Probabilités et Statistiques

Emmanuel PAUL

Chapitre 1 : Statistique descriptive

1 Objectifs des statistiques.

Il s'agit d'étudier un ou plusieurs caractères (appelés aussi variables statistiques) d'une population.

Exemples. On considère la population $E = \{e_1, e_2, \cdots e_n\}$ des n étudiants de première année du département d'informatique de l'IUT de Toulouse. On peut étudier plusieurs caractères :

- C : $E \longrightarrow \mathcal{C}$ (un ensemble de couleurs) qui à chaque étudiant e_α associe $C(e_\alpha)$ = la couleur de ses yeux. C'est un caractère **qualitatif**.
- N : $E \longrightarrow \mathbb{N}$ (ensemble des entiers naturels) qui à chaque étudiant e_α associe sa note $N(e_\alpha)$ obtenue au contrôle d'algèbre linéaire du premier semestre. C'est un caractère **quantitatif discret** (les valeurs sont isolées) **à une dimension** (une valeur par individu).
- T : $E \longrightarrow \mathbb{R}$ (ensemble des nombres réels) qui à chaque étudiant e_α associe sa taille $T(e_\alpha)$. C'est un caractère **quantitatif continu** (le caractère peut à priori prendre toute valeur dans un intervalle de $\mathbb{R}$ donné) à une dimension.
- (T,P) : $E \longrightarrow \mathbb{R}^2$ (ensemble des couples de nombres réels) qui à chaque étudiant e_α associe sa taille et son poids $(T(e_\alpha), P(e_\alpha))$. C'est un caractère **quantitatif continu à deux dimensions.**

Exercice 1. *Les variables statistiques suivantes sont-elles discrètes ou continues ?*
- le nombre d'actions vendues chaque jour à la bourse de Paris ;
- les températures et pressions enregistrées chaque heure dans une station météo ;
- la durée de vie d'un lot d'ampoules électriques fabriquées par une usine ;
- le revenu mensuel de la population ouvrière en France.

La **statistique descriptive** met en ordre les données brutes d'un paramètre, notamment par des représentations graphiques, et fournit des indicateurs de position (valeur moyenne etc...), de dispersion autour de la valeur moyenne (écart-type...), ou d'indépendance (dans le cas de plusieurs caractères).

La **statistique mathématique (ou inférentielle)** fait des estimations sur un caractère uniquement à partir d'une connaissance partielle du caractère sur un échantillon. Elle nécessite l'utilisation de la théorie des probabilités. Elle permet des estimations (sondages etc...), et devient une aide à la décision (tests statistiques).

2 Statistique descriptive à une dimension.

Soit $E = \{e_1, \cdots e_n\}$ une population et $X : E \to \mathbb{N}$ une distribution statistique quantitative discrète. Soit $x_1, x_2, \cdots x_k$ les valeurs prises par cette distribution rangées par ordre croissant.

> ***Exercice 2.*** *Donner les valeurs de n et k pour l'exemple précédent de la distribution N des notes (supposées entières ou demi-entières).*

Pour représenter un caractère discret d'une population, on regroupe par classe C_i tous les individus dont le caractère prend la même valeur x_i, $i = 1, \cdots k$. On note n_i l'**effectif** de cette classe. L'**effectif total** est

$$n = n_1 + n_2 + \cdots n_k = \sum_{i=1}^{k} n_i.$$

La donnée des classes et de leurs effectifs est la **distribution statistique** associée à X. L'**effectif cumulé** des i premières classes est :

$$N_i = n_1 + n_2 + \cdots n_i = \sum_{k=1}^{i} n_k.$$

La proportion de la population prenant la valeur x_i est donnée par la **fréquence** :

$$f_i = \frac{n_i}{n}.$$

La proportion de la population prenant une valeur inférieure ou égale à x_i est donnée par la **fréquence cumulée** des i premières classes :

$$F_i = f_1 + f_2 + \cdots f_i = \frac{N_i}{n}.$$

On la prolonge pour tout x réel par la fonction en escalier : $F(x) = \sum_{x_j \leq x} f_j$ (appelée **fonction de répartition**).
La proportion de la population dont le caractère prend une valeur dans $]a, b]$ (attention aux bornes !) est donnée par

$$F(b) - F(a).$$

Exemple 1. *La population étudiée est un ensemble de 30 familles. Le caractère discret étudié X est le nombre d'enfants. Les classes et leurs effectifs sont donnés par le tableau suivant :*

classes	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
valeurs	0	1	2	3	4	5	6	7
effectifs	5	7	8	4	2	2	1	1
effectifs cumulés								
fréquences								
fréquences cumulées								

