclus dans l'échantillon. Les autres entiers de l'échantillon correspondent aux numéros: 4 + 10 = 14 puis 14 + 10 = 24.



2.2.3. Tirage par grappes

A la différence du sondage élémentaire, dans le tirage par grappes, les individus de l'échantillon sont tirés non pas un à un mais par "lots" appelés "grappes" e.g. la SE est un ensemble de ménages. Pour chaque grappe tirée, on fixe le nombre d'individus devant être enquêtés e.g. 10 ménages par SE.

Ce procédé permet de simplifier l'établissement de la base de sondage et de diminuer le coût de réalisation de l'enquête sur le terrain par une économie des frais de déplacement pour un même nombre d'unités d'observations.

Le sondage aréolaire est un cas particulier de tirage par grappes, la grappe étant ici constituée par une aire déterminée par des limites géographiques aisément identifiables ne nécessitant pas de fréquentes mises à jour de la base de sondage

2.2.4. Tirage à plusieurs degrés

Ce procédé consiste à désigner les unités de l'échantillon en cascade tel que:

- au premier degré, on désigne au hasard un échantillon d'unités primaires (PSU)
- au second degré, on tire dans chaque PSU un échantillon d'unités secondaires (SSU)
- au troisième degré, on tire dans chaque SSU un échantillon d'unités tertiaires
- etc.

Par exemple, pour établir un échantillon d'enfants de moins de 5 ans, au lieu de dresser la liste de tous les enfants de moins de 5 ans et tirer l'échantillon, on peut

- au premier degré, tirer un échantillon de SE constituant les PSU
- au second degré, tirer dans chaque SE un échantillon de ménages
- au troisième degré, tirer dans chaque ménage un échantillon d'enfants de moins de 5 ans.

2.3. Taille de l'échantillon

En matière d'échantillonnage, il importe de bien en déterminer la taille (n). Celle-ci détermine la précision de l'analyse, sans pour autant être en lien direct avec la taille de la population-mère. Il existe deux méthodes de calcul de la taille de l'échantillon:

- à partir de la proportion (p) estimée de la population présentant la caractéristique étudiée

$$n = \frac{t^2}{\varepsilon^2} p(1-p) = \left(\frac{t}{\varepsilon}\right)^2 p(1-p)$$

où t est le niveau de confiance de la t-distribution e.g. $t = 1.96$ au seuil ε pour un degré de liberté de plus de 120 à l'infini

- à partir de l'écart-type estimé (σ) de la moyenne de la caractéristique étudiée

$$n = \frac{t^2}{\varepsilon^2} \sigma^2 = \left(\frac{t}{\varepsilon}\right)^2 \sigma^2$$

Ainsi, on peut établir, à titre indicatif, le tableau de calcul de la taille de l'échantillon pour une proportion d'individus présentant la caractéristique étudiée, pour $t = 2$:

Erreur d'échantillonnage ε	Proportion p (1-p)				
	50%(50%)	60%(40%)	70%(30%)	80%(20%)	90%(10%)
1%	10 000	9 600	8 400	6 400	3 600
5%	400	384	336	256	144
10%	100	96	84	64	36

2.4. Inférence statistique

Si la théorie de l'échantillonnage permet de déduire les caractéristiques des échantillons à partir de celles de la population-mère, la théorie de l'estimation (ou inférence statistique) permet d'induire les caractéristiques de la population-mère à partir des données d'échantillons représentatifs.

L'estimation ponctuelle consiste à estimer un paramètre θ inconnu par un nombre $\hat{\theta}$. L'estimation par intervalle de confiance consiste par contre à estimer un paramètre θ par un intervalle (θ_1, θ_2) contenant θ avec une certaine probabilité: $P(\theta \in (\theta_1, \theta_2)) = 1 - \varepsilon$ où ε est le seuil critique ou risque d'erreur.

L'hypothèse de base de l'inférence statistique est de considérer qu'il existe un processus inconnu qui génère les données dont on dispose. Ce processus peut être décrit par une distribution de probabilité caractérisée par certains paramètres à déterminer e.g. moyenne et variance. Le but de l'inférence est justement de faire des déductions sur les inconnues du processus partant des données d'échantillon. On obtient des estimateurs pour les lesquels on établit des intervalles de confiance ensuite on teste certaines hypothèses.

Supposons que l'on veuille estimer une caractéristique (e.g. moyenne, variance) θ d'une variable aléatoire X pour une population considérée. Soit un échantillon de taille n de la v.a. *X:* $X_1, X_2, X_3, \ldots, X_n$ et soit $\theta(X_1, X_2, \ldots, X_n)$ une fonction de valeurs X_i $(i = 1,2,\ldots,n)$, donc aléatoire aussi. On dit que $\theta(X_1, X_2, \ldots, X_n)$ est un estimateur de θ. La valeur numérique de θ que donne un échantillon particulier est appelée

estimation (ponctuelle) de θ. L'intervalle de confiance de θ est établi en construisant 2 fonctions θ_1 et θ_2 qui encadrent θ avec une certaine probabilité.

2.4.1. Propriétés des estimateurs

Un bon estimateur est sans biais et de faible dispersion. L'estimateur $\hat{\theta}$ est **sans biais** (ou sans distorsion) de θ si $E(\hat{\theta}) = \theta$, le biais ou erreur systématique étant défini par: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Note. Un estimateur $\hat{\theta}$ est asymptotiquement sans biais lorsque l'espérance de sa limite de distribution est égale à θ:

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta$$

L'estimateur $\hat{\theta}$ est dit **efficace** ou de moindre dispersion, s'il a une plus faible variance. On peut tolérer un biais moindre si on obtient une réduction de la variance. On cherchera alors à minimiser la variabilité (mesurée par la variance) de $\hat{\theta}$:

$$\min V(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2.$$

L'erreur totale (MSE) se décompose en erreur aléatoire et erreur systématique telle que:

$$E(\hat{\theta} - \theta)^2 = E\left[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)\right]^2$$

$$= E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 = V\left(\hat{\theta}\right) + B^2\left(\hat{\theta}\right)$$

Note. Un estimateur asymptotiquement efficace est un estimateur consistant pour qui la variance de sa limite de distribution est plus petite que la variance de la distribution de tout autre estimateur consistant.

