Introduction

La statistique a envahi aujourd'hui tous les champs scientifiques. Les statistiques, dans le sens populaire du terme, **traitent des populations,** ce qui est très difficile dans le cas de la Bio-statistique. La statistique constitue, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence.

Le traitement A est-il plus efficace que le traitement B ?

Leur objectif consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. *On peut alors chercher à extrapoler une information obtenue à partir de l'échantillon.*

On trouve des applications de la statistique dans tous les domaines : industrie, environnement, médecine, finance, marketing, ….

Ce module se compose de trois chapitres :

1- Généralités et notions de base

2- Statistiques descriptives (**à une et à deux dimensions**)

3- Statistique inférentielles (tests de comparaisons : **tests paramétriques** et **non paramétriques**)

Objectifs :

- Connaitre le vocabulaire particulier de la statistique
- Comprendre les principes du traitement des données
- Le choix de la méthode statistique opportune à chaque situation particulière
- La réalisation des calculs et des tests de base pour une et deux variables
- La réalisation de ces calculs simplement et efficacement à l'aide de l'utilisation des logiciels statistiques (excel, R, **past**, **Systat**, XLSTAT, **SPSS ……**)

CHAPITRE I : Généralités et notions de base

La bio-statistique

Définition

Ensemble de méthodes à partir desquelles on recueille, organise, résume, présente et analyse des données afin d'en tirer des conclusions et de prendre des décisions avec prudence.

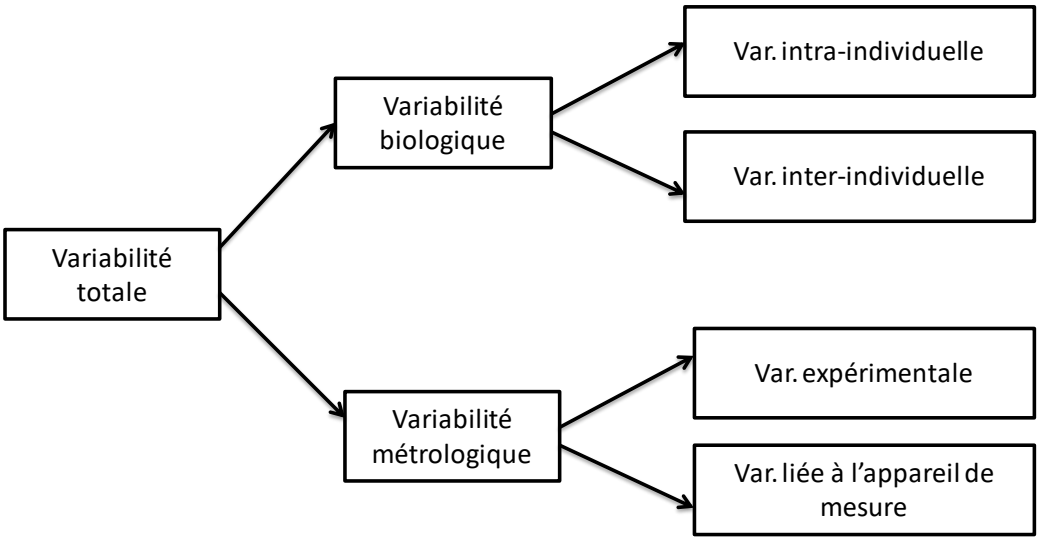
Notions importantes

Parmi les notions importantes nous avons :

La variabilité

Disposition à varié, qualité de ce qui est variable.

La variabilité en biologie est la somme d'une variabilité métrologique et d'une variabilité proprement biologique.



Population : ensemble des individus objets de l'étude, ou ensembles des **éléments** ou d'**individus** de même nature, visés par une problématique scientifique.

Elément : les éléments sont les unités qui composent une population.

Synonymes : Objet, individus, unité statistique, unité d'échantillonnage, sujet, événement, comportement……

Echantillon : C'est une sous ensemble de la population considérée, prélevé pour juger de cet ensemble.

Echantillon représentatif : échantillon qui reflète fidèlement la complexité et la composition de la population. **Le tirage au sor** ainsi que **l'inventaire exhausif (recensement)**, sont deux façons d'obtenir un échantillon représentatif d'une population.

Caractère statistique (ou variable statistique)

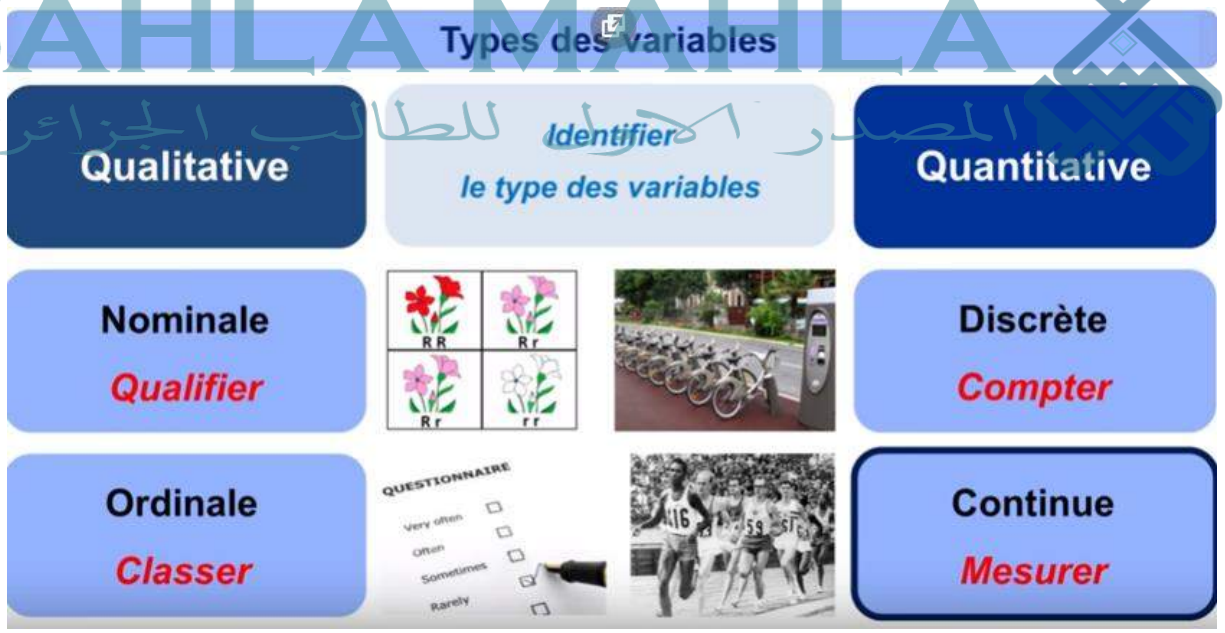
C’est ce qui est observé ou mesuré sur les individus d’une population statistique. Cette variable peut être **quantitative** (numérique) ou **qualitative** (non numérique).

