

# ♥ Introduction aux modèles multivariés ♥

## 1) Rappels

---

### A) Définitions

LA STATISTIQUE : c'est une méthode qui consiste à observer et étudier une ou plusieurs propriétés communes chez un groupe d'être, de choses ou d'entités.

UNE STATISTIQUE : c'est un nombre calculé à partir d'une population (d'êtres, de choses ou d'entités).

POPULATION : c'est une collection (d'être, de choses, ou d'entités) ayant des propriétés communes. Ce terme est hérité d'une des premières applications de la statistique : la démographie

Ex : un ensemble de parcelles de terrain étudiées, une population d'animaux, un groupe de patients présentant une maladie définie, l'ensemble des plantes d'une espèce donnée, une population d'humains habitants un lieu particulier...

INDIVIDU : c'est un élément de la population.

Ex : un patient, un insecte, une plante...

VARIABLE : c'est une des propriétés communes aux individus que l'on souhait étudier.

Elle peut être :

- Qualitative :

Ex : appréciation de la parcelle, l'état de santé de l'insecte, couleur des pétales, appartenance religieuse

- Quantitative (=numérique) continue (= pouvant prendre n'importe quelle valeur réelle) :

Ex : le taux d'acidité du sol, la longueur de l'insecte, la longueur de la tige, l'indice de masse corporelle.

- Quantitative (= numérique) discrète (= dès qu'il y a un saut minimum obligatoire entre deux valeurs successives, ex : les nombres entiers) :

Ex : la somme (sur tous les jours) du nombre de vaches présentes sur la parcelle, l'âge de l'insecte (en jours), le nombre de pétales sur la fleur, le nombre d'année d'études (réussies) depuis la petite école

## B) 2 types de statistiques

Statistique descriptive	Statistique inférentielle
<p>Son but est de décrire, c'est-à-dire de résumer ou représenter par des statistiques les données disponibles quand elles sont nombreuses.</p>	<p>Les données sont considérées incomplètes et elle a pour but de tenter de retrouver l'information sur la population initiale. La prémisse est que chaque mesure est une variable aléatoire suivant la loi de probabilité de la population.</p>
<p><u>Questions types</u> : représentation graphique, paramètres de position et dispersion, divers questions liées aux grands jeux de données</p>	<p><u>Questions types</u> : estimation de paramètres, intervalles de confiance, tests d'hypothèses, modélisation (ex : régression linéaire)</p>

La statistique peut être :

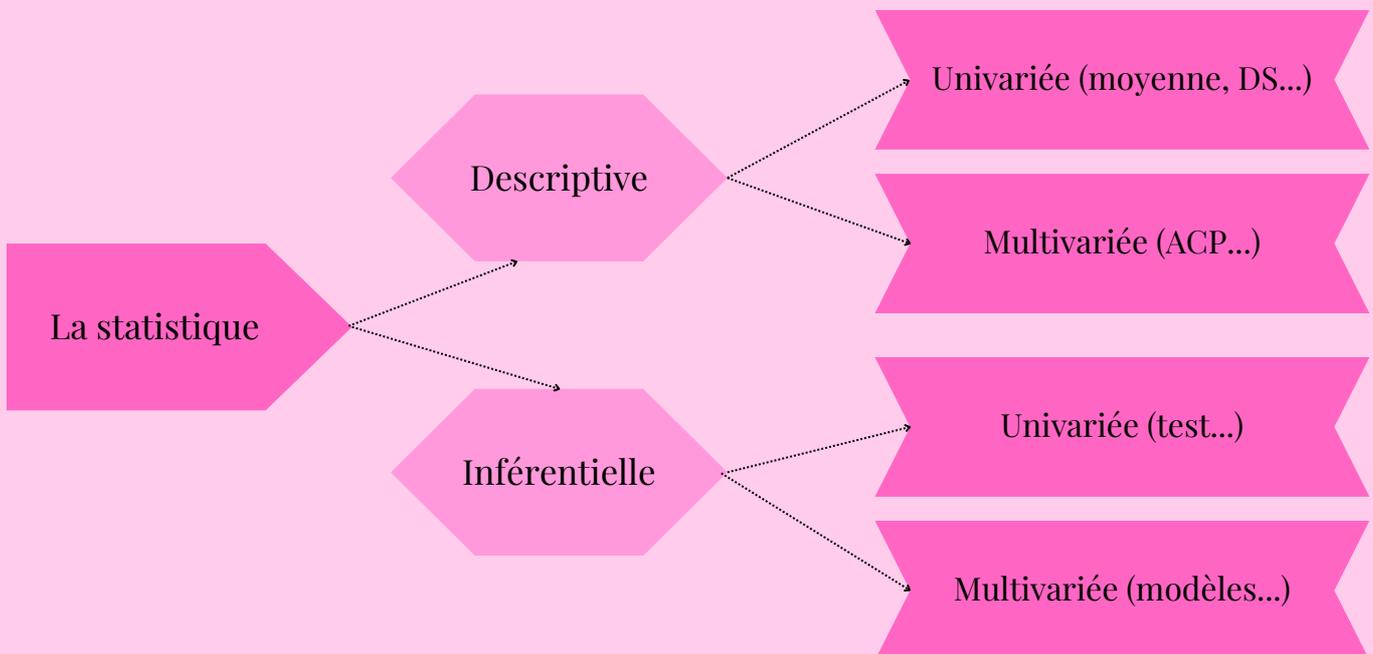
Univariée	Multivariée
Il n'y a qu'une seule variable qui rentre en jeu	Plusieurs variables rentrent en ligne de compte.
<u>Ex</u> : on regarde la proportion d'hommes et de femmes dans un échantillon	<u>Ex</u> : on étudie la variable taille et la variable sexe, on regarde s'il y a un lien entre le fait d'être un homme petit et une femme grande

