Arts et Métiers
Sciences et Technologies

Statistiques

Mohamad FARHAT

2022 - 2023

Objectifs : ANALYSE DES DONNEES

- ✓ Moyenne, Etendue, Variance, Ecart absolu
- ✓ Quartile, Médiane
- ✓ Droite des moindres carrés + Coefficient de corrélation
- ✓ Loi d'échantillonnage de la moyenne et de la fréquence
- ✓ Intervalle de confiance
- ✓ Test d'hypothèse

Arts et Métiers Sciences et Technologies

Introduction

La statistique est la science qui consiste à réunir des données chiffrées, à les analyser et à les commenter.

ANALYSER
Le but de la statistique descriptive est de structurer et de représenter l'information contenue dans les données

Une étude statistique s'effectue sur un ensemble appelé **population** « auxquels on s'intéresse» dont les éléments sont appelés **individus** et consiste à observer et étudier un même aspect sur chaque individu, appelé **caractère**.

Arts et Métiers Sciences et Technologies

Introduction

On distingue deux types de caractères :

Quantitatifs
« Valeurs numériques »

(nombre d'élèves, notes,…)

Qualitatifs
« Valeurs non numériques »

(couleur des yeux, profession,…)

Discrète
« les valeurs du caractère sont isolées »

par exemple :
nombre d'enfants

Continue
« valeurs du caractère sont regroupées en intervalles appelés classes »

par exemple : taille
[1.60,1.70[, [1.70,1.80[…

Arts et Métiers Sciences et Technologies

Série statistique quantitative

On considère comme population 20 adolescents et le caractère est le poids exprimé en Kg :

75; 50; 64; 64; 48; 50; 65; 81; 70; 75; 52; 50; 48 ; 64; 55; 75; 64 ; 50 ;81; 70

La série sera plus lisible

On appelle effectif d'une valeur le nombre d'individus possédant le caractère de cette valeur

Poids X_i	48	50	52	55	64	65	70	75	81	Total=N
Effectifs n_i	2	4	1	1	4	1	2	3	2	20

Diagramme en bâtons

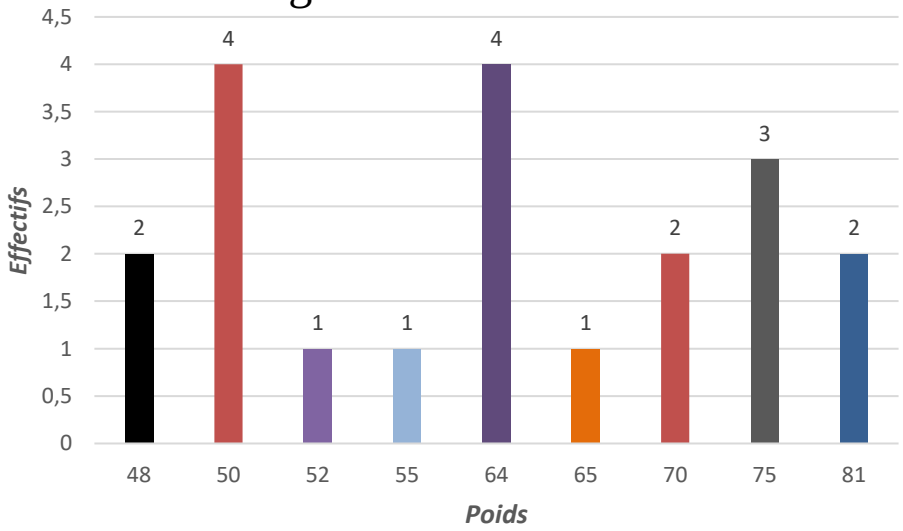
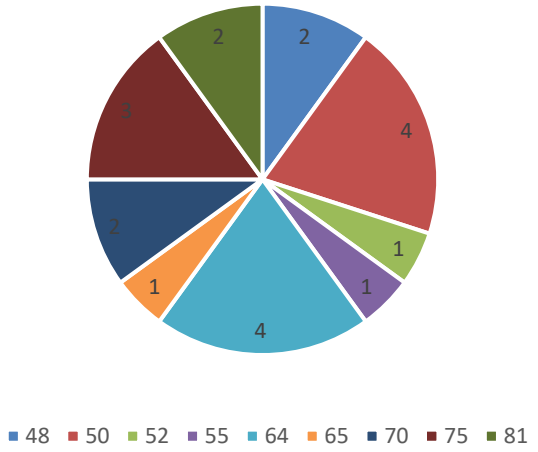


Diagramme circulaire



Arts et Métiers Sciences et Technologies

Série statistique quantitative

On appelle **fréquence** d'une valeur :
Le quotient de l'effectif de cette valeur par l'effectif total de la population.

$$f_i = \frac{n_i}{N}$$

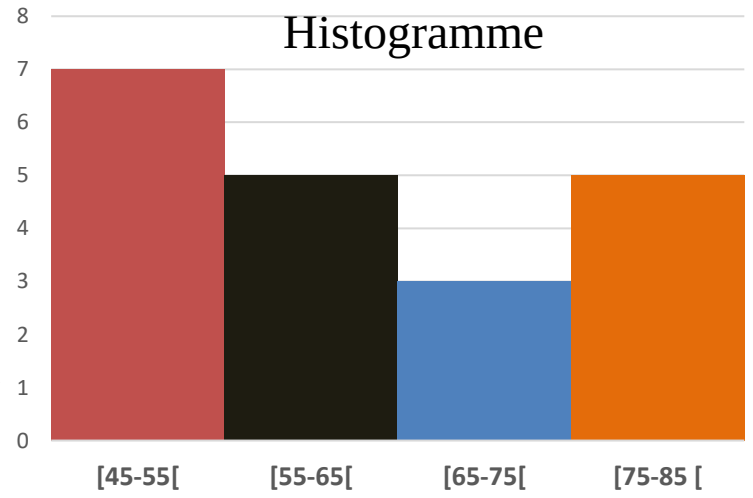
Ainsi, la fréquence de 55 est 1/20

$$\sum_{i=1}^{n} f_i = 1$$

Poids X_i	48	50	52	55	64	65	70	75	81	Total=N
Effectifs n_i	2	4	1	1	4	1	2	3	2	20
Fréquence f_i	0,1	0,2	0,05	0,05	0,2	0,05	0,1	0,15	0,1	1

