INTRODUCTION AUX MODELES MULTIVARIES

Disclaimer : tous les « Points tut' » sont des aides fournies par vos vieux de biostat pour comprendre le cours, ça ne fait pas partie du diapo du prof c'est tout bonus ♥

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'êtres, 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 habitant dans 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 souhaite 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 : taux d'acidité du sol, longueur de l'insecte, longueur de la tige, indice de masse corporelle.
  - **Quantitative** *(= numérique)* discrète *(= dès qu'il y a un saut minimum obligatoire entre deux valeurs successives, ex : 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'années d'études (réussies) depuis la petite école.

Il existe 2 directions en statistique :

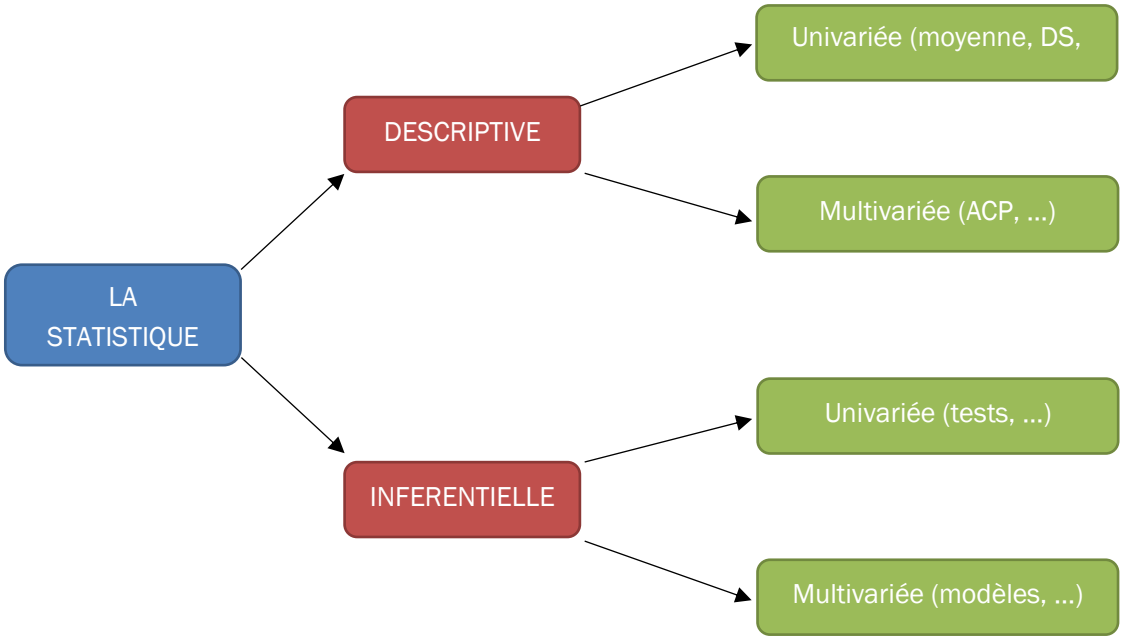
STATISTIQUE DESCRIPTIVE	son but est de **décrire**, c'est-à-dire de résumer ou représenter par des statistiques les **données disponibles** quand elles sont nombreuses.	*Questions types :* représentation graphique, paramètres de position et dispersion, divers questions liées aux grands jeux de données.

STATISTIQUE INFERENTIELLE	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 :* estimations de paramètres, intervalles de confiance, tests d'hypothèses, modélisation *(ex : régression linéaire)*.

La statistique peut être :

- **UNIVARIEE** = il n'y a **qu'une seule variable** qui rentre en jeu.
- **MULTIVARIEE** = **plusieurs variables** rentrent en ligne de compte.
  - 2 variables entre elles = **analyse bivariée**
  - Plusieurs variables = **analyse multivariée**
    - une variable *expliquée*
    - *plusieurs variables explicatives indépendantes deux à deux*

RECAP :



