UNIVERSITE HASSIBA BENBOUALI DE CHLEF

Faculté des Science de la Nature et de la Vie

Département Eau, Environnement et Développement Durable

Polycopié

Statistique appliquée

License : Eau et Sol

Filière : Science Agronomique

Préparé par : Dr Habibi Brahim

Maitre de Conférences classe B

***Département* Eau, Environnement et Développement Durable**

Année universitaire 2019-2020

INTRODUCTION

Les statistiques sont une science qui utilise des méthodes scientifiques pour collecter, organiser, synthétiser, présenter et analyser les données de tel ou tel phénomène. Elles permettent aussi de tirer des conclusions valables et de prendre des décisions raisonnables sur la base de ces analyses**.**

I. L'organisation et le contrôle des données

I.1 L'organisation des données

I.1.1 Acquisition des données

L'acquisition de données consiste à procéder, par le biais d'un instrument de mesure, à acquérir de l'information (par exemple : hauteur d'eau d'une station limnométrique, comptage des basculements d'un pluviographe à augets, vitesse du vent etc…).

I.1.2 Traitement primaire des données

La donnée acquise précédemment nécessite souvent un traitement préalable - ou traitement primaire - afin de la rendre pertinente et exploitable.

Le traitement des données inclut aussi le contrôle primaire des données qui comprend les contrôles de cohérence à l'exclusion de tous traitements statistiques. Il s'agit par exemple, dans le cas d'une acquisition manuelle des données, de les convertir en fichiers informatiques. Dans ce cas, on procède généralement à une double saisie des données puis les fichiers sont comparés afin de déceler d'éventuelles erreurs de saisie.

I.1.3 Contrôle des données

Avant de pouvoir exploiter les données et bien qu'elles soient dans un format adéquat, il importe de contrôler la fiabilité et la précision de ces dernières. Le contrôle permet de valider les données avant leur organisation au sein d'une banque de données pour leur mise à disposition à des fins opérationnelles.

a- Détection des anomalies par cumul des résidus

Quelque soit le soin apporté à la collecte et au traitement des données, il existe toujours des erreurs dans les fichiers. Ces erreurs peuvent être de deux types :

- **Erreurs accidentelles :**

- Erreurs de lecture,
- Années incomplètes non signalées,
- Erreurs de transcription.

- **Erreurs systématiques :**

- Changement de site,
- Utilisation d'une éprouvette inadéquate,
- Modification de l'environnement du poste,

Changement d'observateur

1.4 Organisation des données

Au vu de l'importance quantitative et qualitative des données, il importe de les organiser avec soin. Ceci se fait à partir d'un corpus de documents originels (formulaires de terrain, diagrammes, unité de stockage électronique) constituant les archives qui sont en règle générale accessibles uniquement à un personnel spécifique (responsable du centre de collecte, archiviste…). La traduction des archives sous la forme de fichiers de base génère les "fichiers en l'état" et fournit une indication sur la provenance de la donnée (mesure, calcul, copie etc.) ainsi que sur sa qualité (fiable, complète ou non) et sa précision. Enfin, on constitue un fichier de travail provisoire permettant une visualisation des données et permettant de procéder aux différents tests de qualité et de précision des données qui seront développés tout au long de ce chapitre.

II. Collecte, analyse et extension des données

II.2. Homogénéisation des données en hydrologie

Qu'est ce que l'homogénéisation des données? Pour répondre à cette question qui n'est pas aussi simple que l'on ne croit, il faut saisir et mesurer l'importance des dégâts que l'on peut avoir suite à une information fausse appliquée par un ingénieur pour dimensionner l'ouvrage . Et à ce moment là, il faut revenir et se poser la question : au fait à partir de quelle information de base suis-je parti pour faire mes calculs, est-elle fiable? Tout le problème est là. L'homogénéisation des données est une analyse statistique de l'information aidant à une prise de décision conséquente. Elle consiste en :

- la détection des anomalies dans les séries hydrologiques et d'en chercher la cause ;
- la correction de ces anomalies par des méthodes appropriées ;
- l'extension des séries hydrologiques courtes à partir de séries de base homogènes, soit l'estimation d'une ou de plusieurs observations d'un échantillon à partir d'autres observations prises dans des endroits et à des moments différents sous la condition qu'il existe des liens de dépendance assez étroits.