L'estimateur $\hat{\theta}$ est dit estimateur **convergent** de θ si $V\left(\hat{\theta}\right) \to 0$ quand n croit indéfiniment**.**

Si l'on construit un estimateur $\hat{\theta}_j$ pour chaque échantillon; la séquence des $\hat{\theta}_j$ est dite former une séquence consistante si $\forall \delta$ et ε deux nombres

positifs suffisamment petits, on peut trouver une taille N / pour tout n > N:

$$P\left(\left|\hat{\theta}_j - \theta\right| < \delta\right) > 1 - \varepsilon$$

2.4.2. Estimateurs de la moyenne d'échantillon

Soit une population de taille N individus. On choisit un échantillon de taille n, les résultats dépendant selon que le tirage est avec ou sans remise.

Soit la v.a. X telle que:

$$E(X) = \frac{1}{N}\sum_{1}^{N} X = m$$

$$V(X) = \frac{1}{N}\sum_{1}^{N}(X - m)^2 = \sigma^2$$

i). Tirage avec remise ou échantillon indépendant

Pour la moyenne de l'échantillon, $\overline{X}$, on a:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{1}^{n} X\right) = \frac{1}{n}\sum_{1}^{n} E(X) = \frac{nm}{n} = m$$

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{1}^{n} X\right) = \frac{1}{n^2}\sum V(X) = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

ii). Tirage sans remise ou échantillon exhaustif

On considère l'échantillon E et on définit la variable ε_i de Bernoulli :

$$\begin{cases} \varepsilon_i = 1 \quad si\ l'individu\ \ i \in E \\ \varepsilon_i = 0 \quad si\ l'individu\ i \notin E \end{cases}$$

i	1	2	…	i	…	n	n+1	n+2	…	N
ε_i	1	1	…	1	…	1	0	0	…	0

$$E(\varepsilon) = \frac{1+1+.\cdots+1+0+\cdots+0}{N} = 1\frac{n}{N} + 0\left(1-\frac{n}{N}\right) = \frac{n}{N}$$

La moyenne de l'échantillon est: $\overline{X} = \frac{1}{n}\sum^{n} X = \frac{1}{n}\sum^{N} X\varepsilon$

$$E(\overline{X}) = E\left(\frac{1}{n}\sum^{N} X\varepsilon\right) = \frac{1}{n}\sum^{N} XE(\varepsilon) = \frac{E(\varepsilon)}{n}\sum^{N} X = \frac{1}{n}\frac{n}{N}\sum^{N} X = \frac{1}{N}\sum^{N} X = \frac{1}{N}Nm = m$$

$$V(\overline{X}) = E(\overline{X}-m)^2 = E\left[\left(\frac{1}{n}\sum^{N} X\,\varepsilon - m\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{1}^{N}(X-m)\varepsilon\right)^2\right]$$

$$= \frac{1}{n^2}E\left[\sum (X_i-m)^2 \varepsilon_i^{\,2} + \sum_i\sum_j (X_i-m)(X_j-m)\varepsilon_i\varepsilon_j\right]$$

$$= \frac{1}{n^2}\sum (X_i-m)^2 E(\varepsilon_i^{\,2}) + \frac{1}{n^2}\sum\sum (X_i-m)(X_j-m)E(\varepsilon_i\varepsilon_j)$$

$E(\varepsilon_i^{\,2}) = \frac{n}{N}$ et

$$\varepsilon_i\varepsilon_j = \begin{cases} 1 & si\ \ i,j \in E \\ 0 & \sin on \end{cases}$$

$$P_{ij} = \begin{cases} \frac{n}{N}\frac{n-1}{N-1}\ puisque\ tirage\ sans\ remise \\ 1-\frac{n}{N}\frac{n-1}{N-1}\ \ \sin on \end{cases}$$

$$E(\varepsilon_i \varepsilon_j) = \sum \varepsilon_i \varepsilon_j P_{ij} = 1\frac{n}{N}\frac{n-1}{N-1} + 0\left(1 - \frac{n}{N}\frac{n-1}{N-1}\right) = \frac{n}{N}\frac{n-1}{N-1}$$

$$V(\overline{X}) = \frac{1}{n^2}\frac{n}{N}\sum(X-m)^2 + \frac{1}{n^2}\frac{n}{N}\frac{n-1}{N-1}\sum_i\sum_j(X_i - m)(X_j - m)$$

$$= \frac{1}{nN}\sum(X-m)^2 + \frac{1}{nN}\frac{n-1}{N-1}\sum_i\sum_j(X_i - m)(X_j - m)$$

$$= \frac{1}{nN}\frac{n-1}{N-1}\left(\sum(X-m)^2 + \sum\sum(X_i - m)(X_j - m)\right) + \frac{1}{nN}\left(1 - \frac{n-1}{N-1}\right)\sum(X-m)^2$$

or $$\left(\sum(X-m)^2 + \sum\sum(X_i - m)(X_j - m)\right) = \left(\sum(X-m)\right)^2 = 0$$

d'où:

$$V\left(\overline{X}\right) = \frac{1}{nN}\left(1 - \frac{n-1}{N-1}\right)\sum(X-m)^2 = \frac{N-n}{n\,(N-1))}\frac{\sum(X-m)^2}{N} = \frac{N-n}{N-1}\frac{\sigma 2}{n}$$

Bref $E(\overline{X}) = m$

$$V(\overline{X}) = \begin{cases} \dfrac{\sigma^2}{n} & si\ \acute{e}chantillon\ ind\acute{e}pendant \\ \dfrac{N-n}{N-1}\dfrac{\sigma^2}{n} & si\ \acute{e}chantillon\ exhaustif\ o\grave{u} \end{cases}$$

$\frac{N-n}{N-1}$ est dit coefficient d'exhaustivité (CE)

Puisque CE < 1: à tailles égales, la moyenne d'un échantillon exhaustif est un estimateur plus efficace de la moyenne de la population que celle d'un échantillon indépendant.