II. 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**

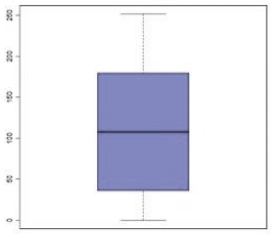
- La courbe qui décrit les variations de Y en fonction de X s'appelle **courbe de régression de Y en X**. On peut, en première approximation, chercher à assimiler cette courbe à une droite

A. Régression linéaire

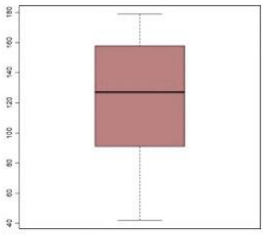
1) Exemple introductif

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 :

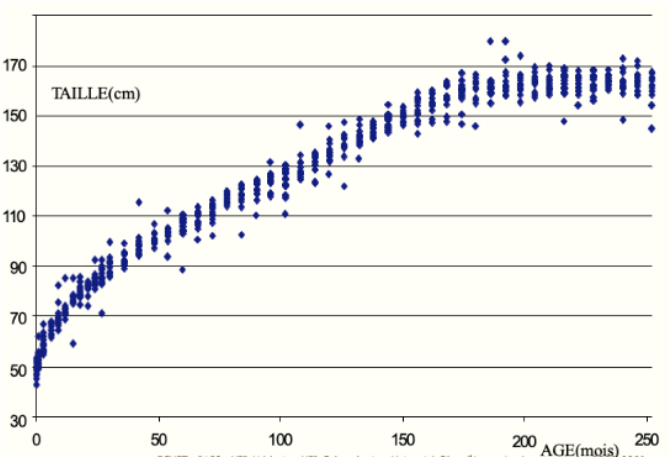
- Existe-t-il un lien entre la taille et l'âge ?
- Quand l'âge augmente, est-ce que la taille augmente aussi ?
- Connaissant l'âge, peut-on prédire la taille ?
  - *On peut y voir un but médical, par exemple : détecter les retards de croissance.*
  - *Autre exemple : cela peut permettre aux médecins 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



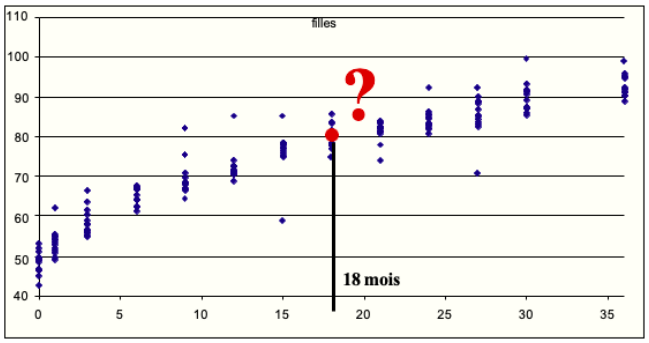
m= 122,83 cm
s^2 = 1317,43 cm²

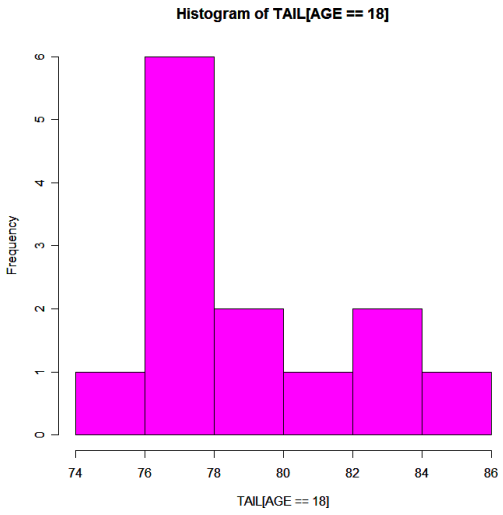


Comment la taille évolue-t-elle en fonction de l'âge ?

- Taille = f(âge) → Autrement dit, pour une variation de Y, quelle est la variation de X ?
- 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) = 9,36cm2

On parle d'une distribution conditionnelle = valeur de la taille sachant l'âge (= T/A).

2) 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 : **Espérance (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)	**l'écart** entre la valeur obtenue par la fonction (y=ax+b) et la vraie valeur observée.

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**, cad 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** *(cad que 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. **εi** représente **l'écart entre la droite et le point i**.