Création de classe ⇒ perte d'informations

Poids	[45-55[	[55-65[	[65-75[	[75-85 [	Total
Effectifs	7	5	3	5	20



Arts et Métiers Sciences et Technologies

Série statistique quantitative

Tableau de répartition

Poids X_i	48	50	52	55	64	65	70	75	81	Total=N
Effectifs n_i	2	4	1	1	4	1	2	3	2	20
Fréquence f_i	0,1	0,2	0,05	0,05	0,2	0,05	0,1	0,15	0,1	1
Porcentages	10%	20%	5%	5%	20%	5%	10%	15%	10%	100%

$$Pourcentage = \frac{n_i}{N} \times 100 = f_i \times 100$$

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

Le Mode : « **valeur dominante » :** est la valeur la plus représentée d'une variable quelconque (non classée) dans une population donnée (la plus grande fréquence)

« Il est toutefois possible qu'il n'y ait aucun mode ou qu'il y ait plusieurs modes »

Dans le cas d'une variable quantitative continue ou discrète (classée) : classe modale : classe dont la fréquence par unité d'amplitude est la plus élevée

« Possibilité de déterminer, dans la classe modale, la valeur exacte du mode »

l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution $X = X_{max} - X_{min}$

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

La médiane M_e : Partage la distribution en deux parties d'effectifs égaux

Effectif cumulé de la k ème individu ou classe : $N_k = \sum_{i=1}^{k} n_i$

Fréquence cumulée de la k ème individu ou classe : $F_k = \frac{N_k}{N}$

Poids X_i	48	50	52	55	64	65	70	75	81	Total
Effectifs	2	4	1	1	4	1	2	3	2	20
Eff cumulé	2	6	7	8	12	13	15	18	20	
Fréquence	0,1	0,2	0,05	0,05	0,2	0,05	0,1	0,15	0,1	1
Fréq cumulée	0,1	0,3	0,35	0,4	0,6	0,65	0,75	0,9	1	

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

La fonction cumulative, appelée aussi fonction de répartition, notée F, est la fonction, qui a tout réel t, associe $F(\mathrm{t})$ la proportion d'individus de la population pour lesquels on a observé une valeur de la variable X plus petite ou égale à t.

Propriétés de la fonction de répartition F :

- F est croissante, i.e pour tous réels $t_1 \leq t_2$, on a $F(t_1) \leq F(t_2)$
- $\forall\ t \leq x_0$, où x_0 désigne la borne gauche de la première classe, $F(t) = 0$.
- $\forall\ t \geq x_n$, où x_n désigne la borne droite de la dernière classe, $F(t) = 1$.
- Lorsque X est une variable continue,

  F n'est connue pour les valeurs de X égales aux extrémités des classes. On considère alors F affine entre ces valeurs, parce que l'on suppose que les classes forment des entités homogènes

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Calcul de la médiane

La médiane M_e : Partage la distribution en deux parties d'effectifs égaux

Il suffit de compter le nombre de valeurs (N) et de les ordonner en ordre croissant :
Si Total=N est :

- *Pair : la médiane est la moyenne des valeurs de rang N/2 et (N/2)+1*
- *Impair : alors (N+1)/2 est le rang de la médiane*

Rang	1	2	3	4	5	6	7
Temps	24,1	25	25,2	25,6	25,7	26,1	27,8

$N = 7$ impair $\Rightarrow$ $\frac{N+1}{2} = \frac{8}{2} = 4$ (rang) $\Rightarrow$ La médiane est de 25,6 secondes

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Calcul de la médiane

La médiane M_e : Partage la distribution en deux parties d'effectifs égaux

La médiane est la plus petite valeur pour laquelle la fréquence relative cumulée atteint au moins 50 %

Poids X_i	48	50	52	55	64	65	70	75	81	Total
Effectifs	2	4	1	1	4	1	2	3	2	20
Eff cumulé	2	6	7	8	12	13	15	18	20	
Fréquence	0,1	0,2	0,05	0,05	0,2	0,05	0,1	0,15	0,1	1
Fréq cumulée	0,1	0,3	0,35	0,4	0,6	0,65	0,75	0,9	1	

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

Les quartiles sont trois valeurs du caractère qui partage la série statistique en quatre groupes de même effectif :

- le 1-ier quartile Q_1 est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.25

- le 2-ième quartile est confondu avec la médiane $Q_2 = M_e$.

- le 3-ième quartile Q_3 est la valeur du caractère à partir de laquelle la fréquence cumulée atteint ou dépasse 0.75

Intervalle interquartile : $Q_3 - Q_1$

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

Les Moyennes

Soit une population de N individus, $x_1, \dots, x_N$

K =nbre de valeurs distinctes de X et f_i = fréquence de x_i

✓ ***Moyenne arithmétique (pondérée)***

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{K} n_i x_i = \sum_{i=1}^{K} f_i x_i$$

« Pour les variables continues classées ou les variables discrètes classées, on utilise le centre des classes »

✓ ***Moyenne quadratique***

$$m_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2} = \sqrt{\sum_{i=1}^{K} f_i {x_i}^2}$$

Arts et Métiers Sciences et Technologies

Série statistique quantitative

Caractéristique

Les Moyennes

Soit une population de N individus, $x_1, \dots, x_N$

K =nbre de valeurs distinctes de X et f_i = fréquence de x_i

✓ ***Moyenne harmonique***

$$m_{-1} = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{1}{\sum_{i=1}^{K} \frac{f_i}{x_i}}$$

Si les observations sont des nombres réels positifs

✓ ***Moyenne géométrique***

$$M_g = \sqrt[n]{x_1 x_2 \quad x_N} = \sqrt[n]{x_1^{n_1} \dots x_K^{n_k}} = x_1^{f_1} \; x_2^{f_2} \; \dots \; x_K^{f_K}$$

$$min(x_1, \dots, x_N) \leq m_{-1} \leq M_g \leq \bar{x} \leq m_2 \leq max(x_1, \dots, x_N)$$

Arts Sciences et Technologies et Métiers

Série statistique quantitative

Caractéristique

Une étude statistique menée sur une population de ménages a montré que

30% de ces ménages ont 1 enfant
40% ont 2 enfants
15% 3 enfants
10% 4 enfants
5% 5 enfants.