Variable quantitative : c’est un paramètre expérimental qui s’exprime par un nombre.

Pouvant être classées en **variables continue** (taille, poids) ou **discontinue (discrète)** (nombre d’enfants dans une famille, nombre d’œufs pondus par un oiseau)

Variable qualitative : une variable qualitative se caractérise par un ensemble discontinu d’états.

Pouvant être classées en variables **catégorielles (nominales)** (couleurs des plumes des oiseaux) ou ordinales (degré des couleurs, résistance d’une plante vis-à-vis un ravageur classée e faible, moyenne, importante).



Notion d’hypothèse :

L’hypothèse est une relation hypothétique (provisoire, postulée par le chercheur).

On distingue deux formes d'hypothèses :

- **Hypothèse nulle (H0)** : $m_1 = m_2$ ou l'absence d'une différence significative entre les moyennes

- **Hypothèse alternative (H1) :** $m_1 \neq m_2$ ou l'existence d'une différence significative entre les moyennes

Seuil de signification :

En statistique, il n'existe pas de règle rigide permettant de tirer une conclusion concernant les hypothèses ; aucun test ne nous fournit une réponse en terme de oui ou non, mais indique dans quelle mesure nous pouvons être certain de tirer des conclusions ; cette mesure se nomme niveau ou seuil de signification, ou encore probabilité d'erreur.

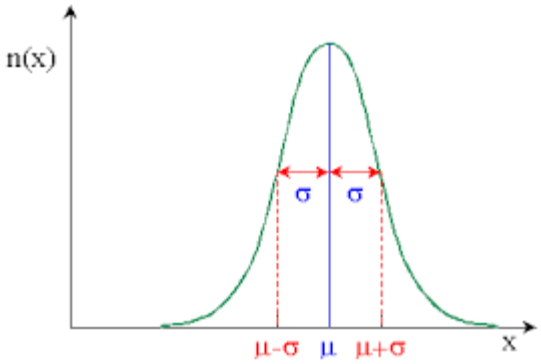
La loi normale (Normalité)

En théorie des probabilités et en statistique, la **loi normale** est l'une des lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Elle est également appelée **loi gaussienne.** En effet, elle correspond au comportement, sous certaines conditions, d'une suite d'expériences aléatoires similaires et indépendantes.

Lorsqu'une variable aléatoire *X* suit la loi normale, elle est dite ***gaussienne*** ou ***normale***

Une distribution normale correspond à la distribution de probabilité d'une variable aléatoire continue dont la courbe est parfaitement symétrique et en forme de cloche.

Les lois de probabilité permettent de décrire de manière théorique le caractère aléatoire d'une expérience qui est considérée comme aléatoire.



Ou plus précisément :

- Dans le cas ou les trois paramètres de position : **moyenne, médiane, mode** sont proche ou égaux.

- Si les données ne subissent pas la loi normale, les scientifiques ont trouvées une solution alternatives est de passer aux méthodes paramétriques.

Rappel :

- Pour augmenter la puissance de vos tests, clarifiez et précisez vos hypothèses grâce à un travail théorique AVANT de faire les expériences,
- Pour augmenter la puissance de vos tests, n'introduisez dans vos analyses QUE des variables dont la présence est JUISTIFIEE PAR LA THEORIE.

Types de test

On parle de **tests paramétriques** lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètre estimé sur l'échantillon moyenne, mode et médiane. La distribution des données suit la loi normale.

Les tests non paramétriques ne font aucune hypothèses sur la distribution sous-jacente des données (la distribution des données ne suit pas la loi normale).

Méthodes statistiques

<table>
<tr><th colspan="4">Méthodes descriptives</th><th colspan="2">Méthodes inférentielles</th></tr>
<tr><td colspan="4">↓
Objectif : recherche des liaisons possibles</td><td colspan="2">↓
Objectif : recherche des différences (présence de signification ou absence)

(tester la normalité tout d’abord pour choisir entre les 2)</td></tr>
<tr><td colspan="2">Descriptive univariée (à une dimension)</td><td>Descriptive bivariée (à deux dimensions)</td><td>Descriptive multivariée (à trois dimensions)</td><td>Paramétrique

(suivent la loi normale)</td><td>Non paramétrique (absence de normalité)</td></tr>
<tr><td>Position

- Moyenne

-Médiane

-Mode

-Quartiles (Q_1, Q_3)</td><td>Dispersion

-Variance

-Ecart type

-Coefficient de variation (CV)</td><td>Corrélation

(selon la normalité des variables)
↙ ↘
r | ∫</td><td>• ACP
• AFC (AFCM)</td><td><u>1-Variables quantitatives</u> :

- test student T

- Anova à un facteur

- Anova à deux facteurs

- Manova

- Ancova (analyse de la covariance)

<u>2-variables qualitatives</u>

- test Khi-deux</td><td>Ces types de tests sont des alternatifs pour les tests non paramétriques dans le cas ou les variables ne suivent pas la loi normal</td></tr>
</table>