2.4.3. Estimateurs de la variance d'échantillon

La variance de la population est: $\sigma^2 = \frac{1}{N}\sum_1^N(X-m)^2$

La variance de l'échantillon est:

$$s^2 = \frac{1}{n}\sum_{1}^{n}(X-\overline{X})^2 = \frac{1}{n}\sum((X-m)-(\overline{X}-m))^2 = \frac{1}{n}\sum(X-m)^2 - (\overline{X}-m)^2$$

L'estimateur de la variance d'échantillon est donné par:

$$E(s^2) = \frac{1}{n}\sum E(X-m)^2 - E(\overline{X}-m)^2 = V(X) - V(\overline{X})$$

i). Echantillon indépendant

$$E(s^2) = V(X) - V(\overline{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Donc l'estimateur sans biais de la variance de la population n'est pas s^2 mais:

$$s^{*2} = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum(X-\overline{X})^2 \quad \Rightarrow \quad V(\overline{X}) = \frac{\sigma^2}{n} = \frac{s^{*2}}{n} = \frac{s^2}{n-1}$$

ii). Echantillon exhaustif

$$E(s^2) = V(X) - V(\overline{X}) = \sigma^2 - \frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N}{n}\frac{n-1}{N-1}\sigma^2$$

L'estimateur sans biais est donc:

$$s^{**2} = \frac{n}{N}\frac{N-1}{n-1}s^2 = \frac{N-1}{N}\frac{1}{n-1}\sum(X-\overline{X})^2 = \frac{N-1}{N}s^{*2} \text{ d'où}$$

$$V(\overline{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N-n}{N-1}\frac{N-1}{N}\frac{s^{*2}}{n} = \frac{N-n}{N}\frac{s^{*2}}{n}$$

$$= \frac{N-n}{N}\frac{s^{*2}}{n} = \frac{N-n}{N}\frac{s^2}{n-1}$$

Note 1. Si l'échantillonnage est un passage de l'univers à l'échantillon, l'inverse est appelé extrapolation. Le coefficient d'extrapolation (N/n) est égal à l'inverse du taux de sondage (n/N).

Note 2. La précision, en terme de variance de $\overline{X}$, est essentiellement liée au nombre d'unités enquêtées n, et du tirage relativement peu au taux de sondage n/N (pas du tout dans le cas avec remise).

Ex.8. Soit E la population des 5 premiers entiers, allant de 1 à 5

1. Calculer la moyenne (m) et l'écart-type (σ) de cette population
2. Quels sont les échantillons de taille 2 qui peuvent en être extraits avec remise
3. Calculer les moyennes $\left(\overline{X}\right)$ et variances $\left(s^2\right)$ de ces échantillons
4. Calculer moyenne $\left(\mu\right)$ et écart-type $\left(s\right)$ de la distribution d'échantillonnage des moyennes puis de la distribution d'échantillonnage des variances
5. Et si le tirage était sans remise

1. $$m = \frac{\sum X}{N} = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2} = \frac{5+1}{2} = 3$$

$$\sigma^2 = \frac{\sum X^2}{N} - m^2 = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N+1}{12}(2(2N+1) - 3(N+1)) = \frac{N^2-1}{12} = 2$$

$$\sigma = \sqrt{2} = 1.41$$

2., 3. et 4.
On construit sur Excel les $5*5 = 25$ échantillons avec remise et on procède aux calculs:

$$\mu = \frac{1}{25}\sum_{i=1}^{25}\overline{X_i} = 3 = m$$

$$s^2 = \frac{1}{25}\sum_{i=1}^{25}\overline{X}_i^2 - \mu^2 = 1 = \frac{\sigma^2}{n} \Leftrightarrow s = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$$

Echantillon	X_1	X_2	$\overline{X} = \frac{\sum X}{n} = \frac{X_1 + X_2}{2}$	$s^2 = \frac{1}{n}\sum X^2 - \overline{X}^2$
1	1	1	1,0	0,00
2	1	2	1,5	0,25
3	1	3	2,0	1,00
4	1	4	2,5	2,25
5	1	5	3,0	4,00
6	2	1	1,5	0,25
7	2	2	2,0	0,00
8	2	3	2,5	0,25
9	2	4	3,0	1,00
10	2	5	3,5	2,25
11	3	1	2,0	1,00
12	3	2	2,5	0,25
13	3	3	3,0	0,00
14	3	4	3,5	0,25
15	3	5	4,0	1,00
16	4	1	2,5	2,25
17	4	2	3,0	1,00
18	4	3	3,5	0,25
19	4	4	4,0	0,00
20	4	5	4,5	0,25
21	5	1	3,0	4,00
22	5	2	3,5	2,25
23	5	3	4,0	1,00
24	5	4	4,5	0,25
25	5	5	5,0	0,00
	=MOYENNE(D2:D26)		3	1,00
	=ECARTYPEP(D2:D26)		1	1,16

5. Si on procède sans remise, le nombre d'échantillons devient: $C_5^2 = 10$ qu'on construit sur Excel puis on procède aux calculs:

$$\mu = \frac{1}{10}\sum_{i=1}^{10} \overline{X_i} = 3 = m$$

$$s^2 = \frac{1}{10}\sum_{i=1}^{10} \overline{X_i}^2 - \mu^2 = 0.75 = \frac{\sigma^2}{n}\frac{N-n}{N-1} \Leftrightarrow s = \sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{2}{2}}\sqrt{\frac{5-2}{5-1}} = \sqrt{\frac{3}{4}} = \sqrt{0.75} = 0.87$$

Echantillon	X_1	X_2	$\overline{X} = \frac{\sum X}{n} = \frac{X_1 + X_2}{2}$	$S^2 = \frac{1}{n}\sum X^2 - \overline{X}^2$
1	1	2	1,50	0,25
2	1	3	2,00	1,00
3	1	4	2,50	2,25
4	1	5	3,00	4,00
5	2	3	2,50	0,25
6	2	4	3,00	1,00
7	2	5	3,50	2,25
8	3	4	3,50	0,25
9	3	5	4,00	1,00
10	4	5	4,50	0,25
	=MOYENNE(D2:D26)		3,00	1,25
	=ECARTYPEP(D2:D26)		0,87	1,17

Ex.9. L'INSTAT a tiré un échantillon exhaustif de 25 mille ménages pour le pays qui compte 2.5 millions de ménages. Sur cet échantillon, on a observé un revenu moyen par tête de 200 mille fcfa, avec un écart-type de 100. Déterminer l'intervalle de confiance se rapportant à l'estimation du revenu moyen par habitant.