o Deux variables entre elles = analyse bivariée

o Plusieurs variables = analyse multivariée

- Une variable expliquée
- Plusieurs variables explicatives indépendantes deux à deux

Schéma récap :



## 2) Régression linéaire simple

Point tut' : en statistique, la régression est une méthode permettant de proposer un **modèle mathématique** pour expliquer les relations entre les observations.

La **régression linéaire simple** consiste à proposer une droite pour expliquer une variable aléatoire **quantitative** par une autre.

Le **coefficient de corrélation linéaire** mesure la **liaison entre 2 variables aléatoires**. Les variables ont un rôle symétrique. Cependant, la question à résoudre peut être plus précise et libellée sous la forme suivante : « *Les valeurs prises par une variable  $Y$  dépendent-elles des valeurs de  $X$  ?* ».

Ici, les deux variables ne sont pas considérées de manière équivalente :

- **Y** (variable à expliquer, également appelée variable dépendante) est la variable dont on veut **expliquer** les valeurs
- **X** (variable explicative, également appelée variable indépendante) est la variable que l'on veut **utiliser** pour expliquer Y

La courbe qui décrit les variations de Y en fonction de X s'appelle **courbe de régression de Y en X**. On peut, en première approximation, chercher à assimiler cette courbe à une **droite**

### A) Exemple introductif

On étudie le lien entre la **taille** et **l'âge** des filles (en mois) sur un échantillon de **637 filles**.

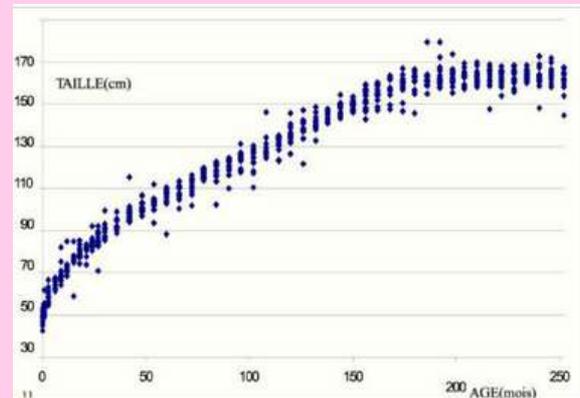
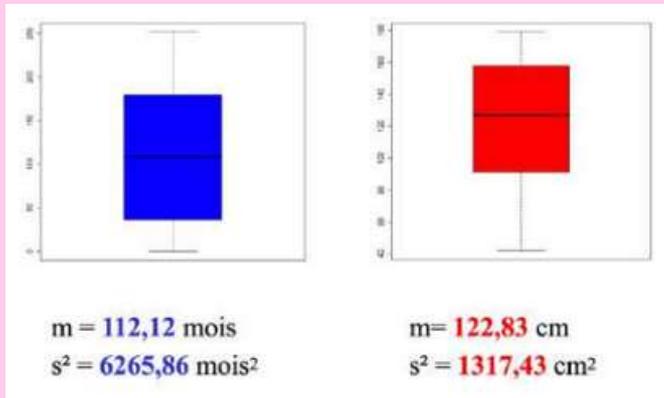
Questions que l'on se pose :