CHAPITRE II : Statistiques descriptives

II.1. statistiques descriptives à une dimension

Paramètres de position	Paramètres de dispersion
La moyenne X ou m = 1 /n Σ x_i	**L'étendue** E = $x_{max} - x_{min}$
Le mode (Mo) : c'est la valeur ou classe correspondant à l'effectif (ou fréquence) le plus élevé.	**La variance** S^2_x = 1/ n Σ *f* $(x_i - x)^2$ (la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité dont elle permet de visualiser si les valeurs sont elles proches ou éloignées ?)
La médiane (Me) : valeur centrale de la série statistique.	**L'écart type :** $s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$ **(racine de la variance) (Plus l'écart-type est faible, plus la population est homogène,** savoir si les valeurs sont très dispersées ou si elles sont proches de la moyenne**)**
Les quartiles Les quartiles partagent la série en quatre groupes. Le premier quartile : c'est la plus petite donnée de la liste telle qu'au moins au quart des données de la liste sont inférieures ou égales à Q_1 = N/4 le troisième quartile : c'est la plus petite donnée de la liste telle qu'au moins les trois quarts des données de la liste sont inférieures ou égales à Q_3 = N x 3/4	**Le coefficient de variation C.V. =** S /m x 100 1) CV < 5% : les valeurs sont très homogènes 2) 5% < CV < 10% les valeurs sont homogènes 3) 10% < CV < 15% les valeurs sont moyennement homogènes 4) 15% < CV < 30% les valeurs sont hétérogènes 5) CV > 30% les valeurs sont très hétérogènes **C.V est pour l'évaluation du degré d'homogénéité des échantillons**

Vous avez l'exemple des boites à moustaches :

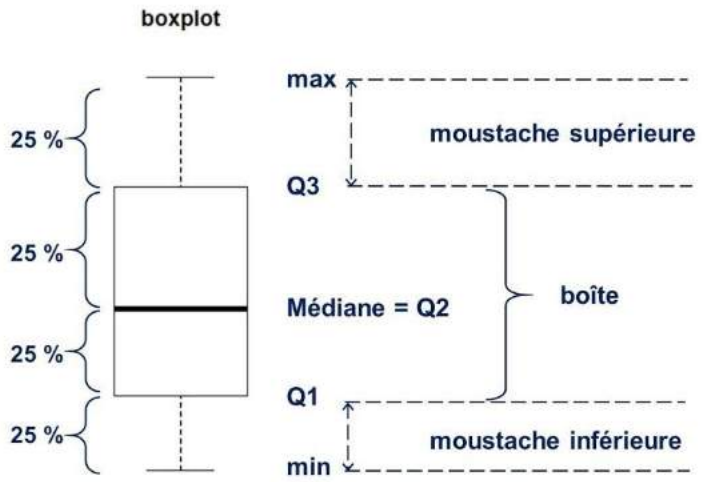
Graphe de la boîte à moustache d'une distribution

▷ statistique descriptive

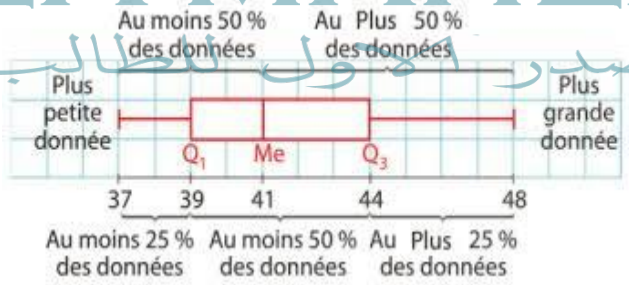
réaliser le graphe de la boîte à moustache d'une distribution :

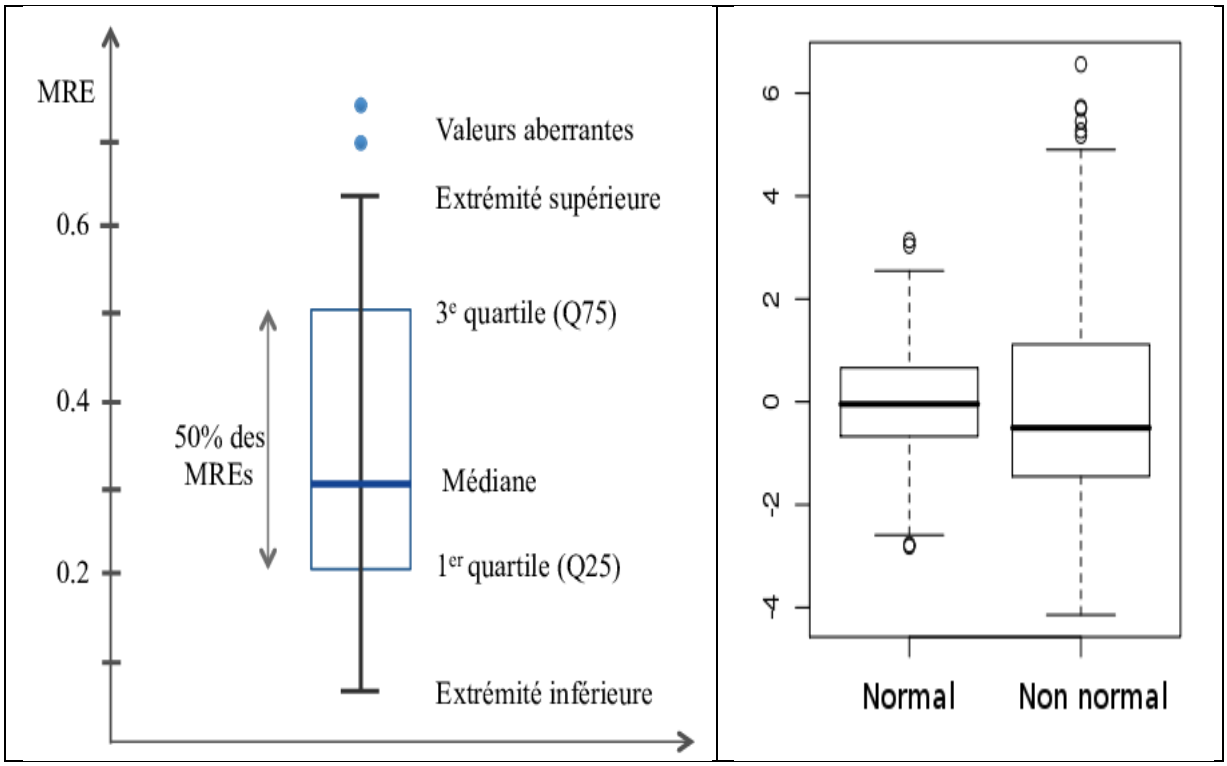
Instruction boxplot

```
> boxplot(data1)
```



Boîte à moustaches, boîte de Tuckey, boîte de distribution, boxplot, ...)