II.2. Détection des erreurs et correction des données

Toute étude hydrologique nécessite la vérification des données utilisées. L'information de base quant à sa qualité revêt une très grande importance. On ne peut espérer à des résultats concluants si la donnée de base n'est pas fiable. De ce fait, l'analyse hydrologique se base sur l'exploitation de données, présentées souvent sous forme de séries statistiques et sujettes la plupart du temps à des erreurs qu'on appelle erreurs systématiques, qu'il convient de détecter et de corriger.

Le contrôle visuel s'avère toujours efficace et permet de déceler à prime abord les hétérogénéités grossières qui peuvent exister et de les corriger avec les originaux.

D'autres hétérogénéités moins évidentes peuvent exister et n'apparaissent pas lors de ce contrôle. Pour celles ci, il est obligatoire de recourir à certaines méthodes statistiques pour les déceler.

Les méthodes d'homogénéisation sont nombreuses et peuvent être graphiques ou analytiques.

1.1 Détection des erreurs

3.1 - Méthode des "doubles masses" (doubles cumuls)

Cette méthode a été longtemps utilisée car sa mise en œuvre est simple et ne nécessite pas de moyen de calcul particulier.

Elle permet de mettre en évidence des erreurs systématiques dans une série de données. Soit deux séries d'observations (xi, yi) sur des variables corrélées entre elles. Il existe alors une relation du type :

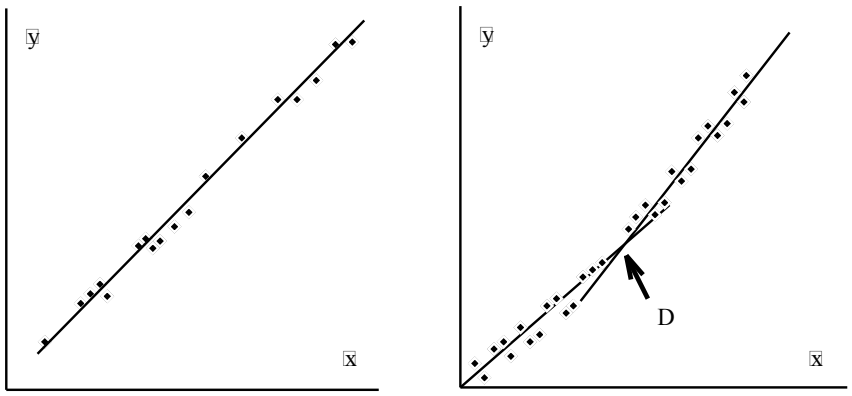
$$\hat{y}_i = ax_i + b$$

Si à partir d'une certaine date, on commet une erreur systématique sur x par exemple, les variables x et y seront encore corrélées mais avec des coefficients a' et b'.