1. Calculer le nombre moyen d'enfants par ménage.
2. Calculer la moyenne quadratique de la variable "Nombre d'enfants par ménage"

Arts Sciences et Technologies et Métiers

Série statistique quantitative

Caractéristique

EXERCICE 2) On achète des Dollars
une première fois pour 100 Euros au cours de 0, 87 Euro le Dollar,
puis on achète
une seconde fois pour 100 Euros également mais au cours de 0, 71 Euro le Dollar.

1 Calculer le montant total des Dollars achetés lors de ces opérations.

2 Le cours moyen du Dollar, C_m, pour l'ensemble de ces opérations est par définition le cours de C_m euro le Dollar, qui aurait permis l'achat en une seule fois, de 255, 79 Dollars pour 200 euros. Calculer C_m.

EXERCICE 3) Un automobiliste parcourt 40 kilomètres à 60km/h puis 40 kilomètres à 120km/h. Calculer V_m la vitesse moyenne en km/h sur l'ensemble de ce trajet de 80km.

Arts et Métiers Sciences et Technologies

Caractéristique : Les indicateurs de dispersion

Soit une population de N individus, $x_1, \dots, x_N$ les valeurs d'une variable quantitative discrète X.
K =nbre de valeurs distinctes de X et f_i = fréquence de x_i

- ***L'étendue*** d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution $X = X_{max} - X_{min}$

- ***Variance :***

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{K} f_i(x_i - \bar{x})^2$$

- ***Ecart-type :*** $\sigma = \sqrt{V(X)}$

- Écart absolu moyen à la moyenne : $e_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}| = \sum_{i=1}^{K} f_i|x_i - \bar{x}|$

- Écart absolu moyen à la médiane : $e_{M_e} = \frac{1}{N}\sum_{i=1}^{N}|x_i - M_e| = \sum_{i=1}^{K} f_i|x_i - M_e|$

Caractéristique : Les indicateurs de dispersion

- *Variance :*

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \sum_{i=1}^{K} f_i(x_i - \overline{x})^2$$

$$\textit{Formule de huygens} : \; V(X) = m_2^2 - \overline{x}^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{K} n_i x_i\right)^2$$

$$= \sum_{i=1}^{K} f_i x_i^2 - \left(\sum_{i=1}^{K} f_i x_i\right)^2$$

Arts et Métiers Sciences et Technologies

Caractéristique : Les indicateurs de dispersion

Soit $\overline{x_T}$ la moyenne arithmétique totale, alors $\overline{x_T}$ est la moyenne arithmétique de $\overline{x_1}$ et $\overline{x_2}$, moyennes arithmétiques des deux groupes, pondérée par les "poids" des deux groupes :

$$\overline{x_T} = \frac{N1}{N_1 + N_2}\overline{x_1} + \frac{N_2}{N_1 + N_2}\overline{x_2}$$

La variance de X, appelée variance totale est définie par :

$$V_T = \overbrace{\left(\frac{N_1}{N_1+N_2}\right)V_1 + \left(\frac{N_2}{N_1+N_2}\right)V_2}^{\text{moyenne des variances de deux groupes}} + \overbrace{\left(\frac{N_1}{N_1+N_2}\right)(\overline{x}_1 - \overline{x}_T)^2 + \left(\frac{N_2}{N_1+N_2}\right)(\overline{x}_2 - \overline{x}_T)^2}^{\text{variance des moyennes de deux groupes}}$$

Variance Intra-groupe — Variance Inter-groupe

Caractéristique : Les indicateurs de dispersion

Exercice

Les 25 étudiants d'un Master sont répartis en deux groupes, 13 étudiants sont dans le groupe 1 et les 12 autres dans le groupe 2. On a relevé leur note à un examen.

1 Calculer les moyennes et les écarts-types pour chacun des deux groupes.
2 Calculer la moyenne arithmétique totale et la variance totale.

Centres des classes	Classes de note	Effectifs du groupe 1	Effectifs du groupe 2
2	$[\,0;4\,[$	0	2
6	$[\,4;8\,[$	1	2
10	$[\,8;12\,[$	10	3
14	$[\,12;16\,[$	2	3
18	$[\,16;20\,]$	0	2

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

Introduction

- Apprendre comment analyser un phénomène quelconque en utilisant des méthodes statistiques
- Analyser des données et construire un modèle empirique (régression linéaire simple, interprétation des effets);

La régression linéaire est une relation stochastique entre une ou plusieurs variables.

Plusieurs domaines : la physique, la biologie, la chimie, l'économie...etc.

1- La régression linéaire simple où on explique une variable endogène par une seule variable exogène

Exemple : la relation entre la variable Revenu et la variable Consommation,

Il s'agit de la régression linéaire simple.

Régression simple $Y = AX + b$

2- La régression linéaire multiple qui représente la relation linéaire entre une variable endogène et plusieurs variables exogènes

Exemple : la demande d'un produit peut être expliquée par les grandeurs Prix, Revenu et Publicité

Régression multiple $Y = A_1X_1 + \cdots + A_kX_k + b$

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

Covariance

On appelle covariance des deux variables aléatoires réelles la quantité

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{xy}$$

Coefficient de corrélation

Le coefficient de corrélation linéaire mesure le degré d'association linéaire :

$$R(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{(\sum (Y_i - \bar{Y})^2)(\sum (X_i - \bar{X})^2)}} \in [-1,1]$$

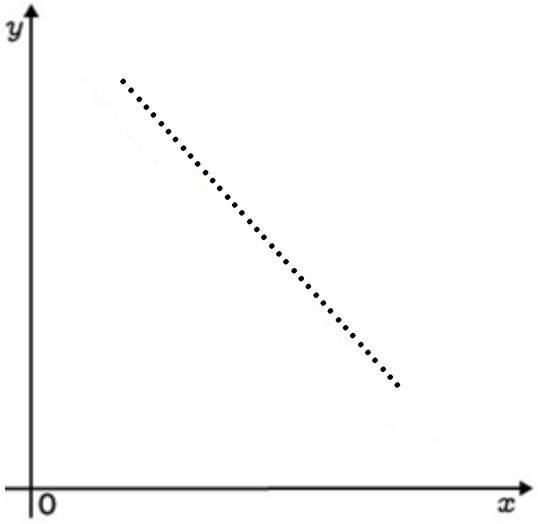
$|R|$ est près de 1 → le lien linéaire entre les deux variables est fort