Petit exo :

Cette boite à moustache représente un diagramme de répartition ou de dispersion des données

L'extrémité de chaque moustache représente le plus(arbres jeunes de 8ans) et le grand âge () de ces arbres

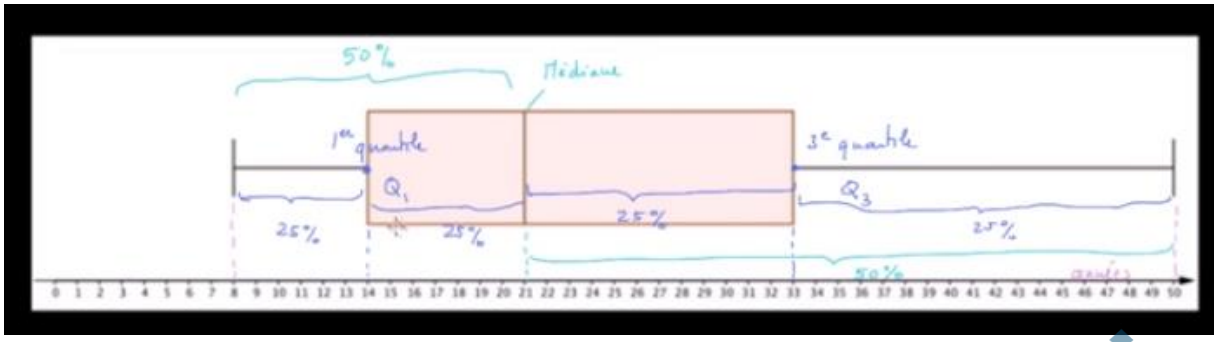
Réponse 1 :

L'étendue = la différence entre la grande et la plus petite valeur

L'étendue = 50- 8= 42 ans

Réponse 2 :

L'âge médian c'est 21Ans, càd que y'a autant d'arbres qui ont moins de 21ans que d'arbres qui ont plus de 21ans



Q1 c'est la médiane des 1er 50%

Remarque :

Intervalles de confiance (marge d'erreur):

Nous avons vu que la moyenne $\overline{X}$ d'un échantillon aléatoire permet d'estimer la vraie moyenne μ de la population.

Nous voudrions estimer également la précision de cette moyenne, c'est-è-dire donner une marge d'erreur ou un intervalle de confiance.
La marge d'erreur prend uniquement en compte l'erreur de l'échantillon.

En statistiques, la **marge d'erreur** exprime la quantité d'erreur d'échantillonnage aléatoire dans l'estimation d'un paramètre, comme la moyenne ou la proportion. Plus la marge

d'erreur est grande, plus l'intervalle est large, moins l'estimation du paramètre est précise et moins on peut avoir confiance que les résultats sont proches des vrais résultats, et ainsi, de la réalité.

Il existe trois différents niveaux de **l'intervalle de confiance**.

Le niveau de **95 %** est le plus répandu (*) **=======>**p= 5%

Le niveau de **99 %** est le plus prudent (**) **======>**p= 1%

Le niveau de **99,9** % est l'idéal (***) **==========>**p= 0,1ù%

Le niveau de 90 % est rarement utilisé.

Pour un niveau de confiance de 99 %, on est sûr à 99 % que la vraie valeur se trouve dans la marge d'erreur de la valeur de l'échantillon.

II.2. Statistiques descriptives à deux dimensions (exemple d'application: corrélation)

La statistique descriptive à deux dimensions a essentiellement pour but de mettre en évidence les relations existantes entre deux séries d'observations considérées simultanément (càd d'étudier sur une même population de n individus, deux caractères différents X et Y et de rechercher s'il existe un lien entre ces deux variables= **dépendance**).

II-1- La covariance est une **mesure de la relation linéaire** entre deux variables aléatoires.

La covariance : C'est la moyenne des produits des écarts pour chaque série d'observation.

$Cov(x,y) = S_{xy} = 1/n \sum(x_i - m)(y_i - \bar{y})$

II.2. La corrélation

En probabilité et en statistique, étudier **la corrélation** entre **deux** ou plusieurs variables aléatoire, c'est étudier **l'intensité de la liaison** qui peut être existée entre ces variables. Donc **la corrélation** est une notion de liaison qui **contredit** leur **indépendance**.

NB : la corrélation est une normalisation de la covariance.

Une mesure de cette corrélation dans le cadre linéaire est obtenue par **le calcul** du coefficient appelé **coefficient de corrélation (**mesure **la dépendance** linéaire entre les variables X et Y**)**.

Ce **coefficient de corrélation linéaire**, est égal est précisément le rapport de la covariance sur le produit des écarts-types de deux variables X et Y.

r = Cov (x, y) / Sx . Sy

La valeur absolue du coefficient, toujours comprise entre 0 et 1.

On a -1 < r < 1. Si r est proche de 1 ou -1, les variables X et Y sont dits : fortement corrélées.

II.2.1. Le coefficient de corrélation

Le coefficient de corrélation est une fonction de la covariance. Il est égal à la covariance divisée par le produit des écarts types des variables. Une covariance positive a donc toujours pour résultat une corrélation positive et de la même façon, une covariance négative a toujours pour résultat une corrélation négative.

Le coefficient de corrélation nous donne des informations sur **l'existence** d'une relation linéaire entre les deux variables. A cet effet, un coefficient de corrélation **nul** ne signifie pas l'absence de toute relation entre les deux variables mais seulement l'absence d'une relation linéaire.

Pour des variables quantitatives, choisissez **le coefficient de corrélation** de :

coefficient de corrélation de Pearson (r)	coefficient de corrélation de Spearman (rho)
↓	↓
Quand les valeurs suivent la loi normale	Quand les valeurs ou bien l'une des deux valeurs ne suivent pas la loi normale

Propriétés

C'est une technique de prévision pour des variables quantitatives. Le but de la corrélation est de savoir s'il existe une relation entre ces deux variables.
En général, cette relation peut être mesurée par un coefficient de corrélation de Pearson qui est noté "**r**".