> Existe-t-il un lien entre la taille et l'âge ? S'il n'existe pas de lien, on obtiendra une droite **parallèle** à l'axe des abscisses (toute variation de X ne produit aucune variation de Y).

> Quand l'âge augmente, est-ce que la taille augmente aussi ?

> Connaissant l'âge, peut-on prédire la taille ? On peut chercher à estimer les zones sans valeur.

- o On peut y voir un but médical : par exemple la détection des retards de croissance.
- o Autre exemple : cela peut permettre aux médecins légistes qui retrouvent un os humain (complet ou fragment) dans la nature, de déterminer l'âge et le sexe.



## Comment la taille évolue-t-elle en fonction de l'âge ?

*Ici, quand on a une variation d'âge (X), on a une variation de taille (Y)*

> Taille =  $f(\hat{\text{âge}})$  *On va essayer de trouver la taille comme une fonction de l'âge ( $Y=aX+b$ ) pour estimer.*

Autrement dit, pour une variation de X, quelle est la variation de Y ?

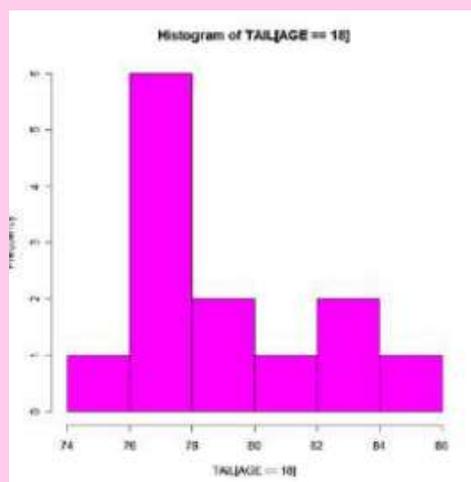
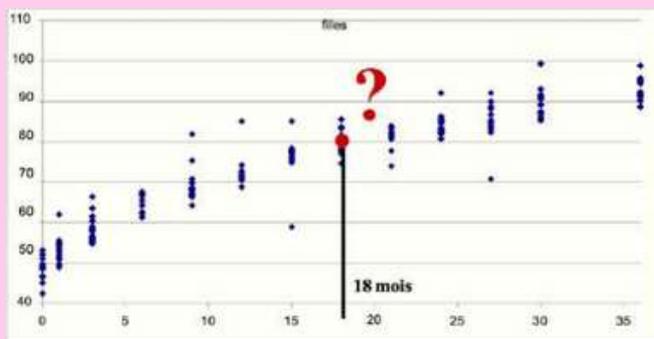
> On parle de **régression de Y en X** :

- Y = taille (cm)

- X = âge(mois)

> On cherche donc à savoir comment évolue la **taille** en fonction de **l'âge** pour chaque valeur d'âge (équation), ou bien encore, quelle est la taille pour un âge donné (valeur et intervalle de confiance).

**Exemple au sein d'un groupe de filles : Chez les filles de 18 mois, on va chercher la taille moyenne, la variance de la taille et la distribution.**



### Méthode pour déterminer l'âge à 18 mois :

- o On stratifie les données.
- o On sélectionne les filles de 18 mois.
- o On calcule les paramètres de la distribution (moyenne et variance), si tant qu'elle soit gaussienne.
- o On calcule un intervalle de confiance à 95% de la moyenne.

### Résultats :

- Moyenne observée =  $M(T/A=18) = 79,23\text{cm}$
- Variance observée =  $V(T/A=18) = 9,36\text{cm}^2$
- > On parle d'une distribution conditionnelle = valeur de la taille sachant l'âge (=T/A)

### Point tut' :

Finalement à nous de choisir le modèle qui convient le mieux. Ici la fonction qui va nous permettre d'avoir la moyenne des **tailles**, connaissant **l'âge**.