Exercice 3.
1- Complétez ce tableau.
2- Quel est le pourcentage de familles admettant deux enfants ?
3- Quel est le pourcentage de familles admettant au plus un enfant ?
4- Quel est le pourcentage de familles admettant 2 à 5 enfants ?

Les **histogrammes** des effectifs, puis des effectifs cumulés (obtenus sous Maple : voir TP1) sont représentés ci-dessous. On a utilisé la convention suivante : la largeur des colonnes contenant chaque valeur est identique pour chaque classe, et la hauteur égale à l'effectif. L'histogramme des fréquences (ou des fréquences cumulées) est identique mais avec une graduation différente de l'axe vertical (les ordonnées sont ici divisées par $n = 30$).

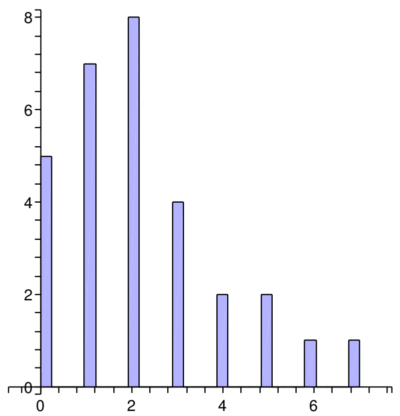

FIG. 1 – Histogramme des effectifs

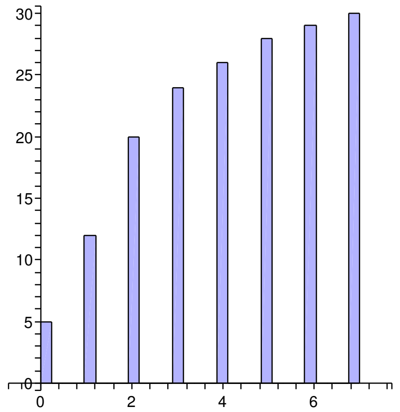

FIG. 2 – Histogramme des effectifs cumulés

Pour un caractère continu, on partitionne l'ensemble de ses valeurs en une collection d'intervalles I_1, I_2, ... I_k de la forme $[c_{i-1}, c_i[$. La classe C_i regroupe les individus dont le caractère prend sa valeur dans l'intervalle $[c_{i-1}, c_i[$. On note à nouveau n_i son effectif. Les notions d'effectifs cumulés, fréquences et fréquences cumulées sont identiques.

La fonction de répartition F est maintenant construite comme suit : on place la valeur 0 au dessus de l'extrémité gauche de I_1, F_1 au dessus de son extrémité droite, puis chaque valeur F_i au-dessus de l'extrémité droite de l'intervalle I_i, en terminant par la valeur 1 à l'extrémité droite du dernier intervalle. On joint les points obtenus.
- La valeur $F(a)$ représente le pourcentage de la population dont le caractère prend une valeur inférieure ou égale à a.
- La valeur 1-$F(a)$ représente le pourcentage de la population dont le caractère prend une valeur strictement supérieure à a.
- La différence des valeurs $F(b) - F(a)$ représente le pourcentage de la population dont le caractère prend une valeur dans l'intervalle $]a, b]$.