R est près de 0 → le lien linéaire entre les deux variables est faible

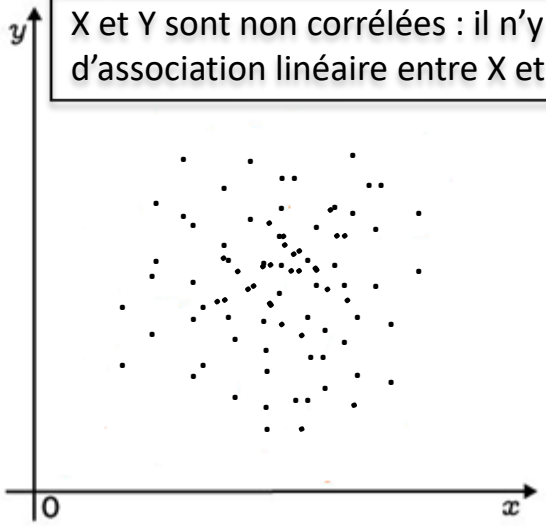
Analyse de données – Analyse des corrélations

Le coefficient de corrélation linéaire mesure le degré d'association linéaire :

$$R(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{(\sum (Y_i - \bar{Y})^2)(\sum (X_i - \bar{X})^2)}} \in [-1,1]$$

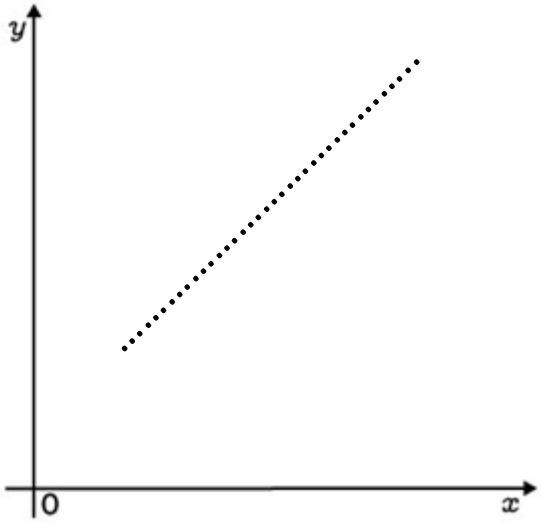


Corrélation linéaire parfaite négative $R = -1$



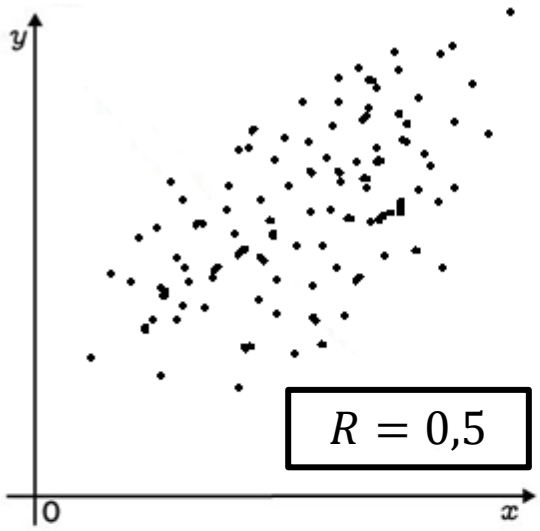
X et Y sont non corrélées : il n'y a pas d'association linéaire entre X et Y

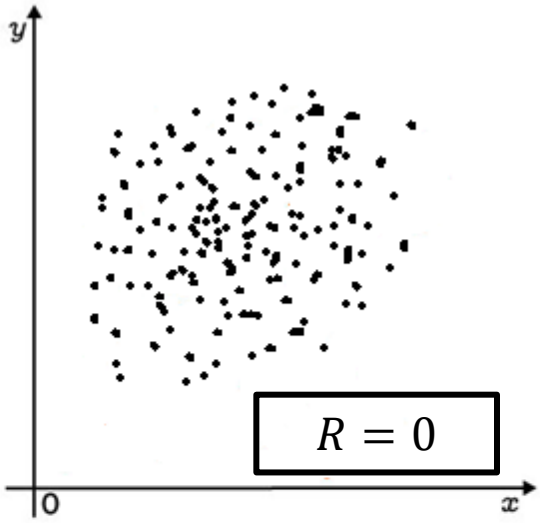
$R = 0$

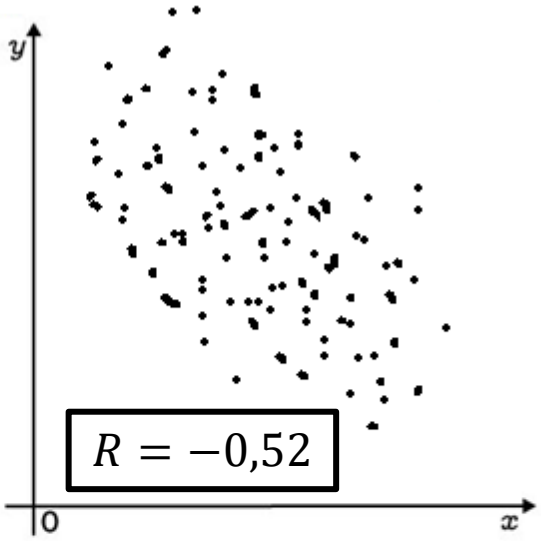


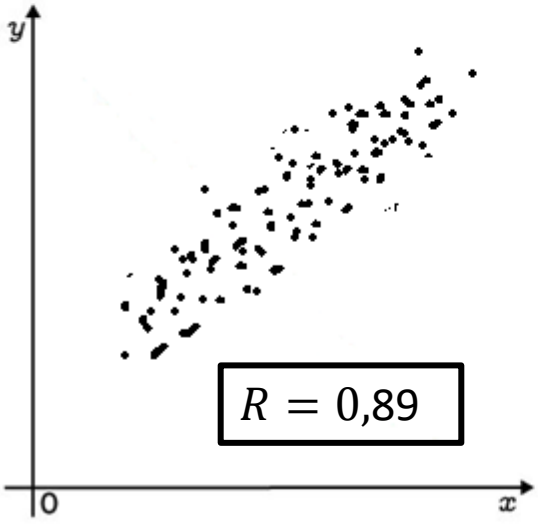
Corrélation linéaire parfaite positive $R = 1$