Pour chaque valeur de X, on a $\boldsymbol{yi = \alpha + \beta xi + \varepsilon i}$.
Or, $\boldsymbol{E(Y/X) = \alpha + \beta X}$.
Donc $\boldsymbol{\varepsilon i = yi - E(Y/X)}$.

On calcule la **somme des carrés des écarts :**

$$\boldsymbol{SCE = \sum_{i=1}^{n} (\varepsilon i)^2.}$$

On cherche à estimer α et β tel que 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** $\boldsymbol{\beta = \frac{cov(XY)}{var(X)}}$ avec :

  cov(XY) = covariance de X et de Y → ***POINT TUT :*** *La* ***covariance*** *indique dans quelles mesures deux variables varient ensembles.*
  var(X) = variance de X
  Dans l'exemple, ß = 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, α = 73,729.

- L'équation finale s'écrit donc :

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

Dans notre exemple, on a 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 : **α = mY – βmX**

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 **ß=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 ß≠0 ?*

Soit b une estimation de ß, la fluctuation de b observée peut être due au hasard.

On note les *hypothèses* :

- **H0 : β =0**, il n'y a pas de lien entre X et Y
- **H1 : β ≠0**, il existe un lien entre X et Y

Sous H0, 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) constante pour tout X
- à X donné, on a un Yi indépendant
  - ⇨ 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 = α + βX comporte 2 paramètres** (α et β).

- **L'hypothèse nulle H0** est que la pente β de la droite de régression de Y en X est égale à 0, cad 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 H1** est que la pente β de la droite est différente de 0.

- **Sous H0**, 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 t0** et à la **comparer à la valeur seuil tα** sur la table de la loi de Student à (n-2) DDL

✓ *Le hasard explique-t-il la fluctuation de b ?*

Intervalle de confiance de la pente : b tend vers t_{n-2} et on a : $b \pm t_{n-2,\frac{\alpha}{2}} * \sqrt{s_b}^2$

Si l'intervalle de confiance à 95% de b ne contient pas la valeur 0, dans ce cas, **b est différent de 0 au risque d'erreur de 5%.**

Intervalle de confiance de la droite : **E(Y/X) = α + βX**, estimé par $m_{y/x} = \alpha + bX$

=> $m_{y/x} \pm t_{n-2,\frac{\alpha}{2}} * \sqrt{s_{my/x}}^2$

Intervalle de prédiction : pour un âge (X) fixé, on prédit la taille (Y) :

Yp = α + bX
$Taille_p$ = 73,73 + 0,44Age

Précision de la prédiction : $y_p \pm t_{n-2,\frac{\alpha}{2}} * \sqrt{s_{y_p}}^2$

✓ *On se pose la question de l'adéquation du modèle, c'est-à-dire, est ce que le modèle est un bon résumé des observations ?*

Pour cela, on va calculer le **pourcentage de variance expliquée R^2** :

$$R^2 = \frac{\textbf{part de la variance expliquée par la régression}}{\textbf{variance totale}} = \frac{\text{écart } (m_{y/x} - m_y)}{\text{écart } (y - m_y)}$$

Variance totale = S^2_Y

Pourcentage de variance expliquée :

$$R^2 = \frac{\Sigma(m_{y/x} - m_y)^2}{\Sigma(y - m_y)^2}$$

Exemple : R² = 88%

$\sqrt{R^2}$ = estimation du coefficient de corrélation entre X et Y

B. La régression logistique

On utilise ce modèle lorsque les conditions d'application de la régression linéaire ne sont pas remplies.

- Variable **à expliquer Y** = binaire (malade ou non).
- Variables **explicatives X** = quantitatives ou qualitatives.

$$Y = f(X1 ; X2 ; \dots ; Xn)$$

Expliquer Y revient à **quantifier l'association de Y pour chaque xi**, ou encore, **prédire Y à partir de nouvelles observations de xi**.