Exemple 2. *La distribution Y des tailles d'une population de 100 collégiens est donnée par le tableau :*

classes	C_1	C_2	C_3	C_4
valeurs en cm.	[150,155[	[155,160[	[160,165[	[165,170[
effectifs	30	25	23	22
effectifs cumulés				
fréquences				
fréquences cumulées				

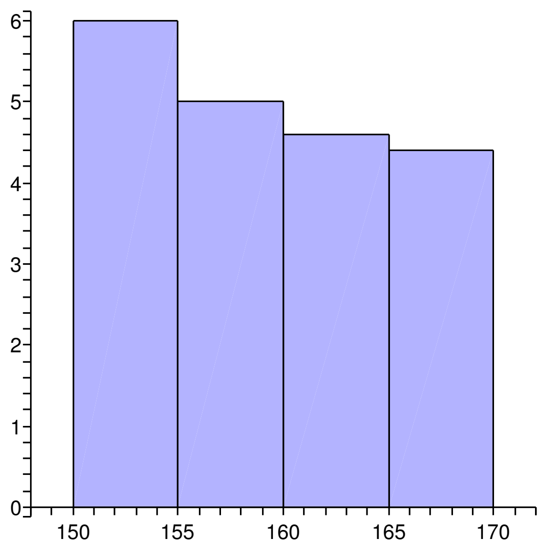

FIG. 3 – Histogramme des effectifs

Remarque : L'aire de chaque rectangle est égale à l'effectif de la classe (d'où la graduation verticale).

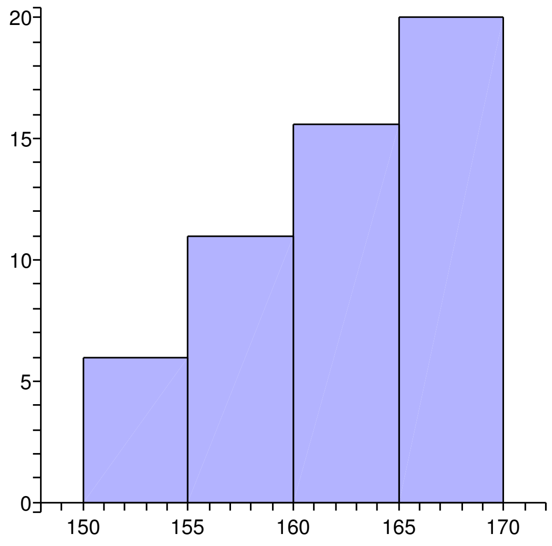

FIG. 4 – Histogramme des effectifs cumulés

Exercice 4.
1- Complétez le tableau de la distribution de Y.
2- Déterminer graphiquement le pourcentage de la population dont la taille est inférieure à 163 cm, à l'aide de l'histogramme des effectifs cumulés ci-dessous (on tracera d'abord le graphe de la fonction de répartition sur l'histogramme des effectifs cumulés.)
3- Déterminer ce pourcentage par le calcul : on cherchera d'abord l'équation $y = ax + b$ du segment de droite concerné, puis l'ordonnée associée à l'abscisse fournie, et enfin l'effectif puis la fréquence correspondante.

3 Paramètres d'une distribution statistique.

3.1 Paramètres de position.

Moyenne. On considère la population $E = \{e_\alpha, \alpha = 1, \cdots n\}$ et le paramètre quantitatif X. Soit $x_1, \cdots x_k$ les valeurs prises par X à valeurs discrètes, et $n_1, \cdots n_k$ la distribution d'effectifs correspondante. Si le caractère est continu x_i désigne le milieu de l'intervalle des valeurs $[c_{i-1}, c_i[$: $x_i = (c_{i-1} + c_i)/2$. Le principal paramètre de position d'une distribution statistique est sa *moyenne* :

$$m = \frac{1}{n} \sum_{\alpha=1}^{n} X(e_\alpha) = \frac{1}{n} \sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i.$$

Remarque : la seconde expression est simplement obtenue en regroupant les termes $X(e_\alpha)$ de la première somme prenant même valeur x_i. Elle est plus rapide à calculer.