Bien que l'échantillon ait été tiré sans remise, on peut en pratique, en raison de la faiblesse du taux de sondage $\left(\frac{25}{2500} = 1\%\right)$ l'assimiler à un échantillon indépendant, dès lors que $\frac{N-n}{N-1} = \frac{2500000-25000}{2500000-1} = \frac{2475000}{2499999} \equiv 1$

La moyenne de l'échantillon suit une loi normale de moyenne m et de variance:

$$V(\overline{X}) = \frac{\sigma^2}{n} = \frac{100^2}{25000} = \frac{2}{5} = 0.4$$

La variable $t = \frac{\overline{X} - m}{\sigma/\sqrt{n}} = \frac{\overline{X} - m}{\sqrt{0.4}}$ a une distribution normale centrée réduite

Au seuil $\varepsilon = 5\%$ $t = 1.96$

$$P\left(\overline{X} - t\frac{\sigma}{\sqrt{n}} \le m \le \overline{X} + t\frac{\sigma}{\sqrt{n}}\right) = 1 - \varepsilon$$

Soit: $199 \le m \le 201$

3. Lois de répartition et tests usuels

Les propositions mathématiques
sont reçues comme vraies parce que
personne n'a intérêt qu'elles soient fausses

Montesquieu

3.1. Lois de répartition

3.1.1. Loi Normale (Laplace-Gauss)

Elle est une loi limite pour beaucoup de lois de probabilités (e.g. la loi binominale) dans certaines conditions de généralisation. Elle s'applique aux phénomènes engendrés par un grand nombre de causes indépendantes et additives et dont l'effet d'aucune n'est prépondérant.

La variable aléatoire X → N(m, σ) si elle a pour densité de probabilité (ou de répartition):

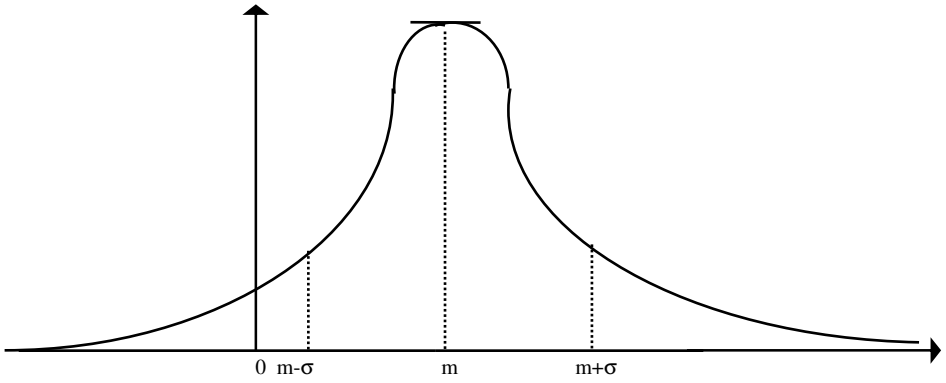
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{X-m}{\sigma}\right)^2\right] \quad avec \quad X \in \Re \quad et \quad \sigma > 0$$

$$f'(X) = \frac{-1}{\sigma\sqrt{2\pi}} \frac{X-m}{\sigma} \frac{1}{\sigma} \exp(\,.\,)$$

$$= \frac{-1}{\sigma^3\sqrt{2\pi}} (X-m)\exp(\,.\,) = 0 \quad \rightarrow \quad X = m$$

$$f(m) = \frac{1}{\sigma\sqrt{2\pi}}$$

$$f''(X) \propto \left[1-\left(\frac{X-m}{\sigma}\right)^2\right] \exp(\,.\,) = 0 \quad \rightarrow \quad X = m \pm \sigma \text{ 2 points d'inflexion.}$$



Ex.10 Etablir le moment d'ordre 2k $\left(m_{2k} = \int_{-\infty}^{+\infty} X^{2k} f(X) dX \right)$

- pour X $\rightarrow$ N(0,1) variable centrée réduite
- pour X $\rightarrow$ N(m, σ) loi normale $\left(\int_{-\infty}^{+\infty} (X-m)^{2k} f(X) dX \right)$

3.1.2. Répartition Gamma (Γ) et Beta (B)

Soit l'intégrale d'Euler: $\Gamma(P) = \int_0^{\infty} t^{p-1} e^{-t} dt \quad (1)$

En intégrant (1) par parties, on obtient:

$$\Gamma(P) = (P-1)\Gamma(P-1) = (P-1)! \text{ avec } \Gamma(1) = 1$$

Posons $t = aX,\ a > 0,\ (1)$ devient: $\Gamma(P) = a^p \int_0^{\infty} X^{p-1} e^{-aX} dX$

i). On dira qu'une variable X suit la loi Gamma **Γ**(p, a) si elle a pour densité: $f(X) = \frac{a^p}{\Gamma(P)} X^{p-1} e^{-aX} \quad pour \quad X \geq 0 \quad a, p > 0$

Note (Si p entier) ***loi d'Erlang*** avec des applications dans les problèmes de files d'attente où P(X > x) est la probabilité d'attendre plus de x minutes avant la P^{è} apparition du phénomène.

Ex.11 Montrer que espérance et variance de X $\rightarrow \Gamma$(p, a)

sont: $E(X) = \frac{P}{a} \quad et \quad V(X) = \frac{p}{a^2}$

ii). Soient 2 variables aléatoires indépendantes: X $\rightarrow \mathbf{\Gamma}$(p, 1) et Y$\rightarrow \mathbf{\Gamma}$(q, 1)

a) La variable $t = \frac{X}{X+Y}$ suit une loi Beta de première espèce de paramètres p et q, de densité:

$$f(t) = \frac{1}{B(p,q)} t^{p-1}(1-t)^{q-1} \qquad t \in [0,1] \text{ où}$$

$$B(p,q) = \int_0^1 X^{p-1}(1-X)^{q-1} dX$$

b) La variable $Z = \frac{X}{Y}$ suit une loi Beta de seconde espèce de densité:

$$f(Z) = \frac{1}{B(p,q)} Z^{p-1}(1+Z)^{-(p+q)} \quad avec \quad Z \in \Re \text{ et}$$

$$B(p,q) = \int_0^{\infty} X^{p-1}(1+X)^{-(p+q)} d\,X$$

Ex.12

1) En intégrant par parties, montrer que: $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$

2) Montrer que: $E(t) = \frac{p}{p+q}$

3.1.3. Loi χ2 (K. Pearson)

Elle est une fonction de répartition continue, définie sur (0, ∞) ayant pour densité:

$$f(X) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} X^{\frac{n}{2}-1} e^{-\frac{X}{2}} \qquad (1) \quad avec \quad X > 0$$

Soient n v.a. normales centrées réduites, indépendantes:

$X_k \to N(0,1) \quad k = \overline{1,n}$

$$f(X_k) = \frac{1}{\sqrt{2\pi}} e^{\frac{-X^2}{2}}$$

La v.a. Khi-deux est définie comme étant la somme des carrés des X:

$$\chi^2 = \sum X_k^2$$

Notons $F_k(y)$ la fonction de répartition de X_k^2

$$F_k(y) = P(X_k^2 < y) = P(-\sqrt{y} \le X_k \le \sqrt{y}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{\sqrt{y}} e^{\frac{-t^2}{2}} dt = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{\frac{-t^2}{2}} dt$$

La densité s'obtient en dérivant $F_k(y)$ par rapport à y:

$$F_k'(Y) = f_k(Y) = \frac{2}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \Bigg|_0^{\sqrt{Y}}$$

$$= \frac{2}{\sqrt{2\pi}} e^{\frac{-Y}{2}} d\sqrt{Y} = \frac{1}{\sqrt{2\pi}} Y^{\frac{-1}{2}} e^{\frac{-Y}{2}} = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}} Y^{\frac{1}{2}-1} e^{\frac{-1}{2}Y}}{\sqrt{\pi}} = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} Y^{\frac{1}{2}-1} e^{\frac{-1}{2}Y}$$

En effet $\Gamma(\frac{1}{2}) = \sqrt{\pi}$; puisque:

$f(X) = \frac{1}{\sqrt{2\pi}} e^{\frac{-X^2}{2}}$ est une densité de probabilité alors:

$\int_{-\infty}^{+\infty} f(X)dX = 1$ i.e.

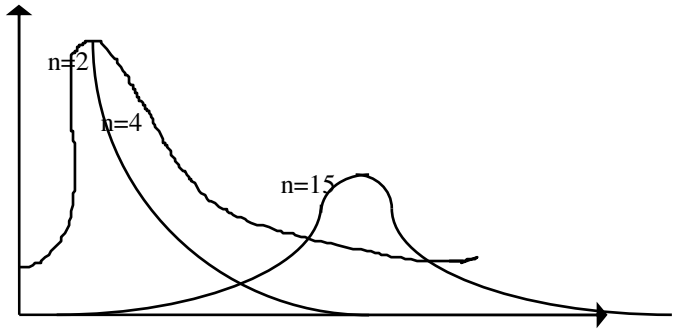
$$\int_{-\infty}^{+\infty} f(X)dX = 2\int_0^{\infty} f(X)dX = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{\frac{-X^2}{2}} dX = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \left(\frac{X^2}{2}\right)^{\frac{1}{2}-1} e^{\frac{-X^2}{2}} d\left(\frac{X^2}{2}\right) \frac{1}{\sqrt{2}}$$

$$= \frac{1}{\sqrt{\pi}} \int_0^{\infty} t^{\frac{1}{2}-1} e^{-t} dt = \frac{\Gamma\left(\frac{1}{2}\right)}{\sqrt{\pi}} = 1 \Rightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$X_k^2 \to \Gamma\left(\frac{1}{2}\right) \Rightarrow \chi^2 = \sum X_k^2 \to \Gamma\left(\frac{n}{2}\right) \qquad f_n(Y) = \prod f_k(Y) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} Y^{\frac{n}{2}-1} e^{-\frac{Y}{2}}$$

ou encore $f(\chi^2) = \frac{1}{2^{\frac{n}{2}}} \Gamma(\frac{n}{2}) \chi^{n-2} e^{\frac{-\chi^2}{2}}$

Note. La distribution χ^2 est dissymétrique avec étalement vers la droite. Elle tend à devenir symétrique lorsque le nombre n de ddl augmente pour être assimilée à la loi normale lorsque n est supérieur à 30.



Ex.13 Montrer que $E(\chi^2) = n \quad et \quad V(\chi^2) = 2n.$

3.1.4. Répartition Student (William Gosset)

Soit $X_k \to N(0, \sigma)$ indépendante, avec $k = \overline{1,n}$

La variable $t = \frac{X}{\sqrt{\frac{\sum X^2}{n}}} \to t(n)$ de densité:

$$f_n(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad (1)$$

$-\infty < t < \infty$

Posons $Y = \sum X_k^2$ alors $t = \dfrac{X}{\sqrt{\dfrac{\sum X^2}{n}}} = \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ où $Y \to \chi^2(n, \sigma)$

La probabilité élémentaire pour le point (x, y) de tomber dans l'intervalle *dx* pour x et *dy* pour y est définie par:

$$P(.) = \frac{1}{2^{\frac{n}{2}} \sigma^n \Gamma\left(\frac{n}{2}\right)} Y^{\frac{n}{2}-1} e^{-\frac{Y}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{X^2}{2\sigma^2}} dX\, dY$$

$$= \frac{1}{\sigma^{n+1} 2^{\frac{n+1}{2}} \sqrt{\pi}\, \Gamma\left(\frac{n}{2}\right)} Y^{\frac{n}{2}-1} e^{-\frac{Y+X^2}{2\sigma^2}} dX\, dY$$

$$De \quad t = \frac{X}{\sqrt{\frac{Y}{n}}} \quad \to \quad X = t\sqrt{\frac{Y}{n}} \quad \text{et} \quad dX = \sqrt{\frac{Y}{n}} dt$$

alors: $$P(.) = \frac{1}{\sigma^{n+1} 2^{\frac{n+1}{2}} \sqrt{\pi}\Gamma\left(\frac{n}{2}\right)} Y^{\frac{n}{2}-1} e^{-\frac{Y+\frac{t^2 Y^2}{n}}{2\sigma^2}} \sqrt{\frac{Y}{n}}\, dt\, dY$$

$$f(t) = \int_0^\infty P(.)\, dY = \frac{1}{\sigma^{n+1} 2^{\frac{n+1}{2}} \sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \int_0^\infty Y^{\frac{n-1}{2}} e^{-Y\frac{1+\frac{t^2}{n}}{2\sigma^2}} dY$$

Soit le changement de variable:

$$Z = Y\frac{1+\frac{t^2}{n}}{2\sigma^2} \quad \text{d'où:} \quad f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Ex.14 Montrer que $V(t) = E(t^2) = \frac{n}{n-2}$

3.1.5. Fisher-Snedecor

On appelle variable F, de Fisher-Snedecor, la variable quotient de 2 χ^2 indépendantes, divisée chacune par son ddl.