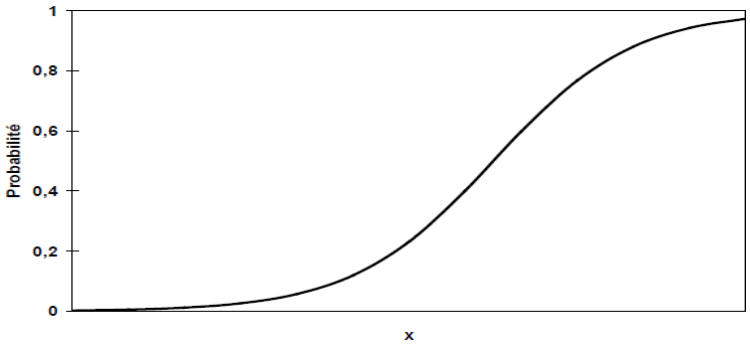
Exemple : Décès en fonction d'une dose de toxique : Comment varie la proportion de décès en fonction de la dose toxique ?

logit(p)=ln(p/1-p)=α+βX

L'estimation d'une probabilité est un rapport

Pour pouvoir transformer un rapport en somme, on passe par la fonction logarithme : **log(A/B) = logA- logB**.
La fonction logit donne le log népérien de la côte d'un évènement, cad le rapport p/1-p.

logit(p) = ln(p/(1-p))



$$p = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

	Chez les exposés	Chez les non-exposés
E	E=1	E=0
Probabilité d'être malade	$p_+ = p(M^+/E=1) = \frac{1}{1+e^{-(\alpha+\beta)}}$	$p_- = p(M^+/E=0) = \frac{1}{1+e^{-\alpha}}$
Probabilité de ne pas être malade	$1-p_+ = p(M^-/E=1) = \frac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}}$	$1-p_- = p(M^-/E=0) = \frac{e^{-\alpha}}{1+e^{-\alpha}}$

L'ODDS RATIO (ou OR) exprime **force du lien entre X et Y**, c'est le rapport des côtes. Il est déterminé à partir de l'estimation des paramètres.

$$OR = \frac{\frac{p_+}{(1-p_+)}}{\frac{p_-}{(1-p_-)}} = e^{\beta}$$

Conditions d'application de la régression logistique :

- Relation linéaire entre logit(p) et X
- Y binomial ou multinomial
- Codage « intelligent » des X catégoriels, afin de pouvoir interpréter les coefficients
- Indépendance des individus

Exemple : Facteurs d'hypotrophie à la naissance : Le poids de la mère est-il un facteur de risque d'hypotrophie ?

Logit(p) = α + β.POIDSMERE

OR = $e^{-0,03}$ = 0,97

IC à 95% de l'OR = [0.94 ; 0.99]

Interprétation : p<0,05, donc on conclut que l'OR est significativement différent de 1, et donc qu'il existe un lien significatif entre le poids de la mère et l'hypotrophie dans le sens suivant : lorsque le poids de la mère augmente, le risque d'hypotrophie diminue.

Pour chaque unité de poids maternel, le risque d'hypotrophie diminue de 0,97. On fait l'hypothèse d'un OR constant, quelque soit le poids maternel. Il s'agit d'une relation linéaire entre le risque d'hypotrophie et le poids maternel. Sinon => modification du codage du poids maternel.

III. Régression linéaire multiple

On peut trouver plusieurs causes dans l'évolution de la taille Y :

- **L'âge** (X1)
- Les **facteurs socio-économiques** (X2)
- Les **taux d'hormones de croissance** (X3)

Dans ce cas, on a :

$$E(Y / X1, X2, X3) = \alpha + \beta1\, X1 + \beta2\, X2 + \beta3\, X3$$

Estimation : α, β_1, β_2, β_3 sont estimés en tenant compte des 3 variables aléatoires X1, X2, X3.
On parle alors **d'ajustement**, et on peut envisager des **interactions** :

$$E(Y / X1, X2, X3) = \alpha + \beta1\, X1 + \beta2\, X2 + \beta3\, X3 + \beta4\, X2\, X3$$

- Tests des ß1, ß2, ß3 à 0
- Interprétation identique
- Adéquation identique
- Approche pas à pas
- Choix des variables : notion de modèle
- Variables très corrélées

Exemple : Prédire l'âge en fonction de 8 mesures : crâne (BIP), tronc (LATHO), membres supérieurs et inférieurs (LOMAIN, PERPOIGN, PERCHEV, PIEDS), globales (STAT, POIDS) sur un échantillon de 1000 enfants de 2 à 16 ans.