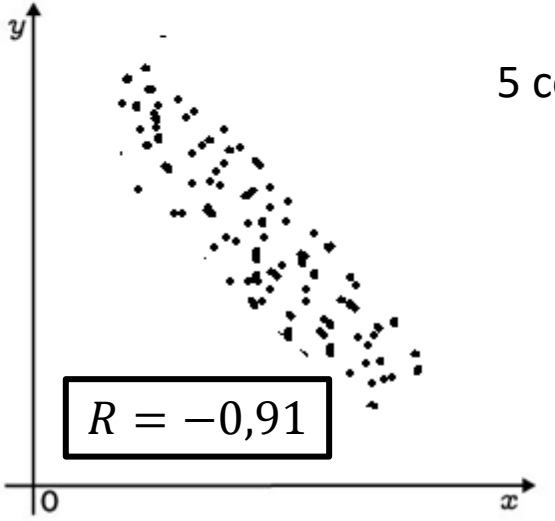
Liaison entre deux variables quantitatives











5 coefficients de corrélation linéaire

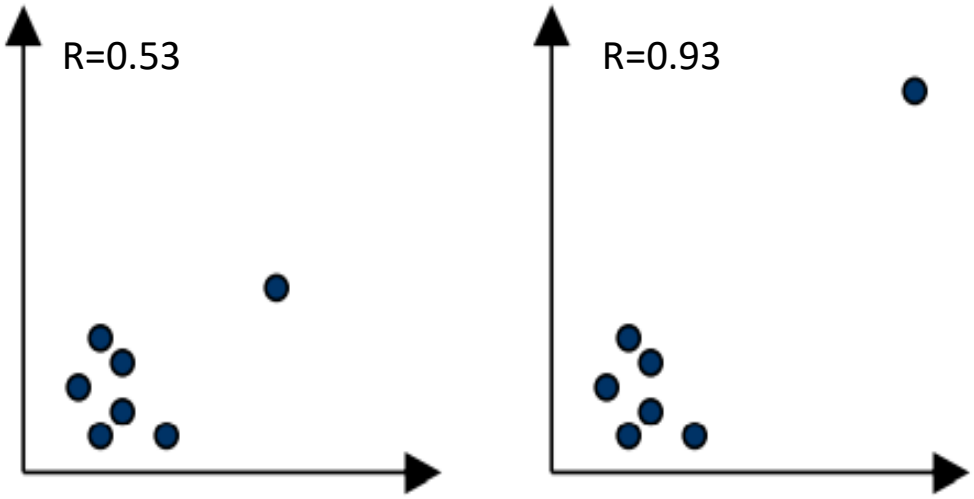
a) −0.91
b) −0.52
c) 0.00
d) 0.50
e) 0.89

Liaison entre deux variables quantitatives

Un coefficient positif indique que les deux variables évoluent dans le même sens

Un coefficient négatif indique qu'il existe une relation inverse entre les variables X et Y

Le coefficient de corrélation est un indice indépendant de la moyenne



Le coefficient de corrélation peut produire de hautes valeurs si des points isolés sont présents.

Liaison entre deux variables quantitatives

Introduction

Régression linéaire simple :

Exemple :

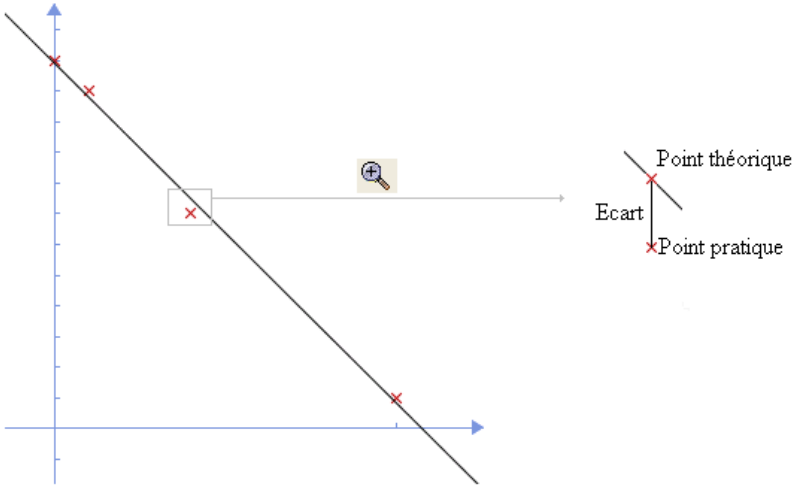
La tension U aux bornes d'une batterie de force électromotrice E et de résistance interne R est $U = E - RI$. On a procédé à différentes mesures :

Intensité mesurée (A) :	0	0,1	0,4	1
Tension mesurée (V) :	12	11	7	1

Le but de la régression simple est de chercher une fonction f telle que :