Soit $F(m,n) = \frac{Y/m}{X/n}$ *où* $X \mapsto \chi^2_n$ *et* $Y \mapsto \chi^2_m$

La variable F, à m et n ddl, a pour densité de probabilité:

$$f_F(t) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} t^{\frac{m}{2}-1}\left(1+\frac{m}{n}t\right)^{-\frac{m+n}{2}} \quad avec \geq 0$$

Posons $\begin{cases} \frac{X}{n} = u \quad \Rightarrow \quad X = nu \quad et \quad dX = n\,du \\ \frac{Y}{m} = v \quad \Rightarrow \quad Y = mv \quad et \quad dY = m\,dv \end{cases}$ *X et Y ont pour densité*:

$$f(x) = \frac{n}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}(nu)^{\frac{n}{2}-1} e^{-\frac{nu}{2}} = \frac{n^{\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} u^{\frac{n}{2}-1} e^{-\frac{nu}{2}}$$

$$f(y) = \frac{m}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}(mv)^{\frac{m}{2}-1} e^{-\frac{mv}{2}} = \frac{m^{\frac{m}{2}}}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} v^{\frac{m}{2}-1} e^{-\frac{mv}{2}}$$

Soit $\frac{v}{u} = t \quad \Rightarrow \quad v = ut \quad et \quad dv = u\,dt$

La probabilité pour (u, v) de tomber dans l'intervalle du et dv est définie par:

$$P(.) = f(u) f(v)\, du\, dv$$

$$P(.) = \frac{n^{\frac{n}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} u^{\frac{n}{2}-1} e^{-\frac{nu}{2}} \frac{m^{\frac{m}{2}}}{2^{\frac{m}{2}} \Gamma\left(\frac{m}{2}\right)} (tu)^{\frac{m}{2}-1} e^{-\frac{mtu}{2}} .u.du.dt$$

$$f(t) = \int_0^{\infty} P(.)\, du = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}}}{2^{\frac{n+m}{2}} \Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} t^{\frac{m}{2}-1} \int_0^{\infty} u^{\frac{n+m}{2}-1} e^{-\frac{u}{2}(n+mt)}\, du$$

Soit le changement de variable: $\frac{u}{2}(n+mt) = z \Leftrightarrow$

$$u = \frac{2z}{n+mt} \quad \Rightarrow \quad du = \frac{2}{n+mt} dz$$

$$f(t) = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}}}{2^{\frac{n+m}{2}} \Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} t^{\frac{m}{2}-1} \int_0^{\infty} \left(\frac{2}{n+mt}\right)^{\frac{n+m}{2}-1} z^{\frac{n+m}{2}-1} e^{-z} \frac{2dz}{n+mt}$$

$$= \frac{n^{\frac{n}{2}} m^{\frac{m}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} (n+mt)^{-\frac{n+m}{2}} t^{\frac{m}{2}-1} \underbrace{\int_0^{\infty} z^{\frac{n+m}{2}-1} e^{-z} dz}_{\Gamma\left(\frac{n+m}{2}\right)}$$

$$f(t) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} t^{\frac{m}{2}-1} \left(1+\frac{m}{n}t\right)^{-\frac{n+m}{2}}$$

$t \mapsto F(m,n)$

Ex.15 Calculer $\mu_r = \int_0^\infty t^r f(t)\,dt$ et en déduire espérance et variance de Fisher.

3.2. Tests usuels

On représente une distribution empirique par une loi de probabilité théorique. L'hypothèse que la population étudiée suit une loi de répartition est testée pour savoir si les écarts constatés entre la distribution réelle et la distribution théorique sont acceptables. Ce sont des tests de légitimité.

Soit une v.a. $X \rightarrow L(\theta)$. On suppose le paramètre $\theta \in \Re$ est réparti entre 2 sous ensembles Θ_0 et Θ_1. Le test consiste à décider sur la base d'un échantillon *$(X_1, ..., Xn)$* de la loi de X, soit:

D_0: accepter $H_0 : \theta \in \Theta_0$

D_1: refuser H_0 : i.e. accepter $H_1 : \theta \in \Theta_1$

On distingue ainsi 2 sous ensembles de réalisation:

- W les réalisations *$(X_1, ..., Xn)$* pour lesquelles on refuse Ho
- $\overline{W}$ les réalisations de *$(X_1, ..., Xn)$* pour lesquelles on accepte Ho.

Ces 2 sous ensembles définissent 2 régions:

- W région critique du test
- $\overline{W}$ région d'acceptation du test.

La décision entraînera 2 erreurs possibles:

- (erreur de première espèce) prendre la décision D_1 alors que Ho est vraie
- (erreur de seconde espèce) prendre la décision Do alors que H_1 est vraie.

Il existe plusieurs méthodes de construction du test, méthodes plus ou moins fonction des conséquences attendues des 2 types d'erreurs.

3.2.1. Tests du khi-deux ($\chi 2$)

i. Test d'indépendance

Il existe des méthodes pour vérifier l'existence ou non des liens entre phénomènes tel que le test du χ^2 et l'ANOVA (Analysis of variance).
On émet l'hypothèse d'indépendance entre 2 variables/caractères (quantitatifs ou qualitatifs) et donc que tout lien observé est dû aux variations aléatoires. Le tableau d'analyse est construit, sur la base de cette hypothèse, en remplaçant les fréquences (ou probabilités) réelles par des fréquences théoriques telles que:

$$n_{ij}^{*} = \frac{n_{i.}n_{.j}}{n} \quad avec \quad i = \overline{1,p} \quad et \quad j = \overline{1,q} \qquad où \quad n = \sum_i \sum_j n_{ij}$$

On calcule la statistique:

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} \quad \mapsto \chi^2_{(p-1)(q-1)\ ddl} \text{ et}$$

si $\chi^2 < \chi^2_{tabulé} \quad \Rightarrow$ H_0 : indépendance.

Ex.16 Pour savoir si le bac influe sur la réussite d'un étudiant en 1ère année universitaire, on construit le tableau suivant des résultats obtenus par 280 étudiants.

