INTRODUCTION AUX MODÈLES MULTIVARIÉS

I. RAPPELS +++

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

Une **POPULATION** 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…*

Un **INDIVIDU** est un **élément de la population**.

> *Ex : un patient, un insecte, une plante…*

Une **VARIABLE** 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.

Il existe 2 directions en statistique :

Statistique descriptive : son but est de **décrire**, c'est-à-dire de résumer ou représenter par des statistiques les **données disponibles** quand elles sont nombreuses.

> Questions types : représentation graphique, paramètres de position et dispersion, divers questions liées aux grands jeux de données.

Statistique inférentielle : 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.

> Questions types : estimation de paramètres, intervalles de confiance, tests d'hypothèses, modélisation (ex : régression linéaire).

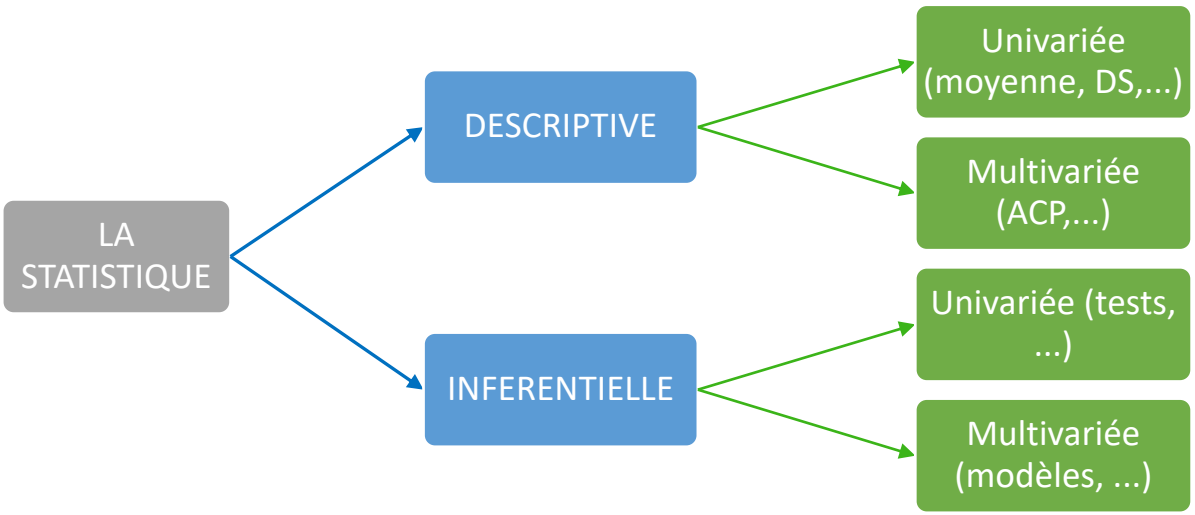
La statistique peut être :

UNIVARIÉE = il n'y a qu'une seule variable qui rentre en jeu. Ex : on regarde la proportion d'hommes et de femmes dans un échantillon.

MULTIVARIÉE = plusieurs variables rentrent en ligne de compte. Ex : 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.

- Deux variables entre elles = analyse **bivariée**
- Plusieurs variables = analyse **multivariée**
  - Une variable expliquée
  - Plusieurs variables explicatives indépendantes deux à deux

Schéma récap :



II. LA RÉGRESSION LINÉRAIRE 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** :

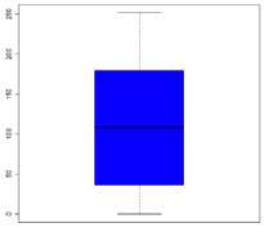
- **Y** *(variable à expliquer, également appelée variable dépendante)* est la **variable dont on veut expliquer les valeurs**
- **X** *(variable explicative, également appelée variable indépendante)* est la **variable que l'on veut utiliser pour expliquer Y**

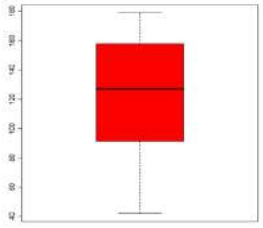
La courbe qui décrit les variations de Y en fonction de X s'appelle **courbe de régression de Y en X**. On peut, en première approximation, chercher à assimiler cette courbe à une droite