$$y_i = f(x_i)$$

La régression linéaire donne $E = 11{,}9$ et $R = 11{,}07$

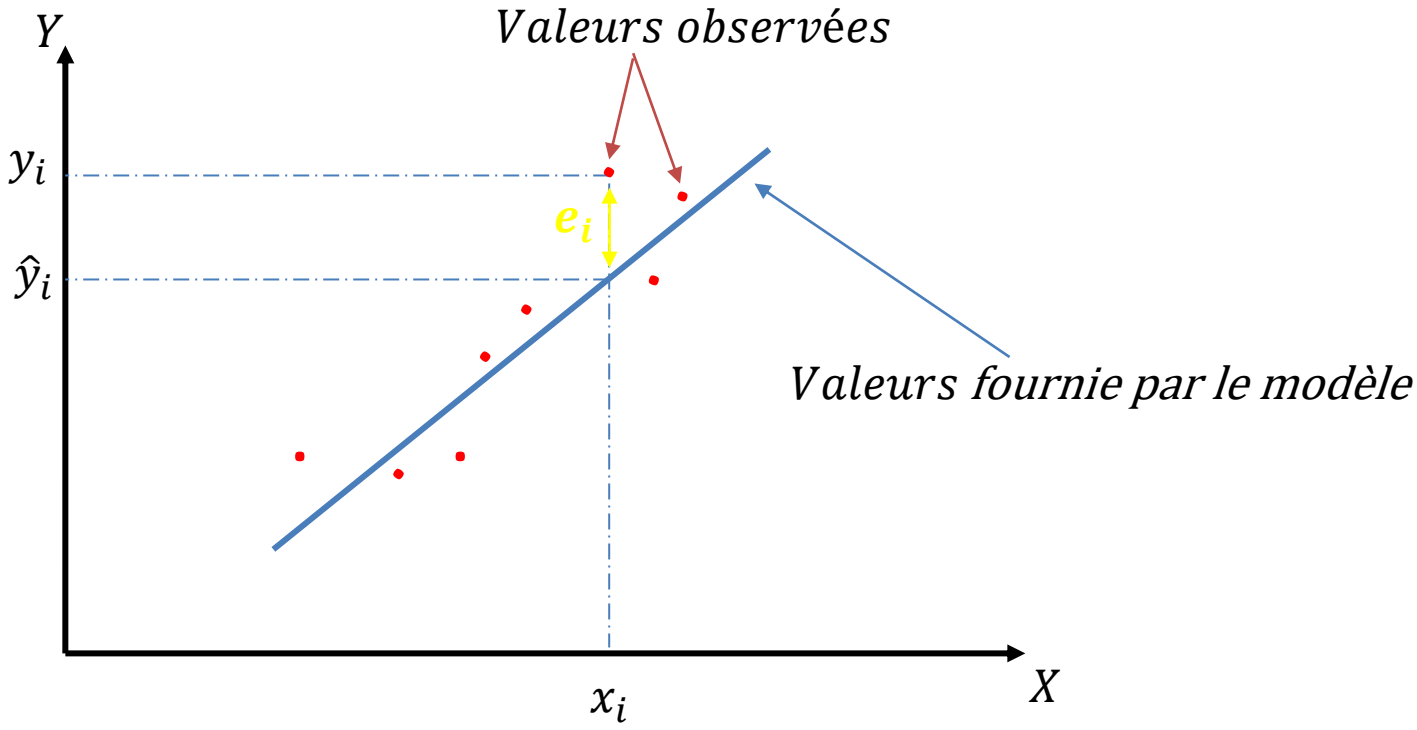


Minimisation de la somme des écarts au carré

Liaison entre deux variables quantitatives

Explication / Prédiction d'une variable à partir de l'autre

$$Y_i = aX_i + b + \varepsilon_i \qquad i = 1,..,n$$



Le résidu ***e*** est une évaluation du terme d'erreur ε

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

Régression linéaire simple :

Le modèle de régression linéaire simple est une variable endogène Y (dépendante ou réponse) expliquée par une seule variable explicative exogène X (indépendante)

$$Y = aX + b + \varepsilon$$

- ε cristallise toutes l'«insuffisance» du modèle
- ε quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle.

Estimation des paramètres par la méthode des Moindres Carrés Ordinaires (MCO)

En minimisant la somme des carrés des erreurs : $Min \sum_{i=1}^{n} {\varepsilon_i}^2$

$$Minimiser \sum_{i=1}^{n} (Y_i - aX_i - b)^2 = Minimiser\ S$$

Liaison entre deux variables quantitatives

Estimation des paramètres par la méthode des Moindres Carrés Ordinaires (MCO)

$$a = \frac{COV(X,Y)}{\sigma^2(X)}$$

$$b = \bar{Y} - a\bar{X}$$

$$\hat{Y} = aX + b$$

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

Exemple 1

L'étude statistique ci-dessous porte sur les poids respectifs des pères et de leur fil aîné.

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

1. Calculez la droite des moindres carrés du poids des fils en fonction du poids des pères.
2. Calculez la droite des moindres carrés du poids des pères en fonction du poids des fils.
3. Montrer que le produit des pentes des deux droites est égal au carré du coefficient de corrélation empirique entre les p_i et les f_i (ou encore au coefficient de détermination)

$$b = \bar{Y} - a\bar{X}$$

$$a = \frac{COV(X,Y)}{\sigma^2(X)}$$

$$\hat{Y} = aX + b$$

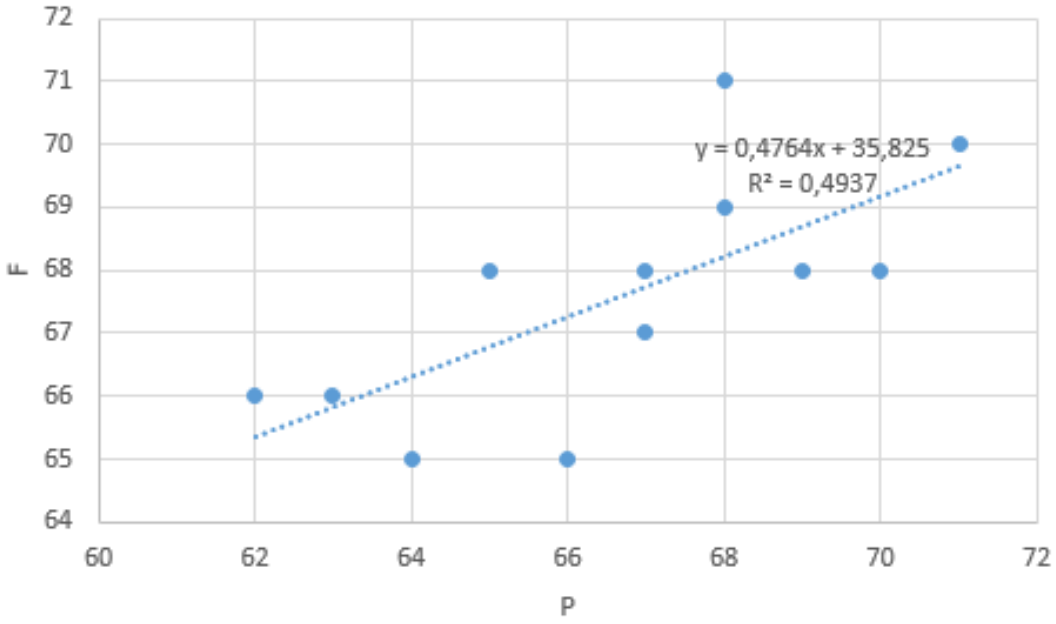
Liaison entre deux variables quantitatives