Dans T/A (taille sachant l'âge), il ne faut pas y voir une probabilité conditionnelle mais simplement le fait d'estimer la valeur de la taille Y pour une valeur de l'âge X.

On choisit le **modèle de la droite affine**, celle-ci passe les couples de points (X,Y).

## B) Fonction de régression

La taille en fonction de l'âge, également écrit  $\text{Moyenne}(\text{taille}/\hat{\text{age}}) = f(\hat{\text{age}})$ , peut s'exprimer par une **fonction f qui est une droite affine de type**

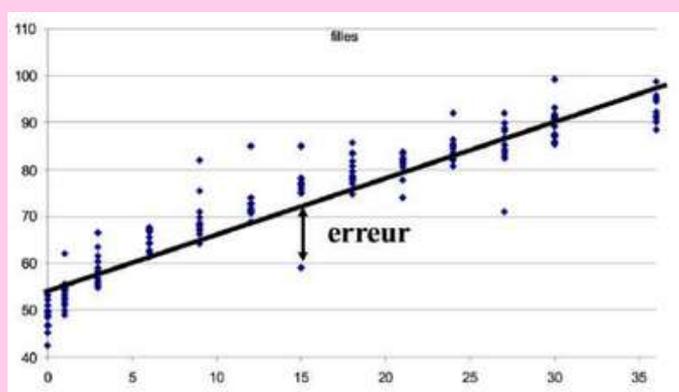
$$y = ax + b$$

On note aussi :

$$\text{Esperance (Taille/Age)} = \alpha + \beta \times \text{Age}$$

Pour chaque sujet, on définit la taille par  $\alpha + \beta \times \text{Age} + \varepsilon$ , avec  $\varepsilon$  qui représente l'erreur individuelle.

L'erreur individuelle ( $\varepsilon$ ) est l'écart entre la valeur obtenue par la fonction ( $y = ax + b$ ) et la vraie valeur observée



### Point tut' :

S'il n'y a pas de lien entre **X et Y** (pas de corrélation), alors **toute variation de X** n'entraîne **aucune variation de Y**.

On obtient donc une droite parallèle à l'axe des abscisses d'équation  $y = \text{constante}$ .

Ainsi, dans  $y = ax + b$ , on a :  $a = 0$ .

Pour chaque individu, par rapport à la moyenne, l'erreur est tant positive que négative, pour **minimiser ces écarts** et s'affranchir du signe, il faut les passer au carré. On va faire la somme des carrés des écarts (**SCE**).

La **REGRESSION LINEAIRE** est le modèle le plus simple pour permettre :

- une interprétation (lien ou non entre les deux variables), permise par la valeur du coefficient de régression qui englobe dans son calcul la pente de la droite, donc la valeur de  $\beta$
- une estimation de  $\alpha$  et  $\beta$  pour que la droite d'ajustement minimise l'erreur individuelle
- la prédiction et l'extrapolation

La **DROITE D'AJUSTEMENT** est aussi appelée droite de régression. On dit qu'elle permet de résumer au mieux le nuage de points.

Point tut' : La régression c'est prouver que l'une des deux variables permet de prédire l'autre, c'est-à-dire montrer qu'à partir de X on peut prédire Y.

On essaie alors de trouver les valeurs de la droite d'équation :  $Y = \alpha + \beta X + \varepsilon$

- Y la variable à expliquer
- X la variable explicative
- $\alpha$  l'ordonnée à l'origine (c'est la valeur de Y pour X=0)
- $\beta$  la pente (c'est la variation moyenne de la valeur de Y pour une augmentation d'une unité de X)
- $\varepsilon$  l'erreur aléatoire

## C) Principe de l'estimation

On veut estimer  $\alpha$  et  $\beta$  tel que  $\varepsilon$  soit le plus petit possible,  $\varepsilon_i$  représente l'écart entre la droite et le point i.

