Généralités, vocabulaire.

L'objet de la statistique est l'étude de certaines caractéristiques des éléments d'un ensemble fini.
Cet ensemble est appelé population, et les éléments sont les individus.

Les caractéristiques étudiées sont appelées caractères ou variables.
Une variable est:
- qualitative si elle ne peut pas faire l'objet de calculs
- quantitative si elle est numérique et si elle peut faire l'objet de calculs

- continue si elle peut prendre toute valeur dans un intervalle
- discrète dans le cas contraire

Les modalités d'une variable discrète sont les valeurs que peut prendre cette variable.

Dans le cas d'une variable continue, ou lorsqu'il y a trop de modalités, on regroupe les valeurs selon des intervalles disjoints. Chaque intervalle est une classe. Pour les calculs on utilise le centre de chaque classe.

On parle de série statistique pour désigner un recueil de données relatives à une variable numérique, ordonnées par ordre croissant des modalités.
Dans ce qui suit, $x_1, x_2, ..., x_p$ désignent les valeurs numériques prises par une série statistique, avec les effectifs respectifs $n_1, n_2, ..., n_p$.

L'effectif d'une modalité est le nombre d'individus présentant cette modalité.

L'effectif total de la population est : $N = \sum_{i=1}^{p} n_i = n_1 + n_2 + ... + n_p$.

La fréquence d'une modalité est le rapport entre l'effectif de cette modalité et l'effectif total de la population.
Une fréquence est un nombre compris entre 0 et 1.

La fréquence de la modalité 1 est donc : $f_1 = \frac{n_1}{N}$

L'effectif cumulé d'une modalité est la somme des effectifs des modalités qui lui sont inférieures ou égales.
L'effectif cumulé de la modalité 3 est donc : $N_3 = n_1 + n_2 + n_3$.
On définit de même la fréquence cumulée.

Les représentations de séries statistiques.

Il existe de nombreuses façons de représenter des données statistiques.

- Diagramme circulaire. (variable qualitative ou quantitative discrète)
Un disque est partagé en secteurs dont les angles sont proportionnels aux effectifs des modalités qu'ils représentent.

- Diagramme en bâtons. (variable qualitative ou quantitative discrète)
La hauteur du segment représentant une modalité est proportionnel à l'effectif de la modalité.

- Histogramme. (variable quantitative continue)
Dans le cas où les valeurs de la variable sont regroupées en classes, on utilise un histogramme où chaque classe est représentée par un rectangle dont l'aire est proportionnelle à l'effectif de la classe.

On appelle polygone des effectifs une ligne brisée rejoignant les sommets des bâtons ou les milieux des côtés supérieurs des rectangles .

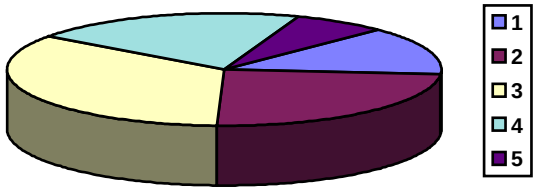
Citons aussi : le diagramme polaire, les pyramides, les représentations symboliques.

Exemples

Exemple 1 :
On considère la série statistique suivante:

modalités	1	2	3	4	5
effectifs	4	7	10	6	2

Représentations possibles:



secteurs circulaires

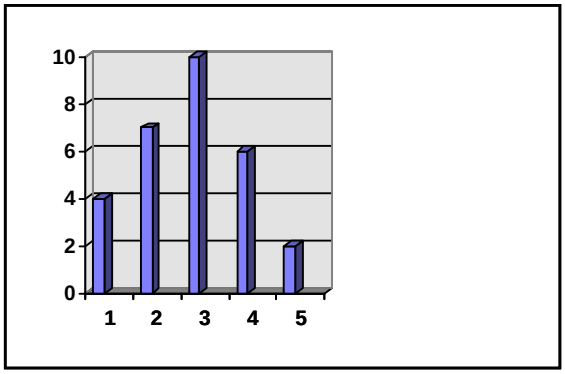


diagramme en bâtons

Exemple 2. Histogramme.

On considère les notes obtenues à un devoir de maths dans une classe de 28 élèves :

note	[0 ; 5 [	[5 ; 8 [	[8 ; 10 [	[10 ; 12 [	[12 ; 15 [	[15 ; 20]
effectif	1	4	7	9	5	2

Etablir un histogramme permettant de représenter cette série statistique.