En moyenne, AGE = α + β_1xBIP + β_2xLATHO + β_3xLOMAIN+ β_4xPERPOIGN + β_5xPERCHEV + β_6xPIEDS + β_7xSTAT + β_8xPOIDS

Les statistiques descriptives nous indiquent que : mean(AGE) = 10,373 et var(AGE) = 11,53541.

Ensuite, on regarde les conditions d'application, les intervalles de confiance des paramètres, ainsi que l'adéquation (R^2).

A. Sélection des variables du modèle

La sélection des variables utiles au modèle se base sur le **principe de parcimonie** *(« Les multiples ne doivent pas être utilisés sans nécessité »)*.

De ce fait, on n'ajoutera pas de nouvelles variables tant que celles présentes suffisent.

C'est ce qu'on appelle la **balance entre l'explication et la prédiction** : si on se retrouve avec trop de variables, notre modèle sera mieux expliqué, mais perdra en prédiction.

On parle aussi **D'OVERFITTING**, ou **D'HYPERADEQUATION**.

Ainsi, la sélection de variables se fait **pas-à-pas** *(stepwise pour les bilingues)*.

- **Ascendant** = on ajoute les variables une à une
- **Descendant** = on retire les variables une à une
- **Double sens**

Le critère de sélection se base sur le calcul d'un **« score » AIC** (Akaike Information Criterion).

$$AIC = 2p - 2ln(L)$$

Avec : p le nombre de paramètres, et L la vraisemblance au modèle.

On veut le AIC le plus **petit** possible.

IV. Régression LOGISTIQUE multiple

L'hypotrophie à la naissance dépend-elle du tabagisme, de l'HTA, de l'âge maternel et du poids maternel ?

Dans ce cas de figure-là, il est nécessaire de faire attention aux **interactions** qu'il peut y avoir entre les variables, notamment ici entre l'HTA et le tabac et entre l'HTA et le poids.

On utilise **l'analyse univariée** grâce au test exact de Fisher, au test du Chi2 de Pearson, et au test t de Student pour l'HTA, le tabac, l'âge et le poids maternel.

Et on utilise des **tests d'interaction** (test exact de Fisher, test de Wilcoxon) pour l'étude des variables HTA.TABAC et HTA.POIDSMAT.

V. Méthodes particulières

- ♥ Données de comptage : **régression de Poisson** *(nombre d'évènements dans le temps)*
- ♥ **Régression non-linéaire**
- ♥ Données censurées (survie) :
  - ☞ Estimation de **Kaplan-Meïer** ou **analyse actuarielle**
  - ☞ Test du **Log-Rank** *(analyse univariée)* ou **modèle de Cox** *(analyse multivariée)*
- ♥ **Séries temporelles** (Box-Jenkins)
- ♥ Variabilité spatiale
- ♥ Analyse factorielle de données : **ACP**, **ACM**, **arbres**, **CHA**, **Kmeans**,…

A. Analyse en composantes principales (ACP)

Dans l'ACP, les variables sont **toutes quantitatives**.
Les *moyennes, variances et corrélations* ont un sens.

On va examiner la structure des données : *ressemblance entre les individus, existence de sous-groupes d'individus, aberrance d'individus.*

On cherche la **corrélation** entre les variables. Cela nous permet d'interpréter facilement la matrice de corrélation.
Si on a **p variables**, il existe :

$$p * (p + 1)/2 \text{ corrélations possibles}$$

Principe de l'ACP : si les données ne comportent que 2 variables, une **simple représentation graphique** suffit pour répondre aux objectifs.

Mais en général, il y a p variables *(on parle d'espace à p dimensions)* et la **représentation sous forme d'axes simples devient impossible**. L'idée est donc d'obtenir des **représentations approchées dans un espace en 2 dimensions**.