Exemple 1

1. Calculez la droite des moindres carrés du poids des fils en fonction du poids des pères.

   La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit

   $f = \hat{\alpha}_1 + \hat{\alpha}_2 p = 35{,}8 + 0{,}48p$



Betta 1

B18 = `=COVARIANCE(A2:A13;B2:B13)/(ECARTYPEP(A2:A13)^2)`

Betta 0

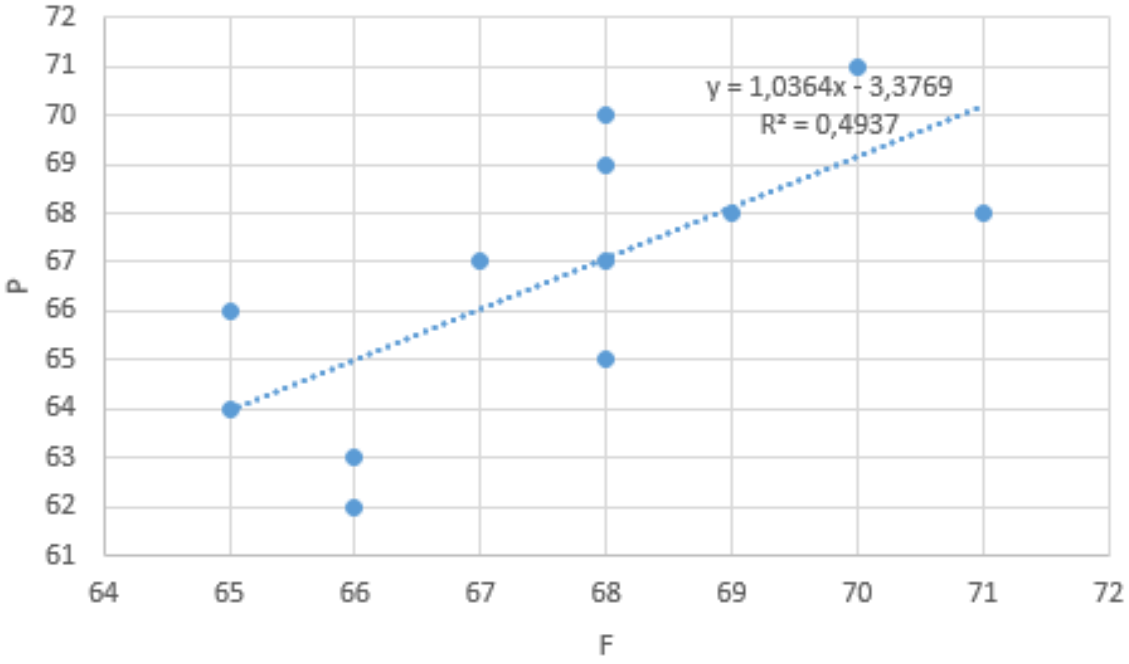
`=MOYENNE(B2:B13)-B18*MOYENNE(A2:A13)`

Liaison entre deux variables quantitatives

Exemple 1

2. Calculez la droite des moindres carrés du poids des pères en fonction du poids des fils.

La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit
$f = \hat{\beta}_1 + \hat{\beta}_2 f = -3{,}38 + 1{,}03f$



Exemple 1

3. Montrer que le produit des pentes des deux droites est égal au carré du coefficient de corrélation empirique entre les p_i et les f_i (ou encore au coefficient de détermination)

Le produit des pentes des deux droites est

$$\hat{\alpha}_2\hat{\beta}_2 = \frac{\left(\sum\ (f_i - \bar{f}\ )\ (p_i - \bar{p})\right)^2}{(\sum\ (f_i - \bar{f})^2)(\sum\ (p_i - \bar{p})^2)} = R^2$$

Où R^2 est le coefficient de détermination, carré du coefficient de corrélation linéaire

Liaison entre deux variables quantitatives

Exemple 2

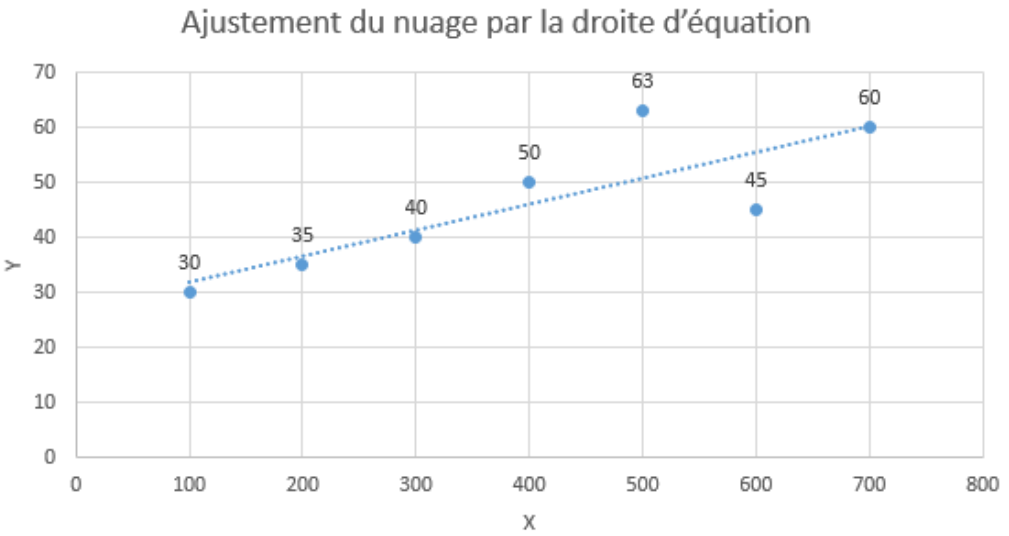
Nous disposons des données qui sont représentés dans le tableau suivant:

X_i	100	200	300	400	500	600	700
Y_i	30	35	40	50	63	45	60

Où X_i désigne les quantités consommées et
Y_i désigne le prix des quantités consommées

Utilisez les fonctions suivantes sur Excel

- $\widehat{\beta_0} = ORDONNEE.ORIGINE(Y;X)$
- $\widehat{\beta_1} = PENTE(Y;X)$



La droite qui ajuste le nuage de point est : $\widehat{Y_i} = 27{,}143 + 0{,}0475\ X_i$

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

$$On\ a\ l'égalité\ suivante : \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \widehat{Y_i})^2 + \sum (\widehat{Y_i} - \bar{Y})^2$$

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{\sum (\widehat{Y_i} - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$0 \leq R^2 \leq 1$$

$$R^2 = 0 \Rightarrow \sum (\widehat{Y_i} - \bar{Y})^2 = 0$$

Modèle Intéressant

Test global de significativité de la régression

Le coefficient de détermination R^2 est-il significatif ??
Le modèle est-il intéressant ??