On peut également construire le diagramme des effectifs cumulés croissants, qui représente la réponse à la question <.
Etablir le diagramme des effectifs cumulés croissants de cette série statistique.

Exercice 1

Utilisation d'un pont suivant l'heure de passage: la population est l'ensemble des véhicules franchissant le pont et la variable quantitative est le temps, variable continue.

heure	[0 ; 6 [	[6 ; 8 [	[8 ; 11 [	[11 ; 14 [	[14 ; 17 [	[17 ; 19 [	[19 ; 24 [
nombre de véhicules en milliers	4,2	6	5,7	6,6	6	8	5,5

Etablir un histogramme permettant de représenter cette série statistique.

Etablir le diagramme des effectifs cumulés croissants de cette série statistique.

Utilisation de logiciels

On sélectionne tout d'abord la plage de cellules contenant les données statistiques
Avec Excel 2003cliquer sur *Assistant graphique* et choisir le type de graphique
Avec Excel 2007 choisir *Insertion* puis *secteurs* ou *colonnes*

Les caractéristiques centrales et de dispersion.

Le mode est la modalité la plus fréquente. Dans le cas où les valeurs de la variable sont regroupées en classes, on parle de classe modale.
La médiane (M_e) d'une série statistique est une valeur du caractère qui partage la population en deux sous-ensembles de même effectif. La médiane est appelée aussi second quartile.
On définit la moyenne $\bar{x}$ et l'écart-type σ d'une série statistique par :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p} \qquad \sigma = \sqrt{\frac{n_1x_1^2 + n_2x_2^2 + ... + n_px_p^2}{n_1 + n_2 + ... + n_p} - \bar{x}^2}$$

Remarque:
Dans le cas d'une répartition normale, le polygone des effectifs présente l'aspect d'une courbe en cloche symétrique par rapport à la moyenne (courbe de Gauss).
Dans ce cas les valeurs comprises entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$ représentent 68 % de l'effectif total, 95 % entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ et 99 % entre $\bar{x} - 3\sigma$ et $\bar{x} + 3\sigma$.

Exercice 2. Un centre de renseignements téléphoniques effectue une enquête auprès d'un échantillon de 100 clients sur le temps d'attente, exprimé en secondes, subi par la clientèle avant d'amorcer la conversation avec un employé. Les résultats de cette étude sont consignés dans le tableau suivant:

Temps d'attente (en secondes)	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20[	[20 ; 25[	[25 ; 30[	[30 ; 35[
Effectif cumulé croissant	10	26	50	74	86	96	100

1. Dresser le tableau statistique dans lequel figurera les classes, les centres des classes, les effectifs et les effectifs cumulés croissants et décroissants.
2. a) Calculer la moyenne et l'écart type de cette série arrondis à 0,1. Commenter les résultats.
b) Calculer $\bar{x} - \sigma$ et $\bar{x} + \sigma$ arrondis à 0,1.
3. a) Tracer le polygone des effectifs cumulés croissants.
b) À l'aide du graphique, déterminer le nombre de clients dont le temps d'attente est inférieur ou égal à $\bar{x} - \sigma$ puis le nombre de clients dont le temps d'attente est supérieur ou égal à $\bar{x} + \sigma$.
c) En déduire le pourcentage des clients dont le temps d'attente est situé dans l'intervalle $[\bar{x} - \sigma$ et $\bar{x} + \sigma\ ]$.
4. Déterminer les quartiles. Commenter les résultats.

Exercice 3. Des études ont été réalisées sur les masses de 57 individus d'une population de petits rongeurs adultes. Les résultats sont les suivants :

Masse (g.)	150	151	152	153	154	155	156	157	158
Effectif	3	4	7	8	12	10	7	5	1

Déterminer la moyenne, l'écart-type et représenter cette série.

Séries statistiques à 2 variables.

Il peut arriver que, pour une population donnée, on s'intéresse simultanément à deux caractères, appelées aussi variables statistiques.
Une série statistique double est souvent donnée sous forme d'un tableau à double entrée.
Les lignes correspondent aux valeurs de l'un des caractères, les colonnes aux valeurs de l'autre caractère.
Le tableau se complète par une ligne et une colonne « total » appelées variables marginales.