1. 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 :

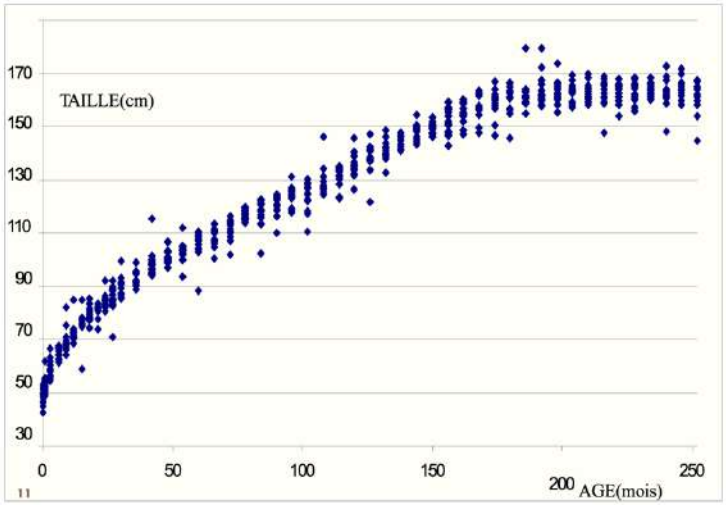
- Existe-t-il un lien entre la **taille** et **l'âge** ? S'il n'existe pas de lien, on obtiendra une droite parallèle à l'axe des abscisses (toute variation de X ne produit aucune variation de Y).
- Quand l'âge augmente, est-ce que la taille augmente aussi ?
- Connaissant l'âge, peut-on prédire la taille ? On peut chercher à estimer les zones sans valeur.
  - On peut y voir un but médical : par exemple la détection des retards de croissance.
  - 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.





m = 112,12 mois
s^2 = 6265,86 mois²

m= 122,83 cm
s^2 = 1317,43 cm²

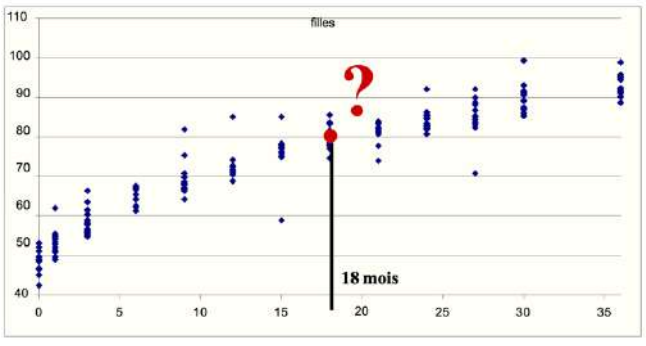


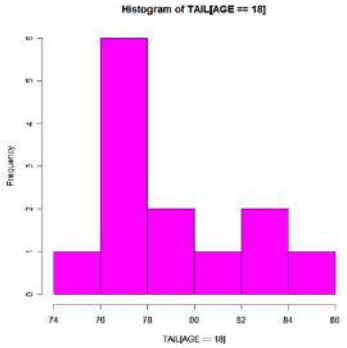
Comment la taille évolue-t-elle en fonction de l'âge ?

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

- Taille = f(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 :

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

Résultats :

- Moyenne observée = M(T/A=18) = 79,23cm
- Variance observée = V(T/A=18) = 9,36cm^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).

2. La fonction de régression

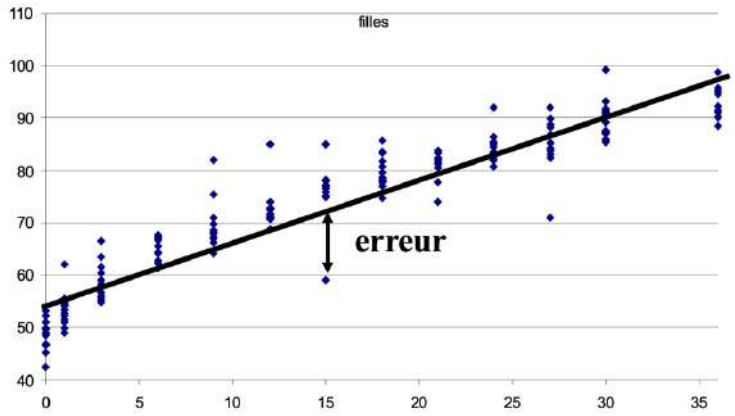
La taille en fonction de l'âge, également écrit Moyenne(taille/âge) = f(âge), peut s'exprimer par une **fonction f qui est une droite affine de type** $y = ax + b$.