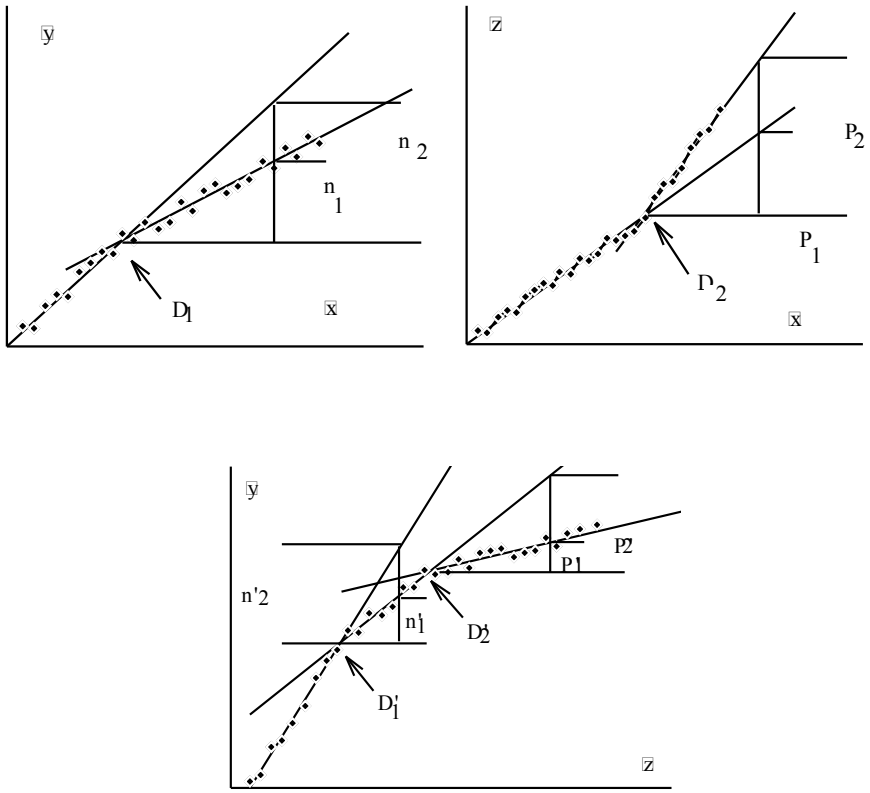
Si on porte dans un graphique xi en fonction de yi, il ne sera généralement pas possible de constater cet écart. Par contre, si on porte non plus xi et yi mais Yi et Xi définis ainsi...

$$Yi = \sum_{j=1}^{i} y_j \quad ; \quad Xi = \sum_{j=1}^{i} x_j$$

on aura alors une fonction monotone croissante en fonction du temps. Si x et y sont stables, les points s'aligneront sur une droite de pente $\frac{\bar{y}}{\bar{x}}$, mais si une des séries subit à partir d'une certaine date, une erreur systématique, on verra les points s'aligner selon deux droites :



Sur la figure de gauche, x et y sont stables alors que sur la figure de droite, on constate qu'à partir du point correspondant à la date D, on a commis une sous-estimation systématique de x ou une surestimation systématique de y. Pour lever cette ambiguïté, on effectuera les doubles masses sur plusieurs postes ou on vérifiera que les cassures se font bien aux mêmes dates et dans les mêmes rapports.



n corrige les données observées en multipliant le rapport de pente $\frac{m_1}{m_2}$ ou $\frac{m_2}{m_1}$ par la valeur erronée respectivement selon que l'on soit après la cassure ou avant la cassure.

4 Détection des anomalies par cumul des résidus

Quelque soit le soin apporté à la collecte et au traitement des données, il existe toujours des erreurs dans les fichiers. Ces erreurs peuvent être de deux types :

- ***Erreurs accidentelles :***
- Erreurs de lecture,
- Années incomplètes non signalées,
- Erreurs de transcription.
- ***Erreurs systématiques :***
- Changement de site,
- Utilisation d'une éprouvette inadéquate,
- Modification de l'environnement du poste,
- Changement d'observateur.

Si deux variables X et Y sont tirées d'une loi de Gauss à deux dimensions, leur régression est linéaire. Les écarts εi = Y (i) – (a X (i) – b) à la droite de régression suivent alors une loi de Gauss de moyenne nulle et d'écart type $\sigma\varepsilon = \sigma\sqrt{1-\sigma^2}$ pour chaque valeur Yi on

calculera le εi correspondant et sa fréquence théorique. Les valeurs de εi ayant des fréquences très rares correspondant à des Yi douteux.

On peut par cette méthode détecter des erreurs accidentelles qui n'apparaissent pas à l'étude des distributions marginales. Dans l'exemple de la figure II.3 l'année 1995 paraît anormale au seuil standard de 95%.