Exemple : Le tableau ci-dessous décrit le sexe et la répartition des loisirs préférés des élèves de plusieurs classes de Terminale.

Sexe \ Loisir	Sport	Lecture	Musique	Danse	Sorties	
Masculin	14	7	14	3	10	
Féminin	13	18	20	8	19	

1. Quel est l'effectif total de la population interrogée? Combien y a-t-il de garçons? Combien de filles ?
2. Quel est le pourcentage de garçons préférant le sport dans l'ensemble de la population?
3. Quel est le pourcentage de garçons préférant le sport parmi les garçons ? Quel est le pourcentage de garçons préférant le sport parmi les élèves préférant le sport?

Exercice 4

Un laboratoire veut tester l'efficacité d'un vaccin sur des souris. Certaines ont été vaccinées, d'autres pas. Toutes ont reçu le virus de la maladie considérée. Certaines ont développé la maladie, d'autres pas.
Voici les informations dont on dispose:

- le laboratoire a effectué cette expérience sur 175 souris au total;
- 90 souris ont été vaccinées
- 120 souris ont développé la maladie et, parmi celles-ci, 65 avaient été vaccinées.

1. Recopier et compléter le tableau ci-après.

	Souris ayant développé la maladie	Souris n'ayant pas Développé la maladie	Total
Souris vaccinées			
Souris non vaccinées			
Total			

2. En arrondissant chaque résultat à l'entier le plus proche, calculer le pourcentage
a. de souris n'ayant pas développé la maladie;
b. de souris ayant développé la maladie, parmi celles qui n'ont pas été vaccinées
c. de souris ayant développé la maladie, parmi celles qui ont été vaccinées.

3. Que pensez-vous de l'efficacité de ce vaccin sur les souris?

Nuage de points, ajustement affine par la méthode des moindres carrés.

.

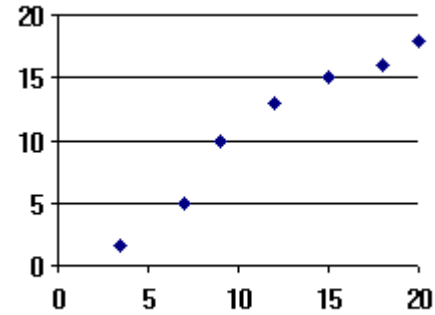
Représentation graphique d'une série double

Lorsqu'une série double porte sur deux caractères quantitatifs x_i et y_i .On peut représenter sur un graphique les points M_i (x_i , y_i). On appelle cette représentation le nuage de points associé à la série.

Considérons par exemple la série suivante:

x_i	3	7	9	12	15	18	20
y_i	2	5	10	13	15	16	18

Le nuage de points correspondant est le suivant:

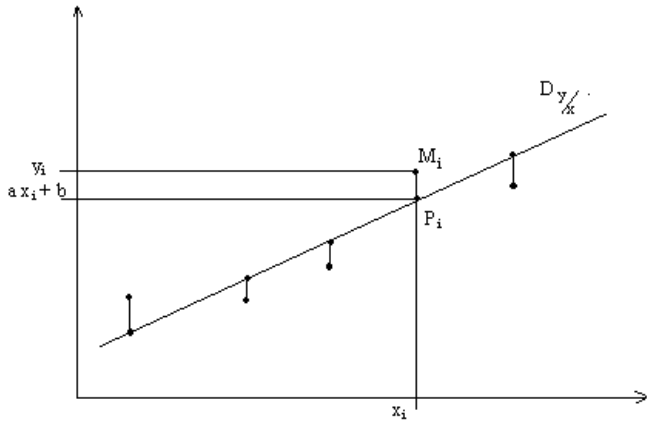


On appelle point moyen du nuage le point G ($\bar{x}$, $\bar{y}$). $\bar{x}$ et $\bar{y}$ étant les moyennes de chacune des variables.

Ajustement affine par la méthode des moindres carrés:

Lorsque les points du nuage sont approximativement alignés, on peut procéder à un ajustement linéaire en traçant une droite le plus près possible de ces points. Cette droite (D) représente une fonction affine
Pour cette raison, on parle aussi d'ajustement affine.
Ce type d'ajustement permet d'évaluer la valeur du caractère y_i pour des valeurs de x_i autres que celles des points du nuage.
Il existe de nombreuses manières d'obtenir une droite d'ajustement. Il est possible de montrer que les meilleures solutions sont obtenues quand la droite D passe par le point moyen de la série.