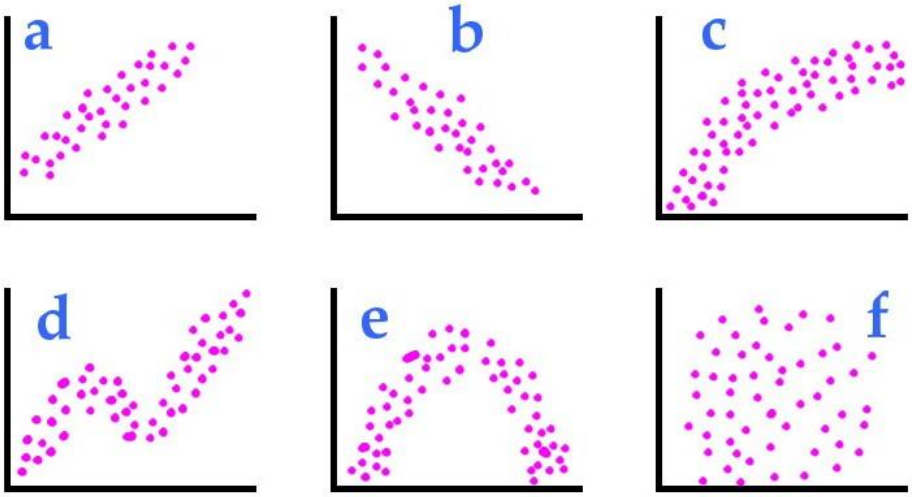
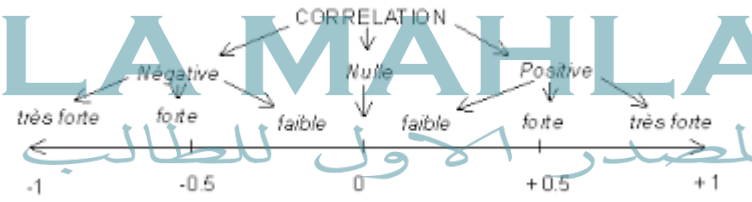


Figure 1 : Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation

- **r** est toujours compris entre **-1 et +1**
- la corrélation dépend toujours de **: l'intensité (**forte ou faible**) et de la signification.**
- **r = 0 ou voisin de 0** signifie **une absence de corrélation (graphique f)**
- **r < 0 : une corrélation négative (**plus je cours vite, moins je vois le paysage et inversement; graphique b**);**
- **r > 0 : une corrélation positive (**plus je cours vite, plus j'ai soif**;** graphiques a ou c)**;**
- **r = 1 ou r = -1 : relation linéaire parfaite (**graphiques a ou b où tous les points seraient parfaitement alignés).



❖ Avec un logiciel statistique, on obtiendra en général le résultat sous forme de "matrice triangulaire" (c'est-à-dire de tableau comme ci-dessous). De plus, on peut demander la "signification statistique" du coefficient de corrélation (cf. "Matrix of Probabilités").

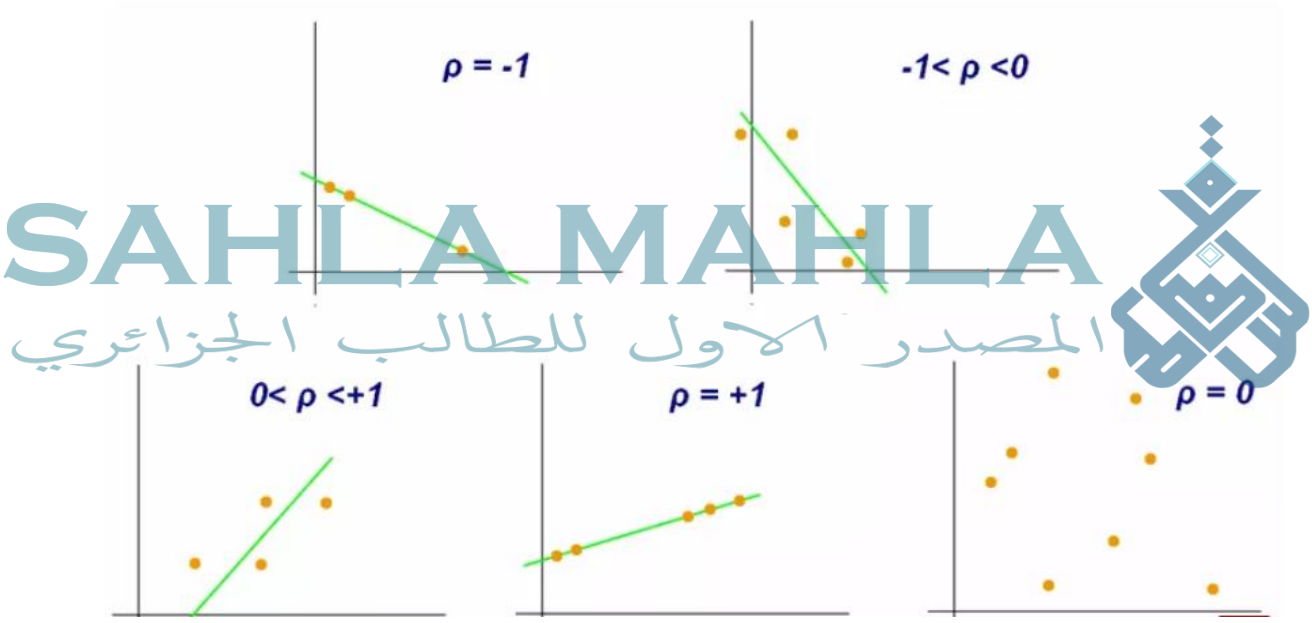
```
Pearson correlation matrix
                 MATH       FRANCAIS
MATH            1.000
FRANCAIS        0.906        1.000


Matrix of Probabilities
                 MATH       FRANCAIS
MATH            0.000
FRANCAIS        0.000        0.000

Number of observations: 15
```

D'après ces tableaux, le coefficient de corrélation entre les notes de math et de français est r = 0.906; cette relation est significative (p < 0.001).