On estime qu'on a p variables, ce qui revient à parler d'une dimension p (Rp). Le but est d'obtenir des représentations en dimension 2 les plus fiables possibles.

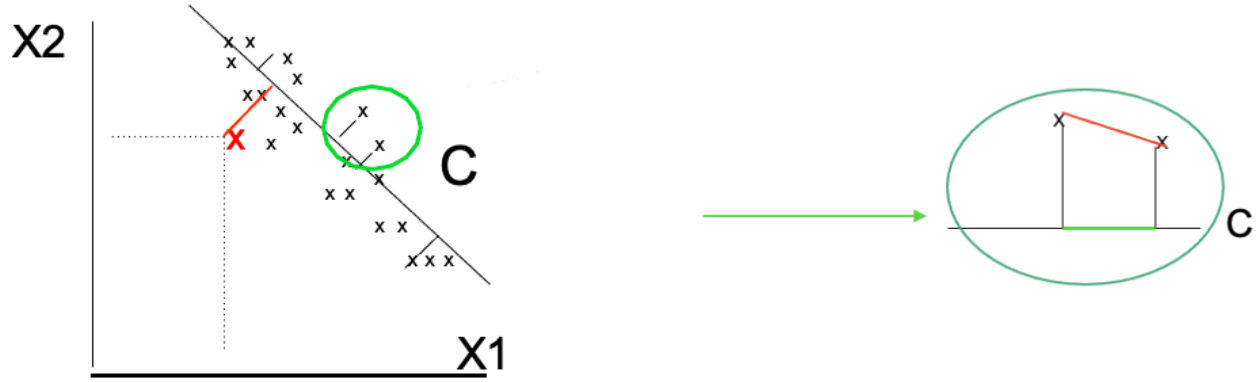
Le critère sur lequel on va se baser va être la **conservation de la variance**, c'est-à-dire qu'on souhaite **conserver la distance entre les individus** lorsqu'on va passer d'une représentation à l'autre.

Pour cela, on construit de **nouvelles variables Cj** qui vont permettre de **maximiser la variance**.

Il existe des contraintes de simplicité : on parle de **combinaisons linéaires des variables initiales**.

$$C1 = A11X1 + A12X2 + \ldots + A1pXp$$

Géométriquement, on a :



Si on considère la nouvelle variable C, l'information est reconstituée de la manière la plus fiable possible au sens de la variance.

- La **première composante principale C1** se définit par la **combinaison linéaire des variables initiales** maximisant la variance.
- La **deuxième composante principale** : **maximise la variance**, et est **non-corrélée à la première composante** *(principe de l'orthogonalité)*.
- Et ainsi de suite…

Au plus, on obtient p composantes principales.

En réalité, s'il existe une liaison entre les variables, **l'essentiel de l'information** *(=la variance)* **est contenu dans les premières composantes principales** *(en général, dans les 2 ou 3 premières composantes principales).*

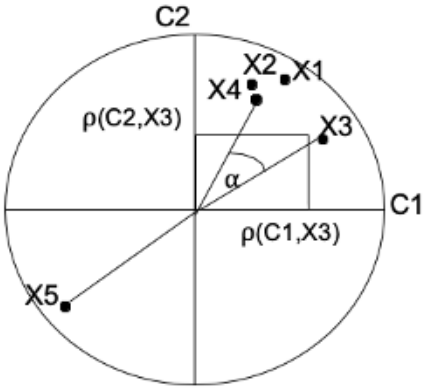
L'analyse des liaisons entre les variables permet d'obtenir une **matrice de corrélation**.

Avec p variables, on obtient **p*(p+1)/2 corrélations possibles**.

Les liaisons se font 2 à 2, il n'y a **pas de liaisons multivariées**.

En ACP, on va <u>représenter les variables</u> sous la forme d'un **cercle des corrélations** *(C1 et C2 étant les deux premières composantes principales).*

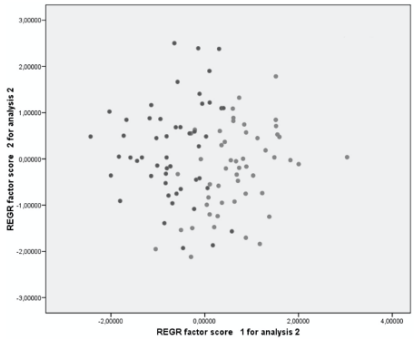
On peut alors montrer que <u>si des variables sont proches de la circonférence</u>, alors le **cosinus de l'angle α est proche du coefficient ρ de corrélation entre ces 2 variables**.