On considère une droite quelconque passant relativement près des points du nuage. A chaque point M_i du nuage on associe un point P_i de la droite ayant même abscisse.



Parmi les droites possibles on cherche une droite d'équation $y = ax + b$ telle que la somme des carrés des différences soit minimale, c'est à dire telle que : $\sum_{i=1}^{p}(M_iP_i)^2 = \sum_{i=1}^{p}(y_i - (ax_i + b))^2$ soit minimale.

Cette droite est appelée droite d'ajustement ou **droite de régression de *y* en *x*** : $D_{y/x}$.

On démontre que cette droite passe par le point moyen G ($\bar{x}$, $\bar{y}$) du nuage et que :

$a = \frac{\sigma_{xy}}{\sigma_x^2}$ $\quad b = \bar{y} - a\bar{x}$ $\quad$ avec $\quad \sigma_{xy} = \frac{\sum_{i=1}^{p} x_i y_i}{\sum_{i=1}^{p} n_i} - \bar{x}\bar{y}$

De la même façon on peut définir une **droite de régression de *x* en *y*** (les écarts considérés étant alors horizontaux) qui a pour équation :

$D_{x/y}$: $x = a' y + b'$ $\quad$ avec $\quad a' = \frac{\sigma_{xy}}{\sigma_y^2}$ $\quad$ et $\quad b' = \bar{x} - a'\bar{y}$

La méthode des moindres carrés crée une dissymétrie entre les deux variables statistiques. Suivant l'utilisation de l'ajustement, on privilégie l'une des deux droites de régression : $D_{y/x}$ pour obtenir des valeurs de *y*, $D_{x/y}$ pour obtenir des valeurs de *x*.

Le **coefficient de correlation** r traduit l'angle formé par les deux droites de régression. $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

si $|r| = 1$ les deux droites sont confondues, l'ajustement affine est parfait

si $|r| > 0{,}8$ il y a une bonne corrélation entre les deux caractères, ce qui signifie que les points du nuage sont suffisamment alignés pour que l'on puisse ajuster le nuage par une droite.
On peut alors, connaissant une valeur de *x* ou de *y*, déterminer la valeur de l'autre variable avec une précision correcte.

Utilisation de logiciels

On sélectionne tout d'abord la plage de cellules contenant les données statistiques
Avec Excel 2003cliquer sur *Assistant graphique* et choisir *Nuage de points*
Avec Excel 2007 choisir *Insertion/Nuage de points*
Pour obtenir une droite d'ajustement selon la méthode des moindres carrés, faire un clic droit sur un des points du nuage et choisir *Ajouter une courbe de tendance,* sélectionner *Linéaire* dans la boîte de dialogue et cocher la case *Afficher l'équation sur le graphique.*

Exercice 5.

La taille moyenne d'un enfant entre 6 et 33 mois est donnée par le tableau suivant :

x	6	9	12	15	18	21	24	27	30	33
y	66	71	74	77	80	83	85	88	90	92

où x désigne l'âge exprimé en mois, y désigne la taille exprimée en cm.

Placer le nuage de points de la série statistique (x ; y) dans un repère orthogonal (unités : 1 cm représente 2 mois en abscisses ; 1 cm représente 10 cm en ordonnées).

1. Donner une équation de la droite de régression $D_{y/x}$.

2. A l'aide de la droite de régression, résoudre graphiquement puis algébriquement les deux questions suivantes :

Quelle est la taille moyenne d'un enfant de 3 ans ?

A quel âge la taille moyenne d'un enfant est-elle de 75 cm ?

Exercice 6.

Le tableau suivant donne, dans une population féminine, la moyenne de la tension artérielle maximale en fonction de l'âge.

Age	36	42	48	54	60	66
Tension max	11,8	13,2	14	14,4	15,5	15,1

1. Représenter graphiquement le nuage de points de cette série statistique dans un repère orthogonal.
On graduera l'axe des abscisses à partir de 36 et l'axe des ordonnées à partir de 11. De plus, on prendra pour unités graphiques
- 0,5 cm pour une année;
- 2 cm pour une unité de tension.