Petite évaluation



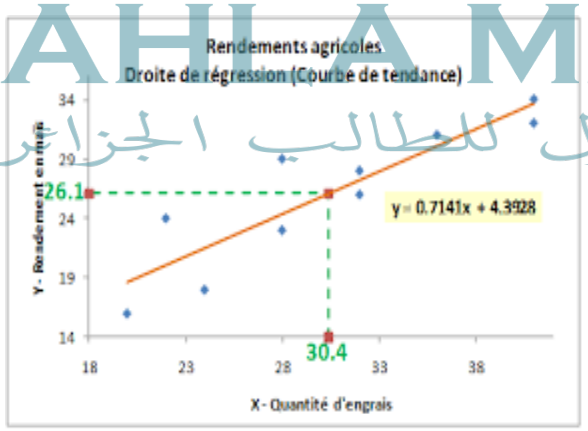
Du droite à gauche :

- Forte corrélation négative avec une pente (r = -1 ; càd : pente -)
- Corrélation moyenne négative
- Corrélation moyenne positive
- Forte corrélation (cas théorique ou bien corrélation parfaite)
- Absence de corrélation

Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite. Cette méthode vise à expliquer un nuage de points par une droite qui lie **y** à **x** (figure 2).

Y = a. **x** + b /



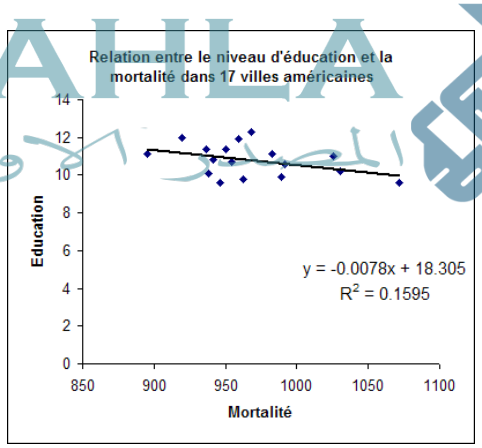
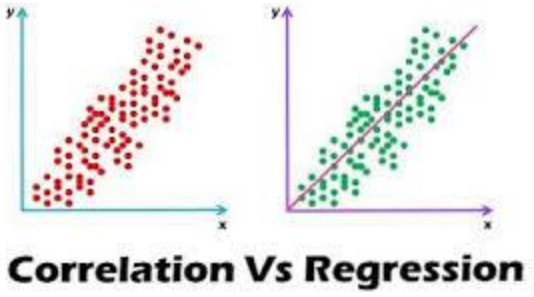


Figure 2 : exemples des droites de régression

Remarque



Corrélation est décrite comme l'analyse qui nous permet de connaître l'association ou l'absence de la relation entre deux variables 'x' et 'y'. À l'autre bout, **Régression** analyse, prédit la valeur de la variable dépendante en fonction de la valeur connue de la variable indépendante.

Le point ci-dessous explique la différence entre les deux:

Corrélation	Régression
Représenter une relation linéaire entre deux variables.	Ajuster une meilleure ligne et estimer une variable sur la base d'une autre variable.

Ajustement linéaire :

L'ajustement linéaire consiste à remplacer le nuage de points par une droite à l'aide d'une équation de la régression.

Remarque :

Le coefficient de corrélation permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir figure 3) :

- Si **r < 0,7 ;** alors l'ajustement linéaire est **refusé (droite refusée).**

- Si **r ≥ 0,7 ;** alors l'ajustement linéaire est **accepté (droite acceptée).**

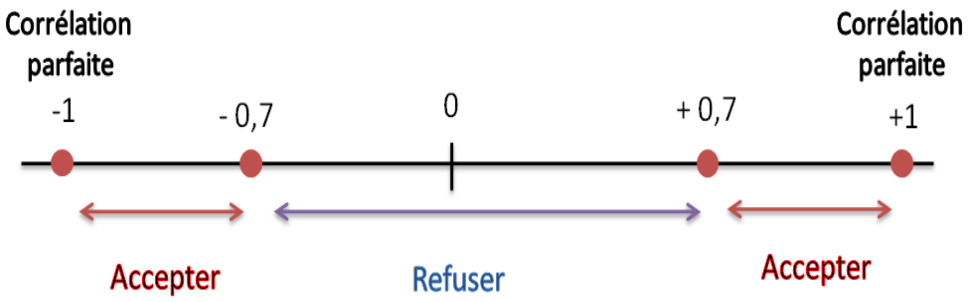


Figure 3 : la zone d'acceptation ou de refus de l'ajustement linéaire

Remarque :

Pour mieux juger la qualité d'une régression linéaire, on définit un autre **indicateur** compris entre 0 et 1, nommé : coefficient de détermination, noté R^2 :

Proche de la **droite de régression** on retrouve ce**: R^2 (coefficient de détermination :** soit le carré du coefficient de corrélation linéaire r), est un indicateur qui permet de juger la qualité d'une régression linéaire simple.

Ce nombre mesure l'adéquation entre le modèle et les données observées ou plus, R^2 est prés de 1 plus le modèle est adéquat, et le contraire est vrai.

Exemple 1 : un coefficient de corrélation **r = 0,9** correspond à un coefficient de détermination **$R^2 = r^2 = 0,81$**.

Cela signifie que **81%** de **la variance** de y est expliquée par la corrélation : la corrélation est bonne

Exemple 2 : Mais un coefficient de corrélation **r = 0,5** correspond à un coefficient de détermination **$R^2 = r^2 = 0,25$**.

Cela signifie que seulement **25%** de la variance de y est expliquée par la corrélation : la corrélation est mauvaise

- Si le **R^2** est nul, cela signifie que l'équation de la droite de régression détermine 0 % de la distribution des points. Cela signifie que le modèle mathématique utilisé n'explique absolument pas la distribution des points.

- Si le **R^2** vaut **1**, cela signifie que l'équation de la droite de régression est capable de déterminer 100 % de la distribution des points. Cela signifie alors que le modèle mathématique utilisé, ainsi que les paramètres **a** et **b** calculés sont ceux qui déterminent la distribution des points.

- En bref, plus le coefficient de détermination **R^2** se rapproche de **0**, plus le nuage de points se disperse autour de la droite de régression. Au contraire, plus le **R^2** tend vers **1**, plus le nuage de points se resserre autour de la droite de régression. Quand les points sont exactement alignés sur la droite de régression, alors **$R^2 = 1$**.