Exercice 5. *Calculer les moyennes m_X et m_Y des distributions statistiques X et Y du paragraphe précédent.*

Remarque. Etant données deux distributions statistiques X et Y sur un même ensemble on peut considérer la distribution somme $X+Y$: on somme les valeurs associées à chaque individu. On peut aussi multiplier une distribution par un nombre réel λ. Notons m_X et m_Y les moyennes de chaque distribution. On vérifie facilement les propriétés de linéarité :

$$m_{X+Y} = m_X + m_Y, \quad \text{et } m_{\lambda X} = \lambda m_X.$$

Médiane. Le second paramètre de position fréquemment utilisé est la *médiane* : c'est la valeur notée $x_{1/2}$ qui partage la population en deux parties de même effectif : les individus dont le caractère est inférieur à cette valeur et ceux pour lesquels le caractère est supérieur à cette valeur. C'est donc la valeur pour laquelle la fonction de répartition vaut 1/2. On la calcule de la manière suivante :

- *Cas discret avec n petit* : on range les valeurs par ordre croissant : $x_1 < \cdots < x_k$ en répétant n_i fois chaque valeur x_i. La médiane $x_{1/2}$ est la valeur séparant cette suite en deux parties d'effectif égal. Dans le cas où n est pair, $x_{1/2}$ tombe entre deux valeurs distinctes : on prend alors pour $x_{1/2}$ le milieu de ces valeurs.

Exercice 6.
- Calculer la médiane de la distribution de la distribution statistique X du paragraphe précédent.
- Que serait devenue cette médiane si la famille de 7 enfants en avait eu 18 ?
- Cela aurait-il changé la moyenne ?
On conclut donc que la médiane, contrairement à la moyenne est insensible aux valeurs "exceptionnelles" qui peuvent parfois provenir d'une erreur de relevé ou d'expérience.

- *Cas continu* : on trace le graphe de la fonction de répartition F. On obtient la médiane $x_{1/2}$ en résolvant l'équation $F(x) = 1/2$. Pour cela, on cherche la **classe médiane**, c'est-à-dire l'intervalle de valeurs $[a, b]$ contenant la médiane : les ordonnées $F(a)$ et $F(b)$ pour la fonction de répartition entourent la valeur 1/2 : $F(a) \leq 1/2$ et $F(b) > 1/2$. La pente p du segment de droite correspondant est donnée par :

$$p = \frac{F(b) - F(a)}{b - a} = \frac{1/2 - F(a)}{x_{1/2} - a}$$

d'où la formule :

$$x_{1/2} = a \ + \ (b - a) \times \frac{1/2 - F(a)}{F(b) - F(a)}.$$

Exercice 7. *Déterminer la classe médiane de la distribution Y, puis obtenir sa médiane, d'abord sur le graphique du paragraphe 2, puis par le calcul.*

Autres paramètres de position :

- on définit de même que la médiane les **quartiles** $x_{1/4}$ et $x_{3/4}$ comme étant les valeurs pour lesquelles F vaut 1/4 ou 3/4. Ils se calculent de la même manière que la médiane : même formule en remplaçant 1/2 par 1/4 (ou 3/4) et en choisissant l'intervalle $[a, b]$ de sorte que les ordonnées par F encadrent 1/4 (ou 3/4). On définirait de même les **déciles**, **centiles**...

- le (ou les) **modes** (ou classes modales) : il s'agit d'une classe d'effectif maximal. Il peut y en avoir plusieurs.

Exercice 8.
- Quels sont les quartiles $x_{1/4}$ et $x_{3/4}$ de la distribution Y ?
- Quelle est la classe modale de cette distribution ?

3.2 Paramètres de dispersion.

Ils mesurent l'éloignement entre les valeurs x_i et la valeur moyenne m, et se calculent donc à partir des écarts $\mid x_i - m \mid$. On moyennise ensuite ces écarts de deux manières différentes :
- l'écart-moyen (peu utilisé) : c'est la moyenne des écarts : $\frac{1}{n}\sum_{i=1}^{k} n_i \mid x_i - m \mid$.
- **l'écart-type** (très utilisé) : c'est la moyenne quadratique des écarts :

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - m)^2}.$$

Remarques :
- le carré de l'écart-type σ^2 est appelé "**variance**" de X.
- On peut aussi calculer σ^2 en utilisant la formule de Koenigs (démontrez-là !) :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - m)^2 = \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i^2\right) - m^2$$

Autres paramètres de dispersion (moins utilisés) :
- l'étendue : $x_{max} - x_{min}$;
- l'écart inter-quartile $x_{3/4} - x_{1/4}$.