2. Calculer le coefficient de corrélation linéaire.

3. Déterminer, par la méthode des moindres carrés, une équation de la droite de régression D de y en x.

On donnera les coefficients de l'équation de D à 10^{-1} près. Tracer D.

a. Déterminer graphiquement, en faisant apparaître les traits de construction utiles, la tension artérielle maximale prévisible pour une personne de 70 ans.

b. Vérifier le résultat précédent par le calcul en utilisant l'équation de la droite de régression D.

Exercice 7.

Un marchand de glaces ambulant vend, sur une plage, des glaces qu'il transporte dans un seau isotherme. Il s'aperçoit qu'à la fin de son circuit les glaces restantes ne supportent pas le retour au fournisseur. Il décide de faire une étude statistique de ses ventes sur une saison pour un circuit sur la plage, en fonction de la température ambiante, afin d'avoir le moins de pertes possible.
Il obtient :

Température t_i en ^{0}C	27	28	29	30	31	32	33	34	35	36
Nombre n_i de glaces vendues	4	9	12	16	25	32	40	49	64	72
$y_i = \sqrt{n_i}$										

1. Le tracé du nuage de points de coordonnées $(t_i ; n_i)$ dans un repère ne fournit pas un alignement suffisant. Le marchand décide de poser $y_i = \sqrt{n_i}$. Compléter le tableau précédent.

2. On choisit pour unité graphique le centimètre ; on commencera la graduation à 25 ^{0}C sur l'axe des abscisses. Construire le nuage de points de coordonnées $(t ; y)$.

3. Calculer le coefficient de corrélation linéaire.

4. Déterminer, par la méthode des moindres carrés, une équation de la droite de régression D de y en t.

5. La météo annonce pour le lendemain une température de 38 ^{0}C ; calculer une estimation de y à deux décimales et en déduire une estimation du nombre de glaces n qu'il peut espérer vendre ce jour-là.

Exercice 8.

Au cours d'une séance d'essai, un pilote d'automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule. Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule.
Pour six expériences, on a obtenu les résultats suivants :

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

Le nuage de points associé ne peut être ajusté linéairement de manière correcte (ce résultat sera admis).
On pose pour les six valeurs de v_i, $x_i = v_i^2$ et on considère la double série (x_i, y_i), $1 \leq i \leq 6$.

1. Compléter le tableau précédent

2. Dans un repère orthogonal, construire le nuage de points associé à cette nouvelle série double.

(Les x_i en abscisses avec 1 cm pour 1000, les y_i en ordonnées avec 1 cm pour 10.)

3.a. Déterminer l'équation de la droite de régression $D_{y/x}$. Tracer cette droite dans le repère précédent.

3.b. A l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m puis la vitesse correspondante du véhicule.

3.c. Quelle est la distance d'arrêt estimée correspondant à une vitesse de 150 km/h ?

3.d. Le manuel du code de la route donne, pour calculer la distance d'arrêt (en mètre), la méthode suivante :
« Prendre le carré de la vitesse exprimée en dizaines de kilomètres par heure. »
Comparer le résultat obtenu au **3.c.** à celui que l'on obtiendrait par cette méthode.

Exercice 9.

Dans une plantation, on a étudié la masse moyenne M par plante obtenue en fin de croissance en fonction de la densité d de plantation. (Les unités sont : 100 g pour M, 1 plante par mètre carré pour d.)

1. Recopier et compléter le tableau suivant en donnant X et Y à 0,01 près :

d	2	4	5	10	15
M	17,68	6,11	4,57	1,58	0,85
$X = \ln d$		1,39			
$Y = \ln M$		1,81			

2.a. Placer dans un repère orthonormal (unité: 5 cm) le nuage de points de la série statistique (X ; Y).

2.b. Déterminer le point moyen de ce nuage.

3. Vérifier qu'une équation de la droite D d'ajustement de Y en X est : $Y = -1{,}5\,X + \ln 50$.

3.a. Tracer D dans le repère précédent.

3.b. En partant de cette équation montrer que la relation entre masse et densité de plantation peut s'écrire : $M = Cd^a$, ou C et a sont des réels à déterminer.

3.c. Déterminer la masse moyenne d'une plante en fin de croissance, pour une densité de 9 plantes/m².