Les anomalies systématiques sont détectées sur la base de l'analyse du cumul des résidus de régression, comme il a été mentionné précédemment, le résidus εi est une variable aléatoire gaussienne de moyenne nulle et d'écart type $\sigma\sqrt{1-\sigma^2}$. On définit alors, la variable Si cumul des i premiers résidus :

Si= $\sum \varepsilon i$

Cette variable Si est une variable aléatoire de moyenne nulle et d'écart type :

$$\sigma S = \sigma\varepsilon\sqrt{\frac{i(n-1)}{n-1}}$$

Si on se fixe à un rapport de 95 %, il y à cinq chances sur cent pour que Si soit extérieur au segment :

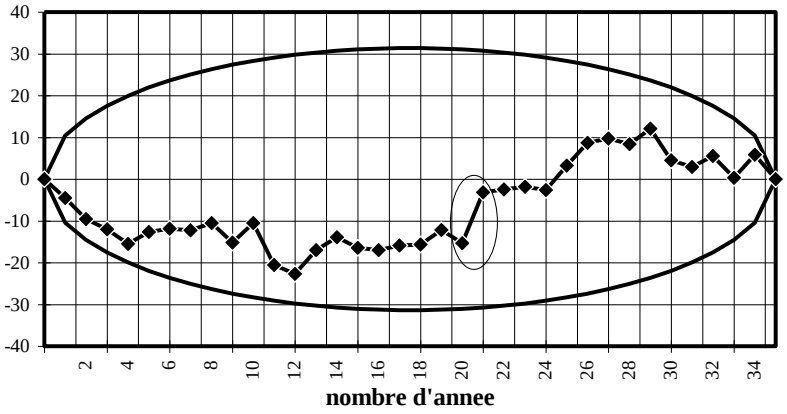
-1.96 $\sigma\varepsilon i$ < εi >1.96 $\sigma\varepsilon i$

On détecte ainsi des anomalies accidentelles qui ont moins de 5% de chance d'être dues au hasard. Pour ce qui concerne les cumuls des résidus, il y a également 95% de chance pour avoir :

-1.96 $\sigma s\varepsilon$ < $S\varepsilon$ >1.96 $\sigma s\varepsilon$

On détecte ainsi les anomalies systématiques qui ont moins de 5% de chance d'être dues au hasard.

1. Les anomalies accidentelles

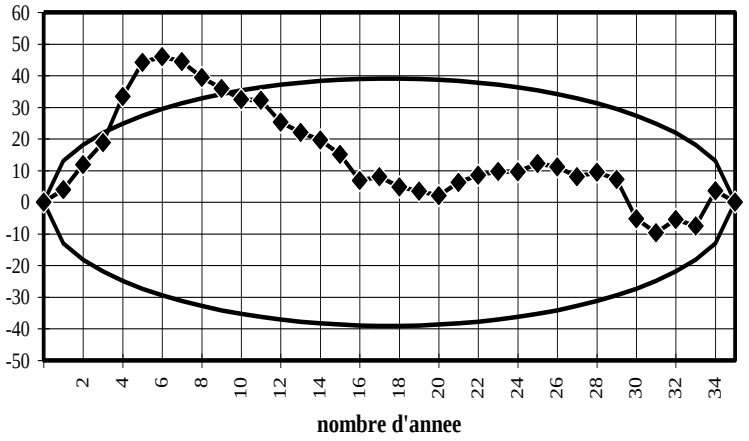


Dans rontation avec les données des stations avoisinantes mais en tenant compte du régime climatique de la région (semi aride) et de l'occurrence des orages pouvant être ponctuelles, ses chiffres peuvent être probables.

Les erreurs ponctuelles ont été corrigées lorsque le cumul des résidus de régression sort en dehors des segments de confiance, dans le cas contraire, ils sont considérés comme fiables.

2. Les anomalies systématiques

La détection des anomalies systématiques de vingt sept stations montre des ruptures de stationnarité comme par exemple la station de El Aouedj.



Les résidus cumulés en partant de1970, le dernier résidu étant celui de 2004, pendant les 06 premières années on remarque que les résidus sont sortie du cercle de confiance et après ils deviennent décroissants. Donc l'anomalie se situe autour 1976.