Exemples des modèles prédictifs :

5. Modèle d'ajustement et d'optimisation de la bioactivité des extraits

Souche bactérienne	Stade phénologique	Variables explicatives	R^2	r	Prédiction de l'activité des composés des extraits au stade phénologique
Salmonella enterica	stade floraison	FLV	0,986	0,993	Y_{iz}=1,50 FLV-2,42
Staphylococcus aureus	stade floraison	FLV	0,997	0,998	Y_{iz} =31,92FLV-96,59

III. Statistiques inférentielles (Tests de comparaison)

But : Les statistiques inférentielles, consistant en des tests permettant de confirmer ou infirmer une hypothèse (**recherche des différences**).

Les tests de comparaisons sont basés sur les deux hypothèses:

Hypothèses

H0 : $m_1 = m_2$ (càd les deux groupes de comparaisons appartiennent à des populations qui possèdent des moyennes identiques) **ex** : 2 produits sont efficaces

H1 : $m_1 \neq m_2$, $m_1 < m_2$, $m_1 > m_2$; **ex** : un produit est plus efficace que l'autre

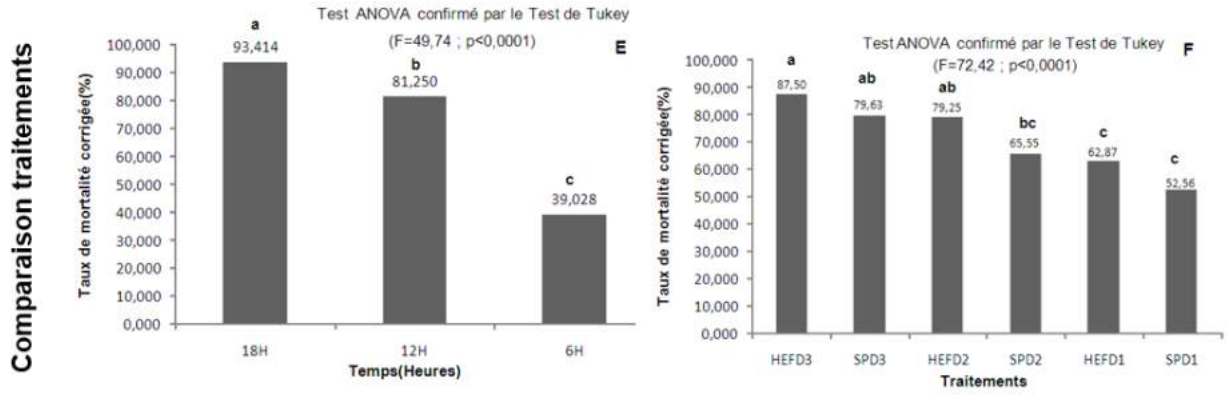
Méthodes inférentielles	
Objectif : recherche des différences (présence de signification ou absence) **(tester la normalité tout d'abord pour choisir entre les 2)**	
Paramétrique **(suivent la loi normale)**	**Non paramétrique** **(absence de normalité)**
1-Variables quantitatives : - Un échantillon (test de conformité) **- Test t ou test de student** Séparés (indépendants) - Deux échantillons Liés (dépendants)	Man withney Wilcoxon

- Anova à un facteur (deux ou +eurs échantillons) →	Kruskal wallis
- Anova à deux facteurs →	Fridman
- Manova	
- Ancova (analyse de la covariance)	
2-variables qualitatives	
- test Khi-deux	

III- 1- Statistiques inférentielles (exemple d'application: ANOVA)

Analyse de la variance à 1 facteur type ANOVA (ANOVA one way)

C'est un modèle statistique utilisé pour comparer les moyennes d'échantillons. Ce test s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielles (appelées alors facteurs de variabilité, leurs différentes modalités étant parfois appelées « niveaux ») qui ont de l'influence sur la loi d'une variable continue à expliquer. On parle d'analyse à un facteur lorsque l'analyse porte sur un modèle décrit par un seul facteur de variabilité.



Autrement dit elle sert à envisager :

⇨ **Dépendance d'une variable quantitative à une (ou deux) variable(s) qualitative(s)**

⇨ **Variable(s) qualitative(s) = facteur(s)** *(expliquant la dépendance)*

L'analyse de la variance est appelé: Bivariée ou multivariée

⇨ **Un seul facteur : analyse bivariée**

⇨ **Si 2 facteurs : analyse multivariée**

⇨ **Variable Dépendante : la variable quantitative continue [VD]**

⇨ **Variable(s) Indépendante(s) : le(s) facteur(s) [VI]**

NOUS ALLLONS TESTER SI

- ⇨ **La dépendance étudiée est-elle ou non significative pour le facteur considéré ?**
- ⇨ **La moyenne de la variable quantitative d'étude est-elle homogène sur l'ensemble des modalités de la variable qualitative ?**
- ⇨ **Rejeter l'hypothèse nulle Ho d'égalité des moyennes**
- ⇨ **Analyse de la variance**
- ⇨ **Test F de Fisher** *(comparant la variance inter-échantillon à la variance intra-échantillon)*

Lorsque F >>1 ⇨ p_{value} << 0,05 Ho rejetée (F = variance inter / variance intra)

But : C'est un test permet de chercher et de comparer la différence entre plusieurs échantillons (moyennes) quantitatifs.

Principe: Leur principe repose sur la comparaison d'un facteur calculé ($F_{calculé}$) (observé) par rapport à un autre facteur théorique (critique) en fonction de degrés de liberté (v_1 et v_2) et au seuil de signification (α).

Si la valeur de *F* n'est pas compatible avec cette loi de Fisher (c'est-à-dire que la valeur de est supérieure au seuil de rejet), alors on rejette l'hypothèse nulle : on conclut qu'il existe une différence statistiquement significative entre les distributions.