Exercice 9. *Calculer les écart-types des distributions statistiques X et Y du paragraphe précédent (voir le mode d'emploi de votre calculatrice).*

4 Statistique descriptive à deux dimensions.

4.1 Distribution statistique à deux dimensions, distributions marginales

Soit $(X, Y) : E \to \mathbb{R}^2$ un caractère à deux dimensions sur une population de n éléments. Soit n_{ij}, $i = 1, \cdots k$, $j = 1, \cdots l$, l'effectif de la valeur (x_i, y_j) (ou dans le cas continu, d'une classe déterminée par le produit de deux intervalles de valeurs). On a : $n = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$ (noté en abrégé : $\sum_{i,j} n_{ij}$). L'application qui, à chaque valeur (ou classe de valeurs) associe l'effectif correspondant est la distribution statistique de (X, Y). On peut aussi définir la distribution des fréquences par $f_{ij} = n_{ij}/n$.

Exemple 3. *On reprend la population des collégiens de l'exemple 2, mais on mesure maintenant les deux caractères (T, P) = (Taille, Poids). Le tableau des effectifs n_{ij} est le suivant :*

P \ T	[150,155[	[155,160[	[160,165[	[165,170[	dist. marg. de P
[40,45[	20	2	0	0	
[45,50[	9	18	5	1	
[50,55[	1	4	12	7	
[55,60[	0	1	6	14	
dist. marg. de T					

Les **distributions marginales** sont les deux distributions à une dimension obtenues en sommant les effectifs sur un des indices :

$$n_{i\cdot} = \sum_{j=1}^{l} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}.$$

De même les fréquences marginales sont :

$$f_{i\cdot} = n_{i\cdot}/n, \qquad f_{\cdot j} = n_{\cdot j}/n.$$

Exercice 10.
- Compléter le tableau de l'exemple 3 par les distributions marginales.
- Comment obtiendrait-on le tableau des fréquences et fréquences marginales ?

Les **fréquences conditionnelles** de Y sachant que $X = x_i$ sont les fréquences obtenues en ne regardant que la i-ème ligne du tableau. Pour tout j (à i fixé) :

$$f_j \mid_{X=x_i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i,\cdot}}.$$

Exercice 11.
Quelle est la fréquence d'apparition d'une taille dans l'intervalle $[160, 165[$ *sachant que le poids d'un individu est compris entre 45 et 50 kg ? entre 50 et 55 kg ?*

4.2 Indépendance.

On dit que X et Y sont *indépendantes* lorsque ces fréquences conditionnelles ne dépendent pas de la condition $X = x_i$, c'est-à-dire lorsque le résultat obtenu est indépendant de l'indice i de la ligne choisie. Dans ce cas, on a

$$\frac{f_{ij}}{f_{i\cdot}} = \frac{\sum_i f_{ij}}{\sum_i f_{i\cdot}} = \frac{f_{\cdot j}}{1}$$

et on a donc : **X et Y sont indépendantes si et seulement si pour tout (i,j),**

$$\boxed{f_{ij} = f_{i\cdot} \times f_{\cdot j}}$$

On peut alors retrouver le tableau de la distribution (X, Y) uniquement en effectuant des produits à partir des distributions marginales.

Exercice 12.
- Dans la distribution précédente, les deux variables sont-elles indépendantes ?
- Quel aurait été l'effectif (théorique) de la case $[50, 55[\times[160, 165[$ *si les variables statistiques avaient été indépendantes ?*

4.3 Paramètres d'une distribution à deux dimensions.

On a d'abord les paramètres (moyennes et écart-types) des deux distributions marginales :

- **Le point moyen :** (m_X, m_Y) avec $m_X = \frac{1}{n}\sum_i n_{i\cdot}x_i$ et $m_Y = \frac{1}{n}\sum_j n_{\cdot j}y_j$.
- **Les deux écart-types :** $\sigma_X = \sqrt{\frac{1}{n}\sum_i n_{i\cdot}(x_i - m_X)^2}$, $\sigma_Y = \sqrt{\frac{1}{n}\sum_j n_{\cdot j}(y_j - m_Y)^2}$.