Bac / Résultat	Mathématiques	Biologie	Technique	Total
réussite	40	60	50	150
échec	20	40	70	130
Total	60	100	120	280

1. Montrer que le tableau des effectifs théoriques tel que:

$$n_{ij}^{*} = \frac{n_{i.}n_{.j}}{n} = \frac{\sum_j n_{ij} \sum_i n_{ij}}{\sum_i \sum_j nij} \qquad \text{est:}$$

Résultat / Bac	Mathématiques	Biologie	Technique	Total
réussite	32	54	64	150
échec	28	46	56	130
Total	60	100	120	280

2. Montrer que

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} = 12.30 > \chi^2_{1X\,2=2} = 9.21 \; au \; seuil \; \varepsilon = 1\%$$

et décider de l'hypothèse d'indépendance entre le résultat à l'Université et le bac.

ii. Test d'adéquation

Soit un n-échantillon *(X_1, X_2, ..., Xn)* de la v.a. X et F une fonction de répartition donnée. On veut tester l'hypothèse Ho: la fonction de répartition de X est F. Les effectifs théoriques sont calculés en combinant effectif total (n) et probabilités lues sur la table de la loi théorique (P_i): $n_i^* = np_i$.

On calcule:

$$\chi^2 = \sum_i^k \frac{\left(n_i - n_i^*\right)^2}{n_i^*} \quad \text{et si} \quad \chi^2 < \chi^2_{(k-1)\,ddl\;\;tabule} \rightarrow \quad H_0 \; :$$

indépendance.

Note

- k est le nombre de classes
- l'effectif de chaque classe doit être > 5, sinon on regroupe 2 (ou plus) classes consécutives.

Ex.17 Pour apprécier le trafic sur le pont, un observateur compte, pendant une heure, le nombre de voitures qui passent devant lui .Il a obtenu :

Nbre de voitures / mn	0	1	2	3	4	5	6	7
Fréquences observées	3	5	12	13	11	10	4	2

Pour pouvoir assimiler le trafic à une loi Poisson de paramètre λ=4, on donne les probabilités $P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$

X	0	1	2	3	4	5	6	7
P(X=x)	0.02	0.08	0.15	0.20	0.20	0.16	0.11	0.06

1. Montrer que le tableau des fréquences théoriques tel que:

$$n_x^* = pn = P(X = x)\sum n_x \quad \text{est:}$$

X	0	1	2	3	4	5	6	7
n*	1	5	9	12	12	10	7	4

2. Montrer que

$$\chi^2 = \sum \frac{(n-n^*)^2}{n^*} = \frac{(3+5)-(1+5)^2}{1+5} + \frac{(12-9)^2}{9} + \frac{(13-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(10-10)^2}{10} + \frac{((4+2)-(7+4))^2}{7+4} = 4.1 < 5.39 = \chi^2(\varepsilon = 2.5\%)$$

et décider de l'ajustement à la loi Poisson

3.2.2. Test de Student

Il s'agit de tester la significance de l'écart entre 2 moyennes :

$$t = \frac{\overline{Z}}{s/\sqrt{n}} = \frac{\overline{X-Y}}{s/\sqrt{n}} = \frac{\overline{X}-\overline{Y}}{s/\sqrt{n}} \mapsto t_{(n-1)\ ddl}$$

Ex.18 Le tableau suivant donne les taux d'échec dans différentes écoles selon le sexe:

Ecole	E1	E2	E3	E4	E5	E6
Garçons	15	18	14	12	15	16
Filles	17	19	13	15	15	20

Les garçons semblent avoir un taux d'échec plus faible; mais la différence est-elle significative (e.g. au seuil 1%) ou relève-t-elle simplement d'évènements aléatoires

Soit Z (différence entre filles et garçons) une v.a. normale $Z \mapsto N(\overline{Z}, s)$:

$$\overline{Z} = \frac{\sum Z}{n} = \frac{9}{6} = 1.5 \ et \ s = \sqrt{\frac{\sum (Z-\overline{Z})^2}{n-1}} = 2.55$$

On teste l'hypothèse : H_0: $E(Z) = \frac{1}{N}\sum Z = 0$

$$t = \frac{\overline{Z} - E(Z)}{SE(\overline{Z})} = \frac{\overline{Z} - E(Z)}{s/\sqrt{n}} = \frac{\overline{Z}}{s/\sqrt{n}} = \frac{1.5}{2.55/\sqrt{6}} = 1.44 < 4.032 = t_{1\%}(5)$$

$\rightarrow$ Ho: donc la différence n'est pas significative.

3.2.3. Test de Fisher

Soit un n-échantillon d'une v.a. X :

	1	2	...	j	...	q
1						
2						
|						
i				X_{ij}		
|						
p						

$X_{ij} \rightarrow N(m_i, \sigma)$ tel que $n = pq$

Le test vérifie l'hypothèse: $H_o : m_1 = m_2 = \text{----} = m_p$

Ex.19 Vingt observations indépendantes sur le rendement du riz à l'hectare sur 4 parcelles selon le type d'engrais utilisé donnent :

Parcelle	S1	S2	S3	S4
1. Sans engrais	7	7	6	4
2. E_2	10	10	9	7
3. $E_3 + E_2$	6	7	5	3
4. E_4	8	6	8	7
5. $E_5 + E_4$	9	7	8	9

1. Montrer que le rendement moyen est $\overline{X} = 7.15$

2. Calculer les quantités:

$$S_1^2 = \frac{1}{p-1}\sum\sum(X_{ij} - \overline{X})^2 = \frac{1}{p-1}\sum q(\overline{X}_i - \overline{X})^2 \quad \text{et}$$

$$S_2^2 = \frac{1}{n-p}\sum\sum(X_{ij} - \overline{X_i})^2 \quad \text{où } p = 5,\ n = 20,\ q = 4$$

3. Montrer que $F = \frac{S_1^2}{S_2^2} = 5.43 > 4.89 = F_{1\%}(4,15)$

et dire si le type d'engrais a une quelconque influence sur le rendement

3.2.4. Méthode de Bayes

On associe aux hypothèses Ho et H_1 des probabilités a priori p_0 et p_1 (=1-p_0). Chaque décision a aussi un coût:

Décision / Hypothèse	D_0	D_1	Probabilité a priori
H_0	C_{00}	C_{01}	p_0
H_1	C_{10}	C_{11}	p_1

Au vu d'un échantillon (X_1 , X_2 , ---- , Xn) on calcule les probabilités a posteriori π_0 et π_1 (= 1-π_0) à associer aux hypothèses / $\pi_0 = \frac{p_0 L_0}{p_0 L_0 + p_1 L_1}$

où L est la fonction de vraisemblance: $\theta \rightarrow L(X_1, ---, Xn ; \theta)$ notée L_o si $\theta \in \Theta_0$ et L_1 si $\theta \in \Theta_1$

La règle de Bayes est de prendre la décision dont l'espérance de coût est la plus faible:

$$\min[E(C(D_0)), E(C(D_1))] \; avec \begin{cases} E(C(D_0)) = C_{00}\pi_0 + C_{10}\pi_1 \\ E(C(D_1)) = C_{01}\pi_0 + C_{11}\pi_1 \end{cases}$$

Ex.20 Une grande entreprise envisage de lancer une campagne publicitaire dans plusieurs pays pour promouvoir ses ventes. On mène la campagne dans 5 pays pilote et les augmentations de ventes enregistrées ont été: 100, 150, 0, 100, 50 unités monétaires.