On note aussi :

Esperance (Taille/Age) = α + β × Age.

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

L'erreur individuelle (ε) 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'entraine aucune variation de Y. On obtient donc une **droite parallèle à l'axe des abscisses** d'équation **y = constante**. Ainsi, dans $y = ax + b$, on a : $\boldsymbol{a = 0}$.

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 **régression linéaire** 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 ß*
- une **estimation de α et ß** 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** $Y = \alpha + \beta X + \varepsilon$, avec :

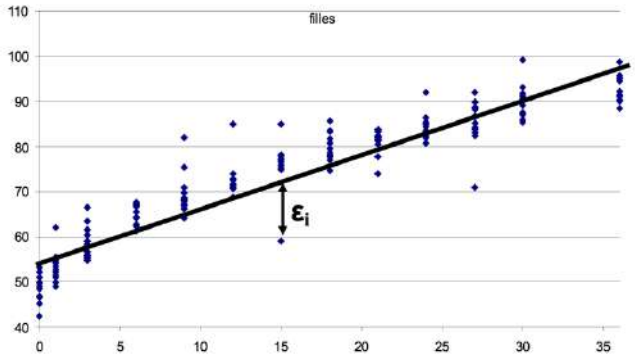
- **Y** la variable à **expliquer**
- **X** la variable **explicative**
- **α l'ordonnée à l'origine** *(c'est la valeur de Y pour X=0)*
- **β** la **pente** *(c'est la variation moyenne de la valeur de Y pour une augmentation d'une unité de X)*
- **ε l'erreur aléatoire**

3. Principe de l'estimation

On veut estimer α et ß tel que ε soit le plus petit possible. **ε_i** représente **l'écart entre la droite et le point i**.

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

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 α et β 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 ε**.

Pour **s'affranchir du signe de l'erreur ε**, 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** $\boldsymbol{\beta = \frac{cov(X,Y)}{var(X)}}$ avec :
  - **La covariance :** cov(X,Y) = covariance de X et de Y. *La **covariance** indique dans quelles mesures deux variables varient ensemble.*
  - **La variance :** 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,* $\beta = \frac{cov(TAIL,AGE)}{var(AGE)} = 0{,}437703$

- **Estimation de l'ordonnée à l'origine α :**

  La droite passe par mY et mX.

  On a $\boldsymbol{mY = \alpha + \beta mX}$, donc $\boldsymbol{\alpha = mY - \beta mX.}$

⇨ *Dans l'exemple,* $\alpha = 73{,}729$

- **L'équation finale s'écrit donc :**

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon,} \text{ ou } \boldsymbol{E(Y/X) = \alpha + \beta X}$$

⇨ *Dans l'exemple, Taille = 73,73 + 0,44 Age + ε ou E(Taille/Age) = 73,73 + 0,44 Age*

Point tut' :

Une **particularité de la droite de régression** est de passer par le point moyen théorique de coordonnées (mx ; my), où mx est la moyenne empirique de X et mY est la moyenne empirique de Y sur l'échantillon.

L'estimation de l'ordonnée à l'origine α est déduit de la pente β et des coordonnées du point moyen (mx ; my) par la formule suivante : $\boldsymbol{\alpha = mY - \beta mX}$

4. Interprétation

De la **pente β** :

- ⇨ **β = 0** : **pas de lien**, évolutions indépendantes
- ⇨ **β < 0** : **évolution** en **sens contraire**
- ⇨ **β > 0** : **évolution** dans le **même sens**

De **l'ordonnée à l'origine** :

- ⇨ **E (Y/X=0) = α**

Test de la pente à 0 : si $\boldsymbol{\beta = 0}$, 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 β, la fluctuation de b observée peut être due au hasard.