Pour chaque valeur de X, on a  $y_i = \alpha + \beta x_i + \varepsilon$

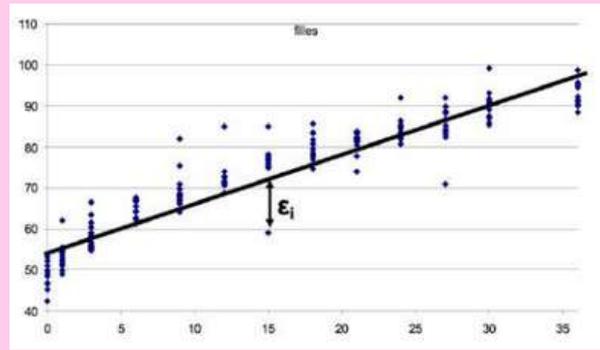
> Or,  $E(Y/X) = \alpha + \beta X$

> Donc  $\varepsilon_i = y_i - E(Y/X)$

> On calcule la somme des carrés des écarts :

$$SCE = \sum_{i=1}^n (\varepsilon_i)^2$$

On cherche à estimer  $\alpha$  et  $\beta$  tel que la SCE soit **la plus petite possible**



Point tut' : La **distance d'un point à la droite** est la distance verticale entre l'ordonnée du point observé et l'ordonnée du point correspondant sur la droite.

Cette distance d'un point à la droite représente **l'erreur  $\epsilon$** .

Pour s'affranchir du signe de l'erreur  $\epsilon$ , on calcule la **somme des carrés des distances de chaque point à la droite (SCE)**.

La droite de régression est alors **la droite qui minimise la somme des carrés des écarts** (donc c'est la droite qui passe le plus proche de chaque point du nuage).

Estimation de la pente :

$$\beta = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

o La covariance :  $\text{cov}(X,Y)$  = covariance de X et de Y. La covariance indique dans quelles mesures deux variables varient ensemble.

o La variance :  $\text{var}(X)$  = variance de X. La variance est une mesure de la dispersion des valeurs d'un échantillon.

Dans l'exemple, l'estimation de beta =

$$\beta = \frac{\text{cov}(TAIL,AGE)}{\text{var}(AGE)} = 0,437703$$

Estimation de l'ordonnée à l'origine  $\alpha$  :

La droite passe par  $mY$  et  $mX$

On a :  $mY = \alpha + \beta mX$

Donc  $\alpha = mY - \beta mX$

Dans l'exemple,  $\alpha = 73,729$

L'équation finale s'écrit donc :

$$Y = \alpha + \beta X + \varepsilon$$

Dans l'exemple, Taille = 73,73 + 0,44 Age +  $\varepsilon$  ou  $E(\text{Taille}/\text{Age}) = 73,73 + 0,44 \text{ Age}$

Point tut' :

Une particularité de la **droite de régression** est de passer par le point moyen théorique de coordonnées  $(m_x ; m_y)$ , où

- $m_x$  est la moyenne empirique de X et
- $m_y$  est la moyenne empirique de Y sur l'échantillon.

L'estimation de l'**ordonnée à l'origine  $\alpha$**  est déduit de la **penste  $\beta$**  et des coordonnées du point moyen  $(m_X ; m_Y)$  par la formule suivante :

$$\alpha = m_Y - \beta m_X$$

#### 4) Interprétation

- De la pente  $\beta$  :

- >  $\beta = 0$  : pas de lien, évolutions indépendantes
- >  $\beta < 0$  : évolution en sens contraire
- >  $\beta > 0$  : évolution dans le même sens

- De l'ordonnée à l'origine :

- >  $E(Y/X=0) = \alpha$

- Test de la pente à 0 : si  $\beta = 0$ , alors il n'y a pas de lien entre Y et X.

- > Le lien entre Y et X est-il significatif ? Autrement dit, est-ce que  $\beta \neq 0$  ?

Soit  $b$  une estimation de  $\beta$ , la fluctuation de  $b$  observée peut être due au hasard.