Les erreurs dues aux anomalies systématiques ne sont corrigées que lorsque les graphes de doubles masses ont donnés des cassures bien déterminées qu'on peut expliquer. Dans les cas douteux, lorsque les données étaient vraiment trop hétérogènes, les stations ont été éliminées.

1.1.2 Méthodes numériques

Elles consistent en l'utilisation de tests statistiques ou tests d'hypothèses ou tests de signification. Ces derniers permettent de comparer une population inconnue d'où est tiré notre échantillon avec une population connue ou hypothétique.

L'hydrologue fait l'hypothèse de travail, selon le résultat du test, il sera amené à rejeter ou non cette hypothèse. Le seuil de signification α est de 5% dans une étude hydrologique.

Nous avons deux types de tests :

- Tests non paramétriques qui ne font aucune hypothèse sur le type de loi concernée.

- Tests paramétriques qui supposent que la loi de la population est connue ;

Les tests paramétriques sont très puissants, mais ils impliquent la vérification de conditions restrictives. Il est indispensable que ces conditions restrictives soient effectivement vérifiées pour que l'utilisation du test ait un sens Toute supposition sur le type de loi de la population dans le seul but d'utiliser un test plus puissant est inutile puisqu'elle conduit à des conclusions sans signification. Dans le cas, où toutes les conditions requises

pour l'utilisation d'un test paramétrique puissant ne sont pas satisfaites avec certitude, il est sûr d'adopter un test non paramétrique moins puissant mais fiable.

Beaucoup de tests statistiques sont applicables dans le domaine de l'hydrologie, nous en citerons quelques uns des plus utilisés.

Tests non paramétriques

Test de Wilcoxon ou Test des rangs

C'est le plus puissant des tests non paramétriques. Soient 2 variables aléatoires Y et X, représentant respectivement 2 séries de précipitations annuelles de taille N1 et N2.

Y étant la série à étudier et X étant la série de base avec N2 > N1. Si l'échantillon Y est issu de la même population que l'échantillon X, l'échantillon nouveau Y U X est également issu de la même population. On classe les éléments de ce nouvel échantillon YUX par ordre décroissant et on associe à chacune des valeurs le rang qu'elle occupe dans cette nouvelle série.

(Si une valeur se répète plusieurs fois, il faut lui associer le rang moyen qu'elle détermine.

On calcule les quantités Wy et Wx :

Wy représente la somme des rang de Y et c'est celle qui nous intéresse et est égale à :

$$Wy = \sum_{i=1}^{n} \text{rang} = 1 + 3 + 4 + \ldots\ldots + 13 + 17 + \ldots + n$$

$$Wx = \sum_{i=1}^{n-1} \text{rang}_x = 2 + 5 + \ldots + 12 + 14 + 15 + 16 + \ldots + n\text{-}1$$

L'hypothèse nulle est vérifiée si :

$$Wmin < Wy < Wmax$$

Avec :

$$Wmin = \frac{(N_1 + N_2 + 1)N_1 - 1}{2} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

Et,

$$Wmax = (N_1 + N_2 + 1)N_1 - Wmin$$

$u_{1-\frac{\alpha}{2}}$ Représente la valeur de la variable centrée réduite de Gauss correspondant à une probabilité de $1 - \frac{\alpha}{2}$.

L'hypothèse d'homogénéité est rejetée si l'une des deux inégalités suivantes n'est pas vérifiée :

$$Wmin < Wy < Wmax$$

Ce rejet se fait au seuil de signification 1- α .

Exemple

débits	**49**	**55**	**63**	**76**	…	**125**	**125**	135	138	…	270	**285**	**330**	375
rang	1	2	3	4	…	26	27	28	29	…	72	73	74	75

La statistique W de Wilcoxon est la somme des rangs du premier échantillon. On a donc :

$W_x = 28 + 29 + \cdots + 72 + 75 = 2174$ et $W_y = 1 + 2 + 3 + 4 + \cdots + 26.5 + 26.5 + \cdots 73 + 74$ =676.

Pour $n_1, n_2 > 10$, on utilise l'approximation suivante :

$$W_x \sim N\left(\frac{n_1(n_1 + n_2 + 1)}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right) = N(1596, 8778)$$

La valeur critique est 1750. Comme $W_x > 1750$, on rejette l'hypothèse nulle ce qui est conforme à notre attente.