Les autres paramètres servent à mesurer le degré d'indépendance entre les deux caractères. On utilise :

- **La covariance** :

$$\begin{aligned} \mathrm{cov}(X,Y) &= \frac{1}{n}\sum_{i,j} n_{ij}(x_i - m_X)(y_i - m_Y) \\ &= \frac{1}{n}(\sum_{i,j} n_{ij}x_iy_j) - m_X \cdot m_Y. \end{aligned}$$

La deuxième expression est une généralisation à deux indices de la formule de Koenigs.

- **Le coefficient de corrélation linéaire** :

$$r(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Propriétés de la covariance et du coefficient de corrélation linéaire :

- Ces deux paramètres sont symétriques par rapport à X et Y. Ils peuvent être négatifs.

- Ils sont inchangés par translations : $\text{cov}(X+a, Y+b) = cov(X,Y)$ (idem pour r) ;

- Sous l'action d'un changement d'échelle on a :
$\text{cov}(aX, bY) = ab \ \text{cov}(X,Y)$ et $r(aX, bY) = \pm r(X,Y)$ suivant le signe de ab.

- On a : $\mid \text{cov}(X,Y) \mid \leq \sigma_X \cdot \sigma_Y$ et donc $\mid r(X,Y) \mid \leq 1$.

- **Si X et Y sont indépendantes, alors cov(X,Y)=r(X,Y)=0.** En effet

$$\begin{aligned} \text{cov}(X,Y) &= (\sum_{i,j} f_{ij} x_i y_j) - \overline{x} \cdot \overline{y} = (\sum_i \sum_j f_{i\cdot} f_{\cdot j} x_i y_j) - \overline{x} \cdot \overline{y} \\ &= (\sum_i f_{i\cdot} x_i)(\sum_j f_{\cdot j} x_j) - \overline{x} \cdot \overline{y} = \overline{x} \cdot \overline{y} - \overline{x} \cdot \overline{y} = 0. \end{aligned}$$

La réciproque est fausse. Si r est nul, on ne peut pas conclure que X et Y sont indépendantes. En effet r ne mesure que les liaisons linéaires (le long d'une droite) et il peut y en avoir d'autres (liaisons quadratiques le long d'une parabole etc...)

- **Si X et Y sont linéairement liés ($Y = aX + b$) alors, $r(X,Y) = \pm 1$.** En effet,

$$r(X, aX+b) = r(X, aX) = \pm r(X,X) = \pm 1.$$

Exercice 13.
- Dans la distribution de l'exemple 3, calculer la covariance et le coefficient de régression linéaire.
- Les deux variables sont-elles proches d'une relation linéaire ?

4.4 Ajustement linéaire d'une distribution à deux dimensions.

On peut représenter graphiquement une distribution à deux dimensions dont tous les effectifs valent 1 par un nuage de points : pour chaque valeur (x_i, y_j) prise par (X,Y), on place un point. Si la distribution est continue (comme sur l'exemple 3), x_i et y_j sont les milieux des intervalles. Si les effectifs sont quelconques les points sont affectés d'un *poids* égal à cet effectif.

Exemple 4. On considère les deux séries statistiques sur une population de 16 individus :

$$X := [0, 1, 2, 2, 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 9, 10]$$

$$Y := [4, 5, 5, 3, 5, 3, 4, 4, 4, 6, 6, 5, 9, 8, 8, 7\ ].$$

Le couple (X,Y) prend donc les valeurs (0,4), (1,5) etc... Les couples (2,5) et (5,4) sont de poids deux. Le nuage de points obtenu est :

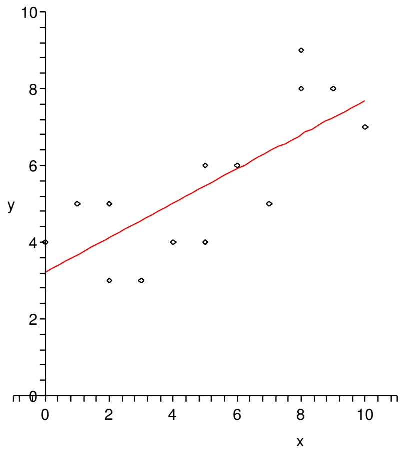

FIG. 5 – Nuage de points représentant (X,Y)

Nous venons de voir que r mesure l'existence de relations linéaires : les points du nuage sont proches d'une droite lorsque $|r|$ est proche de 1. Sur cet exemple, $r \simeq 0,74$. L'alignement n'est pas très bon.