L'augmentation $X_i \left(i = \overline{1, 5}\right)$ est une v.a. $\rightarrow$ N (m ,σ) avec

σ =200 et m = 0 si la campagne est inefficace

m = 200 si elle est efficace .

Le coût moyen de la campagne est estimé à 50 unités monétaires par pays. Faut-il lancer l'opération si la firme pense que la publicité a une chance sur deux de réussir ?

Dans cet exemple, nous avons:

- ✓ 2 hypothèses : $\begin{cases} H_0 : \textit{campagne inefficace} ; m = 0 \\ H_1 \ \textit{campagne efficace} ; m = 200 \end{cases}$
- ✓ 2 décisions possibles : $\begin{cases} D_0 : \textit{ne rien faire} \\ D_1 : \textit{entreprendre la campagne} \end{cases}$
- ✓ les coûts associés : $\begin{cases} C_{00} = C_{10} = 0 \\ C_{01} = 50 - 0 = 50 \\ C_{11} = 50 - 200 = -150 \end{cases}$

Soit finalement:

	D_0	D_1	Probabilité a priori
H_0	0	50	½ = p_0
H_1	0	- 150	½ = p_1

On entreprendra la campagne si:

$$E(C(D_1)) \langle E(C(D_0)) \Leftrightarrow 50\pi_0 - 150\pi_1 \langle 0 \Leftrightarrow \pi_0 \langle 3\pi_1 \Leftrightarrow p_0 L_0 \langle p_1 L_1$$

$$L(X_{1'} - - -, X_{n'} m) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum \left(\frac{X - m}{\sigma} \right)^2 \right)$$

$$L_0 = \frac{1}{200^5 (2\pi)^{5/2}} \exp\left[-\frac{1}{2} \left(\left(\frac{1}{2} \right)^2 + \left(\frac{15}{20} \right)^2 + 0^2 + \left(\frac{1}{2} \right)^2 + \left(\frac{1}{4} \right)^2 \right) \right]$$

$$L_1 = \frac{1}{200^5 (2\pi)^{5/2}} \exp\left[-\frac{1}{2} \left(\left(-\frac{1}{2} \right)^2 + \left(\frac{1}{4} \right)^2 + 0^2 + \left(-\frac{1}{2} \right)^2 + \left(\frac{15}{20} \right)^2 \right) \right]$$

$P_0 L_0 = \frac{1}{2} L_0 < \frac{3}{2} L_0$ d'où la décision de lancer la campagne.

3.2.5. Méthode Neyman-Pearson

☞ On appelle risque de 1ère espèce la probabilité de rejeter à tort l'hypothèse nulle (H_o): $\alpha = P(D_1 \setminus H_0) = P(w \setminus \theta \in \Theta_0)$

☞ On appelle risque de 2nde espèce la probabilité d'accepter à tort l'hypothèse nulle (H_o): $\beta = P(D_0 \setminus H_1) = P(\overline{w} \setminus \theta \in \Theta_1)$

Puisque rejeter à tort l'hypothèse nulle est considérée comme étant l'erreur la plus grave, on fixe un seuil maximum α_0 au risque de 1ère espèce et on cherche un test qui minimise le risque de 2ème espèce: $\min \beta \Leftrightarrow \max(1-\beta) = \max \eta = P(D_1 \setminus H_1) = P(w \setminus \theta \in \Theta_1)$

où η est appelé **puissance du test** i.e. la probabilité de rejeter l'hypothèse nulle avec raison. Ainsi la règle de décision de Neyman et Pearson consiste à déterminer la région critique w pour laquelle la puissance η est max sous la contrainte $\alpha \langle \alpha_0$.

Ex.21 Dans l'exemple précédent, quelle serait la décision si l'on souhaitait avant tout vérifier le risque de ne pas entreprendre une campagne qui serait efficace ?

On teste $H_0 : m = m_0 = 200$ contre $H_1 : m = m_1 = 0$ au seuil de confiance α_0

La v.a. $X \rightarrow N(m, \sigma)$ et l'échantillon $(X_1, \text{------}, X_n)$ a pour vraisemblance:

$$L(.) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (X-m)^2 \right) \text{ avec:}$$

$$L_{H_0}(.) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (X-m_0)^2 \right)$$

$$L_{H_1}(.) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (X-m_1)^2 \right)$$

$$w = \left\{ (X_1,\ldots,X_n) \Big/ \frac{L_0}{L_1} \le k \right\} \text{ où k est une constante}$$

$$\frac{L_0}{L_1} = \exp\left(-\frac{1}{2\sigma^2} \left(\sum (X-m_0)^2 - \sum (X-m_1)^2 \right) \right) \le k \text{ qui donne (par les log.):}$$

$$\sum \left((X-m_0)^2 - (X-m_1)^2 \right) \le -2\sigma^2 \ln k = k_1 \Leftrightarrow \sum \left(m_0^2 - m_1^2 - 2(m_0 - m_1)X \right) \le k_1$$

$$\Leftrightarrow n(m_0^2 - m_1^2) - 2(m_0 - m_1)n\overline{X} \le k_1 \Leftrightarrow n(m_0 - m_1)(m_0 + m_1 - 2\overline{X}) \le k_1$$

$X \to N(m, \sigma) \Rightarrow \overline{X} \to N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ d'où: $Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \to N(0,1)$

$$\alpha = P\left(Z \ge \frac{a - m_0}{\sigma/\sqrt{n}}\right) \Leftrightarrow 1 - \alpha = P\left(Z < \frac{a - m_0}{\sigma/\sqrt{n}}\right) = P(Z < \lambda)$$

Pour $\alpha = 5\% \to P(.) \to \lambda = 1.645 \Rightarrow a = m_0 + \frac{\sigma}{\sqrt{n}}\lambda = 347$

Or $\overline{X} = \frac{1}{5}\sum X = 80 < 37 = a$

$\to$ on accepte H_o: la firme décide de lancer la campagne.