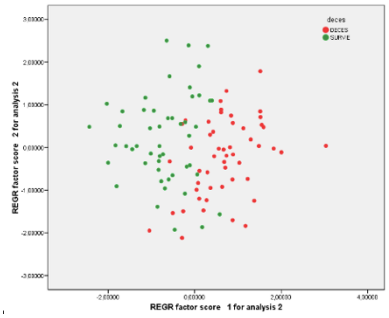
<u>Exemple d'ACP</u> : Infarctus du myocarde

- ☞ *Variables <u>numériques</u> : fréquence cardiaque, index cardiaque, index systolique, pression diastolique, pression artérielle pulmonaire, pression ventriculaire, résistance pulmonaire*
- ☞ *Variable <u>qualitative</u> : décès*

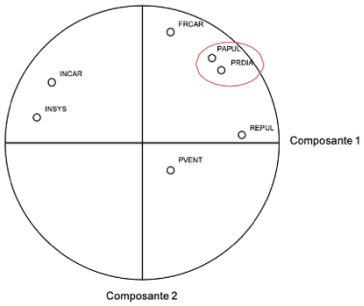
Ici, les objectifs vont être de vérifier la cohérence des données, rechercher les individus exceptionnels (en multivarié), rechercher l'existence de profils d'individus différents (sur p variables, donc en multivarié), et utiliser la variable « décès » comme variable illustrative.



Nuage des individus



Nuage des individus avec l'ajout d'une variable illustrative (vers l'inférentiel)



Cercle des corrélations entre les variables

VI. Stratégie d'analyse

STATISTIQUES DESCRIPTIVES	- Moyennes, pourcentages, intervallles de confiance, médianes - Graphiques (boxplot, histogrammes)
ANALYSES UNIVARIEES	- Descriptives : statistiques et graphiques par groupe, survie (Kaplan-Meier)
	- Tests statistiques (± séries appariées) : pourcentages (test du Chi-2, test exact de Fisher), moyennes (test t de Student, ANOVA, Wilcoxon, Krustal-Wallis), corrélation de Pearson ou de Spearman, LogRank (survie), interactions en fonction de la biologie, séries chronologiques, corrélations spatiales,…
ANALYSE MULTIVARIEE	- Choix de la méthode (R linéaire, R logistique, modèle de Cox,…)
	- Choix des variables initiales : variables connues dans la littérature, variables avec un sens biologique, variables p<0,2 ou p<0,25 pour les tests univariés
	- Méthode pas-à-pas, avec les interactions, choix du critère statistique - Garder les variables sélectionnées par la méthode pas-à-pas, et les variables biologiquement pertinentes
	- Vérification de la qualité du modèle - Interprétation du modèle final

FIN

Places aux dédicaces :

Toute l'équipe de biostat*royale* vous souhaite plein de courage sachant que vous entamez clairement la période la plus difficile du S1. Soyez malins, et surtout **BOSSEZ LA BIOSTAT !!!** car si vous pouvez vous servir des UE spé (abordables ++ à l'examen) c'est toujours des UE en + validées *(et on vous veut en P2 + que tout)*♥ ♥ ♥

On reste là pour vous pour des questions de cours, ou sur le chill pour des questions plus variées. N'hésitez pas à nous bombarder de questions s'il le faut on est là pour ça. Ce cours est loin d'être simple on le sait, donc si besoin on remontera vos questions aux professeurs.

Ne vous en faite pas, on vous prépare aussi des surprises pour bosser au mieux la biostat' c'est promis. Enfin, sachez que nous serons là pour vous à tout moment avant le concours, juste après (on viendra vous faire coucou quand même) et tout le reste de l'année même si y a plus de biostat'*(snif)*.

