Chapitre 5 : Statistiques -02-12-10-
Terminale ES 1, 2010-2011, Y. Angeli

1. Rappels : statistiques à une variable

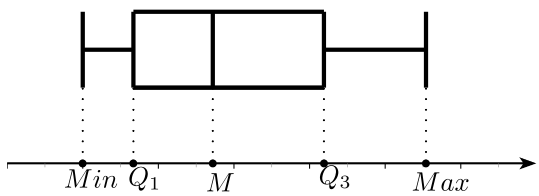
• La *population* désigne l'ensemble étudié, un *individu* est un éléméent de cet ensemble. L'*effectif total* N est le nombre d'individus de la population.
• Le *caractère* ou la *variable statistique* x désignent la propriété étudiée.
• Le terme x_i de la série statisitique est la valeur prise par le caractère x pour l'individu numéro i de la population.
• Les *modalités* $a_1, ..., a_k$ sont différentes valeurs prises par le caractère x. L'*effectif* n_j d'une modalité a_j représente le nombre d'individus de la population dont le caractère vaut a_j. (c'est-à-dire le nombre de fois que a_j apparaît parmi les x_i). En particulier : $N = n_1 + ... + n_k$
• La *moyenne* de la série statistique est le nombre

$$\bar{x} = \frac{x_1 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{n_1 a_1 + ... + n_k a_k}{N} = \frac{1}{N}\sum_{j=1}^{k} n_j a_j$$

• La *variance* V de la série statistique est le nombre :

$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2\right) - \bar{x} = \frac{1}{N}\sum_{i=j}^{k} n_j(a_j - \bar{x})^2 = \frac{1}{N}\left(\sum_{j=1}^{k} n_j a_j^2\right) - \bar{x}^2$$

• L'*écart type* σ est la racine carrée de la variance : $\sigma = \sqrt{V}$. (L'écart-type et la variance mesurent la *dispersion* de la série autour de la moyenne : un grand σ indique une série dispersée, un petit σ une série plus compacte.)
• On suppose la série des x_i rangée par ordre croissant. La *médiane* M est la valeur "centrale" : $M = \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1})$ si N est pair, $M = x_{\frac{N+1}{2}}$ si N est impair.
• Le *premier quartile* Q_1 est la médiane de la série des x_i où $i > \frac{N+1}{2}$.
• Le *troisième quartile* Q_3 est la médiane de la série des x_i où $i < \frac{N+1}{2}$.
• L'*intervalle interquartile* est $]Q_1, Q_3[$. L'écart interquartile est $Q_3 - Q_1$.
• Le *diagramme en boîte* résume les informations sur la distribution de la série :



2. Exemple : Structure de l'emploi du temps par genre

Le tableau suivant a été publié par l'organisme Eurostat le 8 mars 2006 (journée de la femme). Il présente la structure des emplois du temps des femmes et des hommes agés de 20 à 74 ans, en heures par jours, selon un sondage réalisé en 2005.

	Emploi,études		**Travail domestique**		**Travail total**		
	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Différence
Belgique	2.1	3.5	4.5	2.6	6.6	6.1	0.5
Allemagne	2.1	3.6	4.2	2.4	6.3	6.0	
Espagne	2.4	4.7	4.9	1.6	7.3	6.3	
France	2.5	4.1	4.5	2.4	7.0	6.5	
Italie	2.1	4.4	5.3	1.6	7.4	6.0	
Lettonie	3.7	5.2	3.9	1.8	7.6	7.0	
Lituanie	3.7	4.9	4.5	2.2	8.2	7.1	
Hongrie	2.5	3.8	5.0	2.7	7.5	6.5	
Pologne	2.5	4.3	4.8	2.4	7.3	6.7	
Slovénie	3.0	4.1	5.0	2.7	8.0	6.8	
Suède	3.2	4.4	3.7	2.5	6.9	6.9	
Royaume-Uni	2.6	4.3	4.3	2.3	6.9	6.6	
Norvège	2.9	4.3	3.8	2.4	5.7	6.7	
Moyenne	2.7			2.3	7.2		
Variance	0.28			0.13			
Écart-type	0.53			0.36			
Minimum	2.1			1.6			
Quartile 1	2.25			2			
Médiane	2.5			2.4		6.6	
Quartile 3	3.1			2.55			
Maximum	3.7			2.7			