On veut déterminer la droite qui ajuste le mieux le nuage, c'est-à-dire la droite D qui minimise les distances verticales entre les points du nuage et D : c'est la **droite de régression ou d'ajustement linéaire, ou encore droite des moindres carrés** de Y par rapport à X. Elle passe par le point moyen (m_X, m_Y) et son équation est donnée par

$$\boxed{y = ax + b, \text{ avec } a = \frac{cov(X,Y)}{\sigma_X^2} \text{ et } b = m_Y - am_X.}$$

Nous calculerons cette droite au TP1. Elle est dessinée sur le graphique ci-dessus. On calcule de même la droite de régression linéaire de X par rapport à Y en échangeant les rôles de X et Y. Elle minimise les distances horizontales entre les points du nuage et D.

Exercice 14.
- Sur l'exemple 4, calculer l'équation de la droite d'ajustement linéaire de Y par rapport à X.
- Tracer cette droite sur le nuage de points pondérés. Vérifier qu'elle coïncide avec celle indiquée ci-dessus.

Relations non linéaires. Il se peut que X et Y soient approximativement liés par des relations non linéaires. Par exemple : $Y = aX^2 + bX + c$, $Y = b + a\ln(X)$, $Y = be^{aX}$, etc... Le nuage de points s'ajuste alors de manière plus satisfaisante sur les courbes correspondantes. On peut se ramener à la recherche d'une relation linéaire par un changement de variable : $X' = X^2$, $X' = \ln(X)$, $Y' = \ln(Y)$... (voir le T.D.1), ou obtenir directement des approximations de degré 2, 3 etc... (voir le T.P.1).

Test sur le chapitre 1

1. On considère une distribution statistique dont les valeurs sont x_i, $i = 1, \cdots k$ et les effectifs correspondants n_i.
   Qu'est-ce qu'une fréquence ?
   Quelle formule permet de calculer la moyenne ? l'écart-type ?
   Que mesurent ces deux paramètres ?
2. Quelle est la définition de la médiane dans le cas d'une variable continue ?
3. On considère une distribution statistique de deux paramètres (X, Y) dont les fréquences sont $f_{i,j}$.
   Comment calcule-t-on les fréquences marginales ?
   Sous quelle condition X et Y sont-elles indépendantes ?
4. Quelle formule donne la covariance de (X, Y), le coefficient de corrélation linéaire ?
   Que mesure ce coefficient ?
5. Qu'est-ce que la droite de régression linéaire ?
   Comment la calcule-t-on ?

Chapitre 1 : Travaux dirigés

1. Un hypermarché note pour chaque classe de prix le nombre d'articles vendus :

Classe de prix	effectifs	eff. cumulés	fréquences	fréq. cumulées	carrés des écarts
[50,150[	80				
[150,250[	160				
[250,350[	720				
[350,450[	1680				
[450,550[	2720				
[550,650[	1760				
[650,750[	640				
[750,850[	160				
[850,950[	80				

   (a) Complétez le tableau (sauf la dernière colonne).

   (b) Tracer l'histogramme des fréquences, des fréquences cumulées, et la courbe de la fonction de répartition.

   (c) Calculer la moyenne $\overline{x}$ de cette distribution et déterminer la classe modale.

   (d) Déterminer la médiane, d'abord graphiquement, puis en interpolant la fonction de répartition F.

   (e) Remplir la dernière colonne (carrés des écarts à la moyenne) et calculer les paramètres de dispersion : écart-type σ et étendue.

   (f) Quel est le pourcentage d'articles dont le montant est inférieur à 750 euros ?
   Entre 550 et 750 ?
   (plus difficile :) moins de 700 ? (faire une interpolation).

2. Voici le temps moyen en heures que passent 30 internautes chaque semaine sur le web :
   3,4,4,5,5,5,5,5,5,6,6,6,6,7,7,7,7,7,8,8,9,10,10,10,10,10,10,12,55,60.