On note les hypothèses :

- $H_0$  :  $\beta = 0$ , il n'y a pas de lien entre X et Y
- $H_1$  :  $\beta \neq 0$ , il existe un lien entre X et Y

Sous  $H_0$ , et si les conditions d'application sont respectées, on a une statistique :

$$t_0 = \frac{b - \beta}{\sqrt{s_b^2}}$$

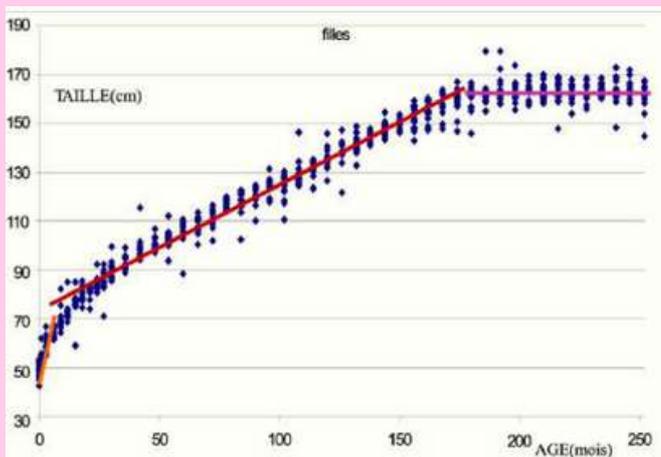
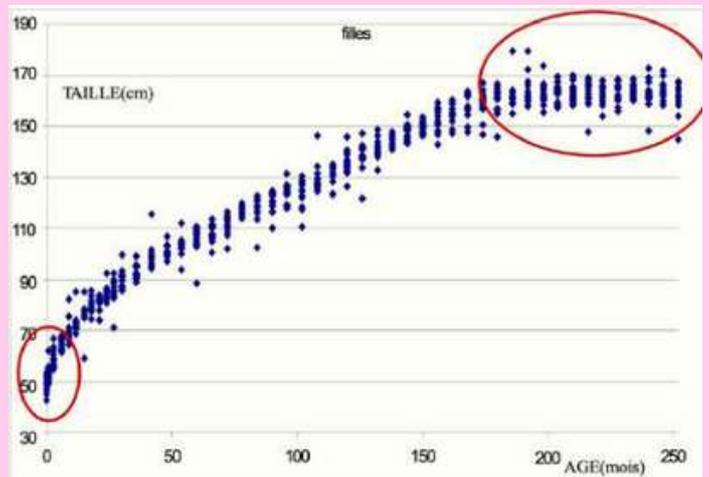
qui suit une loi de Student à  $n-2$  DDL, avec :

- >  $L(Y/X)$  qui tend vers  $N$
- >  $V(Y/X)$  constantes pour tout  $X$
- > à  $X$  donné,  $Y_i$  indépendants

### La régression est linéaire

Point tut' :

Sur le graphique, à 0 mois (naissance), on a un regroupement de points, il semble ne pas y avoir de lien entre taille et âge. Après 200 mois, on peut tracer une droite parallèle à l'axe des abscisses



Courage c'est bientôt la fin ;)

Point tut' :

On veut appliquer un test statistique qui est le **test de la pente de la droite de régression**. La droite de régression d'équation  $Y = \alpha + \beta X$  comporte 2 paramètres ( $\alpha$  et  $\beta$ ).

- **L'hypothèse nulle  $H_0$**  est que la pente  $\beta$  de la droite de régression de Y en X est égale à 0, c'est-à-dire que Y est égal à  $\alpha$ , ou encore que la droite de régression est horizontale et qu'il n'y a pas de liaison entre X et Y.
- **L'hypothèse alternative  $H_1$**  est que la pente  $\beta$  de la droite est différente de 0.

**Sous  $H_0$** , le rapport de l'estimateur de la pente b sur son écart-type suit **une loi de Student à (n-2) DDL**, où n est l'effectif de l'échantillon.

Le test de la pente consiste à calculer la grandeur  $t_0$  et à la comparer à la valeur seuil  $t_{\alpha}$  sur la table de la loi de Student à (n-2) DDL

Voilaaaaa c'est fini pour ma deuxième (et dernière ) fiche de biostat sniff, mais je continue de vous embêter en SE mouahaha

Grosse dédi à vous qui venez (ou pas) de finir ce cours pas évident, je vous rassure les QCM sont en général assez similaires et ya pas besoin d'apprendre les démos et les exemples par coeur :)

Je sais que cette année peut vous paraître insurmontable mais je crois fort en vous, vous êtes capables à 100% de réussir ayez confiance, ça va le faire !