Analyses de variance en conditions non paramétriques

Conditions à réaliser pour ANOVA en conditions paramétriques
- variances comparables entre les différents groupes
- distribution normale pour chacun des groupes

Si un des critères au moins n'est pas réalisé, faire un test de Kruskal-Wallis : il y a un différence significative entre les groupes si la probabilité associée est inférieure à 5 %

Statistique descriptive multidimensionnelle (exemple d'application: ACP)

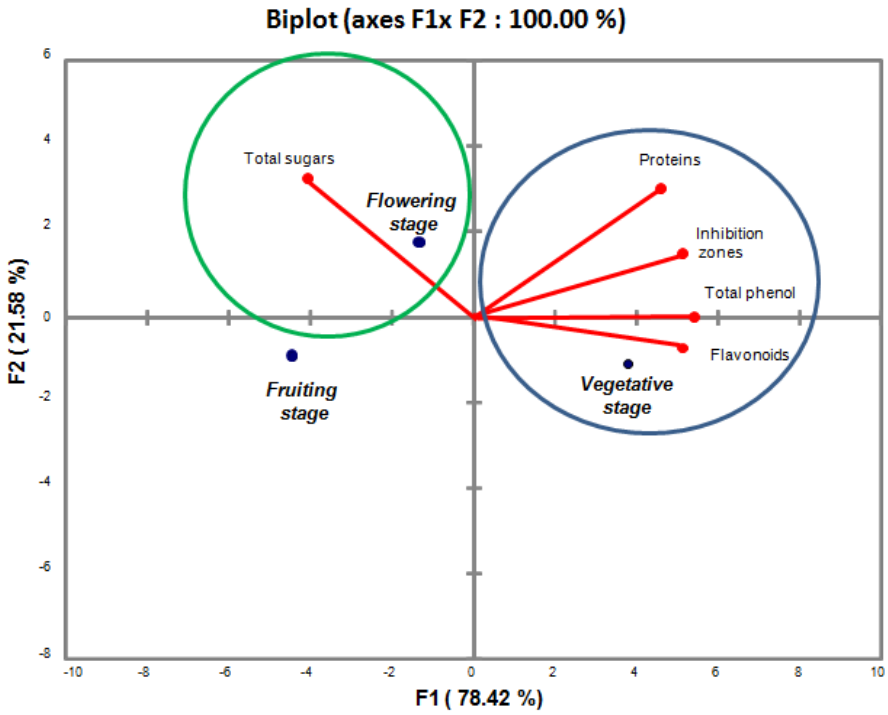
Analyse en Composantes Principales (ACP)

L'objectif général cette fois-ci est de tenter de regrouper les individus en groupes assez homogènes.

But : résumer le maximum d'informations possibles et en perdant le moins possible pour :

- Faciliter l'interprétation d'un grand nombre de données initiales ;
- Donner plus de sens aux données réduites

L'ACP permet donc de réduire des tableaux de grandes tailles tout en conservant un maximum d'information.



Exemple de projection ACP avec interprétation

Les projections des valeurs de contenu des extraits en métabolisme primaire et secondaire sur le premier axe (76.32%) ont permet de constater la présence d'une corrélation positive entre les phénols totaux (TP) et les flavonoïdes (FLV) par rapport au stade végétatif, alors que le stade floraison est corrélé positivement avec les métabolites primaires notamment les sucres totaux (SGR). Ces tendances ont été confirmé par les valeurs respectives du coefficient de corrélation de Pearson (r=0.973 ; r=0.994)

Analyse des résultats :

Analyser les résultats d'une ACP, c'est répondre à trois questions :

1- Les données sont-elles factorisables (réalisables) ?
2- Combien de facteur à retenir ?
3- Comment interpréter les résultats ?

1- ***Les données sont-elles factorisables (réalisables) ?***

- Pour répondre a cette question, dans un **premier temps**, il convient d'observer la matrice des corrélations. Si plusieurs variables sont

- **corrélées (> 0,5),** la factorisation est possible. Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.
- Dans un **deuxième temps**, il faut observer **l'indice de KMO** (Kaiser-Meyer-olkin) qui doit tendre vers 1. Si ce n'est pas le cas, la factorisation n'est pas conseillée. Pour juger de **l'indice de KMO,** on peut l'échelle suivante **:**

- **> 0,9 merveilleux !**
- **> 0,8 méritoire (idéal)**
- **> 0,7 moyen (acceptable)**
- **> 0,6 médiocre**
- **> 0,5 misérable** } **l'analyse ne peut pas être conduite**
- **< > 0,5 merdique !**

Enfin : on utilise le test de sphéricité de Bartlett:

* si la signification (sig.) tend vers 0,000 ; c'est **très significatif**
* inférieure à 0,05 **significatif**
* entre 0,05 et 0,10 **acceptable**
* au-dessus de 0,10 ; **on rejette**

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,889
Bartlett's Test of Sphericity	Approx. Chi-Square	66,059
	df	6
	Sig.	,000

Si l'ACP satisfait au moins deux de ces trois conditions, on peut continuer

1- Combien de facteur à retenir ?

* on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l'information.

* d'autre source dit que la variance expliquée totale (% variance expliquée : minimum 60%)

* autres disent au moins 50%

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,876	71,892	71,892	2,876	71,892	71,892
2	1,120	27,992	99,884	1,120	27,992	99,884
3	,004	,089	99,974			
4	,001	,026	100,000			

Axe 1 — Axe 2 — Σ axes

Extraction Method: Principal Component Analysis.

1- Comment interpréter les résultats ?

C'est la phase la plus délicate de l'analyse.

Si la variance expliquée est trop faible, on peut choisir d'exclure certaines variables.

Pour choisir les variables à éliminer, on observe leur qualité de représentation : plus la valeur associée à la ligne «extraction » est faible, moins la variable explique la variance.

- Ce tableau est pour éliminer les variables qui causent le faible % des axes

Qualités des représetation

	Initial	Extraction
MATH	1,000	,999
PHY	1,000	,999
FRANS	1,000	,999
ANGL	1,000	,998
CLASSE	1,000	,411

C'est des valeurs de cosinus carré

Très éloignés on peut les éliminés

Extraction Method: Principal Component Analysis.

Il faut toujours tenir compte du positionnement de chaque variable sur chaque axe : les variables à éliminer sont les variables qui sont:

- Soit proche du centre sur l'ensemble des axes retenus
- Soit au milieu d'un quart de cercle sur les axes retenus

La qualité de la représentation d'un point sur un axe : un individu donné est d'autant mieux représenté sur un axe, que son cosinus carré est proche de 1 ; il est d'autant mal représenté que son cosinus carré est proche de O.