Statistiquement, le test s'écrit :

H_0: *(Hypothèse nulle) Le modèle n'amène rien dans l'explication de Y* $(\hat{\beta}_1 = 0)$
$H_1 : \hat{\beta}_1 \neq 0$ *Le modèle est globalement significatif*

Statistique de test :

$$F_{obs} = (n-2)\frac{R^2}{1-R^2}$$

Risque de 5% Fisher (1,n-2) →

EXCEL : INVERSE.LOI.F.N(0.95;1;n-2)

OU

INVERSE.LOI.F.DROITE(0.05;1;n-2)

Loi de F_{obs} sous H_0

La statistique F_{obs} suit la loi de Fisher à $(1, n-2)$ ddl

Plus la valeur de F_{obs} est grande et plus elle est en faveur de H_1

Région critique au risque α : $F_{obs} > F_{1-\alpha}(1, n-2) \Rightarrow$ ***Rejet de H_0 au seuil de α***

Liaison entre deux variables quantitatives

Arts et Métiers Sciences et Technologies

Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A :	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B :	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression. Calculer le coefficient de détermination. Commenter.

2. Deux stagiaires semblent se distinguer des autres. Les supprimer et déterminer la droite de régression sur les dix points restants. Calculer le coefficient de détermination. Commenter.

Liaison entre deux variables quantitatives

Arts Sciences et Technologies et Métiers

Le tableau ci-dessous donne la production annuelle d'une usine de pâte à papier

2010	2011	2012	2013	2014	2015	2016	2017
325	351	382	432	478	538	708	930

1. *Tracer le nuage de points correspondant sous Excel*
2. *Un ajustement affine vous semble-t-il adéquat?*
3. *Pour chaque année, on note pi la production de la pâte à papier et mi=ln(pi). Tracer le nouveau nuage de points (i,mi) et calculer le coefficient de corrélation linéaire de la série double (i, mi). Qu'en pensez-vous?*
4. *Donner une équation de la droite d'ajustement par les moindres carrés de mi en i.*
5. *Tracer les résidus ei et commenter*
6. *Quelle production peut-on prévoir en 2024?*

Statistiques inférentielles

Arts Sciences et Technologies et Métiers

En statistique,
en général impossible d'étudier un caractère sur toute
une population de taille N élevée

On suppose que les paramètres du caractère étudié dans la population sont connus, on en déduit les propriétés sur les échantillons prélevés.

- échantillonnage de taille n : ensemble des échantillons (aléatoires) de taille n.

Loi d'échantillonnage des Moyennes

Population de taille N,

X variable aléatoire définissant le caractère étudié avec E(X) = m et σ(X) = σ.
Soient $X_1, \dots, X_n$ les n v.a. de même loi X correspondant à n épreuves indépendantes.

$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ associe à tout échantillon de taille n, la moyenne de cet échantillon.

Proposition si $n \geq 30 \rightarrow \bar{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$

Arts et Métiers Sciences et Technologies

Loi d'échantillonnage de la fréquence

Population de taille N, étude d'un caractère dont on suppose qu'un individu le possède avec la probabilité $\boldsymbol{p}$.

$\boldsymbol{S}$ la v.a. qui, à tout échantillon aléatoire prélevé avec remise, de taille $\boldsymbol{n}$, associe le nbre d'individus possédant le caractère dans l'échantillon. Alors $S \sim B(n; p)$.

Proposition

$$\boldsymbol{Si\ F = \frac{S}{n}(proportion\ du\ caractère\ dans\ l'échantillon)}$$

$$\boldsymbol{suit\ approximativement\ la\ loi\ Normal} \sim \boldsymbol{N\left(p, \frac{p(1-p)}{n}\right)}$$

Statistiques inférentielles

Arts et Métiers Sciences et Technologies

Estimation ponctuelle

Estimation d'une valeur caractéristique d'une v.a. de la population mère à partir de l'étude d'un échantillon d'effectif n

	Population mère	Echantillon
Effectif	N	n
Moyenne	m	$\overline{x}_e$
Ecart type	σ	s
Fréquence	ρ	f

Proposition

- **Meilleure estimation de m : $\overline{x}_e$**
- **Meilleure estimation de σ: $s\sqrt{\frac{n}{n-1}}$**
- **Meilleure estimation de ρ : f**

Limite de l'estimation ponctuelle : n'indique pas le risque de se tromper ⇒ détermination d'un intervalle contenant la valeur de la moyenne/de la fréquence avec un risque d'erreur décidé à l'avance

Intervalle de confiance de la moyenne

Intervalle de confiance de la moyenne m au seuil de risque α :

$$\text{intervalle } I \text{ t.q pour } 100 \times (1-\alpha)\%, \qquad m \in I$$

α : seuil de risque/risque d'erreur et 1−α : coeff de confiance.

Proposition

Intervalle de confiance de la moyenne m au seuil de risque α :

$$\left[\bar{x}_e - t\frac{\sigma}{\sqrt{n}} \; ; \bar{x}_e + t\frac{\sigma}{\sqrt{n}}\right]$$

avec t est le nombre tel que $P(-t \leq Z \leq t) = 1-\alpha$
où Z suit la loi normale N (0; 1).

Arts et Métiers Sciences et Technologies

Intervalle de confiance de la moyenne

Proposition

Intervalle de confiance de la fréquence p au seuil de risque α :

$$\left[f - t\sqrt{\frac{f(1-f)}{n-1}} \; ; f + t\sqrt{\frac{f(1-f)}{n-1}} \right]$$