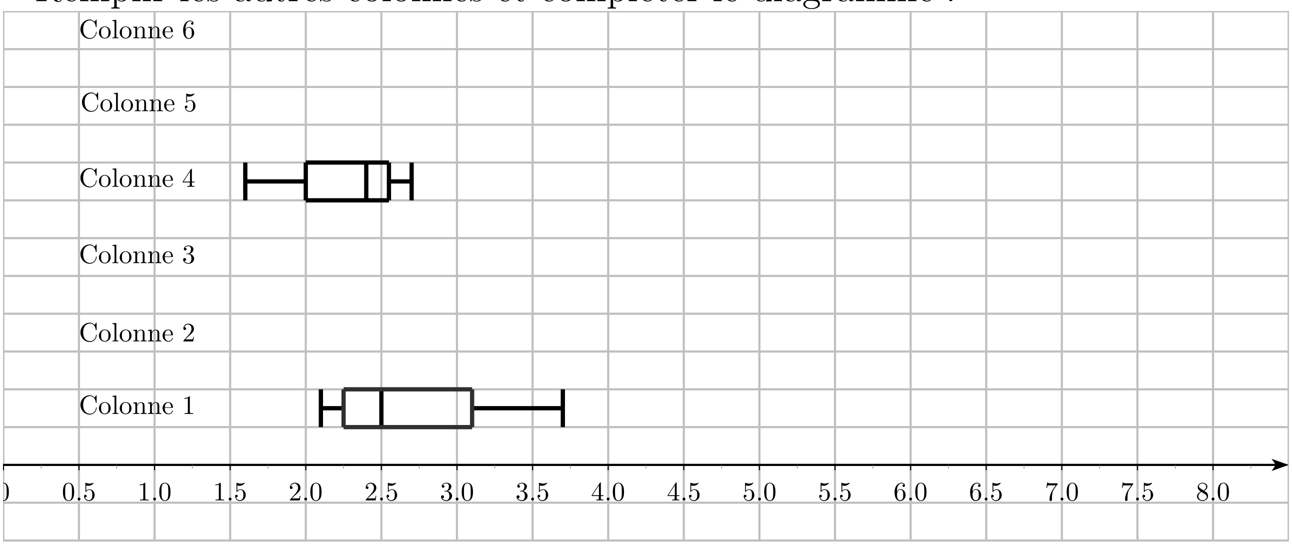
• **Travail payé et études des hommes** : Ranger les termes de la série de la seconde colonne par ordre croissant :

$$\leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq \quad \leq$$

Moyenne= ....

Variance=...

• Remplir les autres colonnes et compléter le diagramme :



3. Statistiques à deux variables

On considère, pour une population donnée, deux caractères dont les valeurs sont $x_1, ..., x_N$ et $y_1, ..., y_N$.

Définition. • Dans un repère orthogonal, l'ensemble des points $M_i(x_i, y_i)$ est le *nuage de points* associé à la série statistique à deux variables.
• Le *point moyen* associé à cette série est le point $\bar{M} = (\bar{x}, \bar{y})$ où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des séries x et y.
• La *covariance* des séries x et y est le nombre

$$C_{x,y} = \text{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right) - \bar{x}\bar{y}$$

Une covariance positive indique une tendance croissante du nuage, une covariance négarive une tendance décroissante.

4. Ajustement affine

Lorsque le nuage présente une forme allongée, il est naturel de rechercher une droite aussi proche que possible de l'ensemble des points du nuage. Réaliser un *ajustement affine* d'une série statistique à deux variables, c'est trouver une droite qui approche au mieux l'ensemble des points du nuage.

Définition. On considère l'ajustement affine d'une série statistique par une droite $\mathcal{D}$ d'équation $y = ax + b$. Soit $M_i'(x_i, ax_i + b)$ le point de $\mathcal{D}$ d'abscisse x_i. La différence des ordonnées de M_i et M_i' est $r_i = y_i - (ax_i + b)$. Elle permet de mesurer la différence entre la réalité et le modèle pour le point i de la série. Le nombre r_i est le ième *résidu.*

Théorème. Il existe une droite, appelée *droite des moindres carrés* ou *droite de regression*, telle que la somme des carrés des résidus

$$\sum_{i=1}^{N} A_i P_i = \sum_{i=1}^{N}(y_i - ax_i - b)^2$$

soit minimale. Elle passe par le point moyen du nuage. Son coefficient directeur a et son ordonnée à l'origine b sont

$$a = \frac{C_{x,y}}{V_x}, \quad b = \bar{y} - a\bar{x}.$$

5. Exemple : inscrits au pôle emploi depuis septembre 2008

D'après l'insee, le nombre d'inscrits au pôle emploi, en milliers, depuis le mois de septembre 2008 est :

Mois	09-08	10-08	11-08	12-08	01-09	02-09	03-09	04-09	05-09	06-09	07-09	08-09
Mois n$^o x_i$	0	1	2	3	4	5	6	7	8	9	10	11
Inscrits y_i	3299	3348	3387	3438	3524	3605	3688	3786	3843	3851	3888	3923
Résidu r_i												

Les coordonnées du point moyen sont $(\bar{x}, \bar{y}) =$ ....

La covariance de la série à deux variables est $C(x, y) =$ ....

La droite d'ajustement des moindres carrés a pour équation $y = ax + b$ avec

$$a = \frac{C(x,y)}{V(x)} = \qquad\qquad b = \bar{y} - a\bar{x} =$$

La somme des carrés des résidus r_i est $\sum_{i=0}^{11} r_i^2 =$

Combien de milliers d'inscrits peut-on prévoir pour septembre 2009 ?
Pour novembre 2009 ?