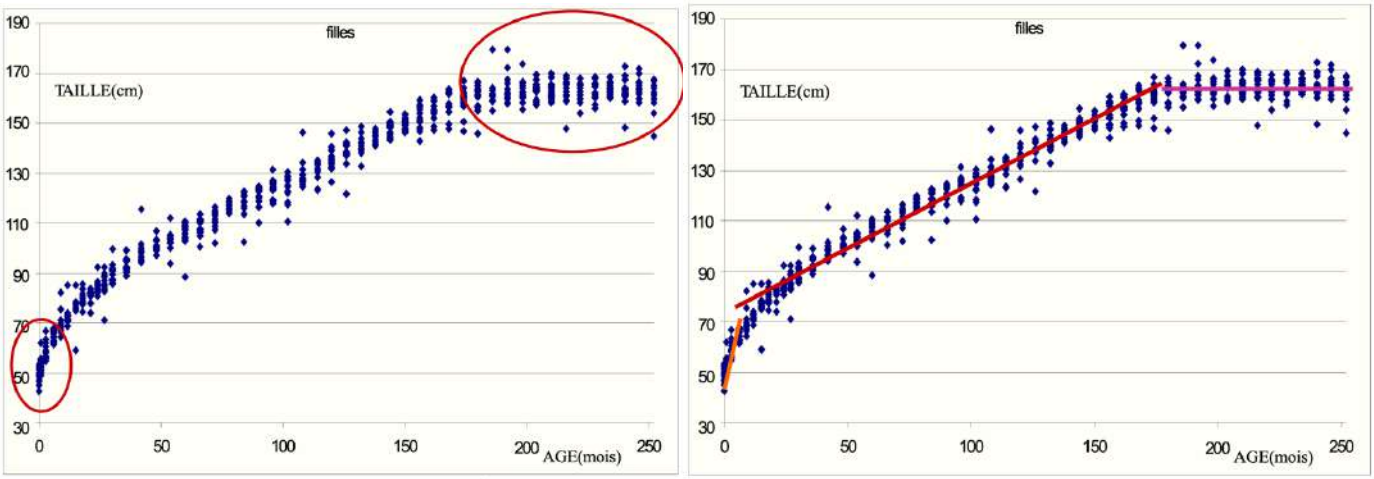
On note les *hypothèses* :

- **H_0 : $\boldsymbol{\beta = 0}$**, il n'y a **pas de lien** entre X et Y
- **H_1 : $\boldsymbol{\beta \neq 0}$**, 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, Yi indépendants

La **régression est linéaire**.



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** (α et β).

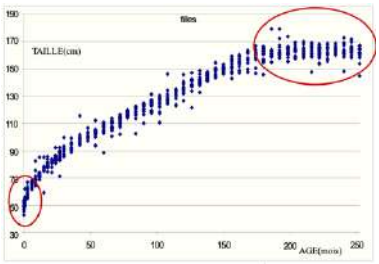
L'hypothèse nulle H_0 est que la pente β de la droite de régression de Y en X est égale à 0, c'est-à-dire que Y est égal à α, ou encore que la droite de régression est horizontale et qu'il n'y a pas de liaison entre X et Y.

L'hypothèse alternative H_1 est que la pente β 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_α** sur la table de la loi de Student à (n-2) DDL

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.



FIN

Cette année, le professeur ne présente pas le cours dans son intégralité, seul le modèle de la **régression linéaire simple** est à connaître. Cela correspond aux 27 premières diapos (14 et 22 sautées, mais j'ai conservé dans la fiche).

Ce cours peut paraître complexe aux premiers abords, les points tut' sont là pour vous aider, si ça n'aide pas on passe. N'hésitez pas à poser des questions sur le **forum** ou pendant les permanences Discord.

Ce n'est pas toujours évident mais **TRAVAILLEZ LA BIOSTAT** ! Tout n'est pas simple mais une grande partie des QRUs sont abordables et assez répétitifs. Il y a des points faciles à prendre pour **l'examen**, ça peut faire la différence.

Courage.