   (a) Calculer la moyenne, la médiane et le (ou les) mode(s) de cette donnée statistique. Lequel de ces paramètres de position vous parait ici le plus représentatif ?

   (b) Calculer et comparer son écart-type et son demi-écart interquartile ($|x_{3/4}-x_{1/4}|/2$). Lequel de ces paramètres de dispersion vous parait ici le plus représentatif ?

3. On considère les relevés de notes X d'une classe de 35 étudiants :

16.5	13.5	2.5	8.5	17.5	9	16	9.5	10.5	9.5
15	11.5	8.5	6	5.5	6.5	7.5	12	5	12.5
7	9.5	5	16	7	16.5	11	11.5	18.5	13.5
15	11.5	15	9	7					

   (a) Etablir le tableau d'effectifs correspondant, ainsi que le diagramme en bâtons.

   (b) Déterminer la moyenne m de la classe. On pose $Y = X - m$. Ecrire le tableau correspondant à Y puis calculer sa moyenne. Pouvait-on le prévoir ? Enoncer et démontrer un résultat général.

(c) Déterminer l'écart-type σ de la classe. On pose $Z = \frac{X-m}{\sigma}$. Dresser le tableau correspondant à Z. Calculer la moyenne et l'écart-type de Z. Enoncer et démontrer un résultat général concernant Z.

*Remarque : Z est appelé "**variable centrée réduite**" associée à X. Elle calcule à partir de l'origine "moyenne", dans l'unité "écart-type". Cette normalisation opermet de comparer différentes distributions : voir l'exercice suivant.*

4. Utilisation d'une variable centrée réduite. Une étudiante a obtenue la note de 84/100 à un examen de mathématiques pour lequel la note moyenne était de 76 avec un écart-type de 10. A l'examen final d'informatique, elle a obtenu la note de 90/100 pour un examen de moyenne 82 et d'écart-type 16. Dans quelle matière est-elle la meilleure ?
5. Le tableau ci-dessous donne les valeurs expérimentales de la pression P d'une masse de gaz donnée en fonction de son volume V :

V (cm^3) :	54,3	61,3	72,4	88,7	118,6	194
P (kg/cm^2) :	61,2	40,5	37,6	28,4	19,2	10,1

(a) Dessiner le nuage de points de cette distribution.

(b) On recherche une relation de la forme $P = CV^\alpha$, où C et α sont deux constantes. Par quel changement de variable de la forme $Q = f(P)$ peut-on ramener cette relation à une relation linéaire ?

(c) En calculant le coefficient de corrélation linéaire entre les variables Q et V, est-il raisonnable de chercher un tel ajustement ?

(d) Si oui, déterminer C et α à l'aide des données expérimentales.

(e) Estimer P lorsque $V = 100\ cm^3$.

Travail personnel :

Au cours d'une séance d'essais, un pilote automobile doit stopper le plus rapidement possible son véhicule lorsqu'il recoit un signal sonore. On mesure la vitesse du véhicule v_i juste avant le freinage et la distance d'arrêt y_i correspondante. Les six essais donnent le tableau :

vitesses (en km/h) v_i :	27	43	62	80	98	115
distance d'arrêt (en m.) y_i :	6,8	20,5	35,9	67,8	101,2	135,8

On veut vérifier que la distance d'arrêt est proportionnelle au carré de la vitesse. On pose donc $x_i = v_i^2$.

1. Tracer le nuage de points (x_i, y_i) avec les unités 1 cm pour 1000 en x et 1 cm pour 10 en y.
2. Calculer le point moyen de cette distribution, les variances et la covariance de cette distribution, à 10^{-2} près.
3. Calculer le coeffcient de régression linéaire : peut-on approximer linéairement cette distribution statistique ?
4. Déterminer l'équation $y = ax + b$ de la droite de régression linéaire (a à 10^{-4} près, et b à 10^{-2} près). Tracer cette droite dans le repère précédent.
5. A l'aide de cette équation, déterminer la vitesse correspondant à une distance d'arrêt de 180 m.