Test de la médiane ou Test de Mood

Ce test permet de vérifier si une série de données est homogène. Soit un échantillon x1, x2, x3,……xn ; déterminons sa médiane M après avoir classé l'échantillon par ordre croissant.

La médiane M est une constante de telle sorte que 50% des xi lui soient inférieures et 50% des xi lui soient supérieures.

Remplaçons donc la série des valeurs non classées par une suite de signe :

$$\begin{aligned} &+ \quad \text{pour les } x_i > M \\ &- \quad \text{pour les } x_i < M \end{aligned}$$

Calculons les quantités Ns et Ts, avec :

Ns : nombre total de séries de + ou de -

Ts : Taille de la plus grande série de + ou de -

Ns suit approximativement une loi normale de moyenne $\frac{N+2}{2}$ et de variance $\frac{1}{4}(N-1)$ et Ts suit une loi binomiale. Ceci a permis d'établir que pour un seuil de signification compris entre 91 % et 95 %, les conditions du test sont les suivantes :

$$\begin{aligned} Ns &> \frac{1}{2}\left(N + 1 - u_{1-\frac{\alpha}{2}}\sqrt{N+1}\right) \\ Ts &< 3.3\,(\log_{10} N + 1) \end{aligned}$$

Remarque : Un des plus puissants tests paramétriques est celui de Pearson ou test du χ^2. Vu son importance dans la vérification de l'adéquation d'une loi théorique.

Exemple

***Exemple* :** On veut vérifier l'homogénéité de la série des débits de pointe de la Viège sur la période totale d'observation.

On va classe les donnes plus petit jusqu a la grand

débits	240	171	186	158	…	145	155	230	270		330	55	63	49
signe	+	+	+	+	…	-	+	+	+		+	-	-	-

On a que $N_s = 22$ et $T_s =$ 9. Comme $N_s < \frac{1}{2}\left(n+1-1.96\sqrt{n-1}\right)$ =29.5, on rejette l'hypothèse nulle

Test paramétrique

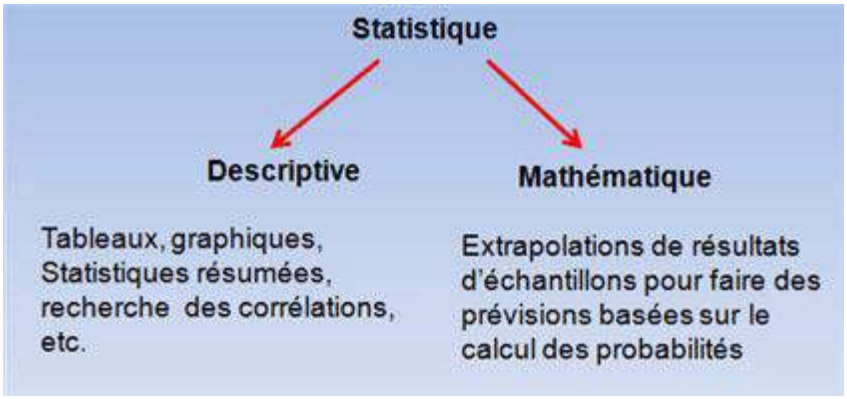
Ces tests sont beaucoup plus utilisés dans l 'ajustement théorique d'une loi de probabilité à une distribution empirique. est de Pearson ou du X^2 (Analyse fréquentielle

Statistique descriptive

Définition du champ de la statistique descriptive On divise généralement l'étude de la**statistique générale** en deux parties :

- ✓ La **statistique descriptive**, qui est un ensemble de méthodes permettant de décrire les **unités statistiques** qui composent une **population**

- ✓ La **statistique mathématique** dont l'objet est de formuler des lois à partir de l'observation d'**échantillons**, c'est-à-dire de tirages limités effectués au sein d'une population. La statistique mathématique intervient dans les **enquêtes** et les **sondages**. Elle s'appuie sur la statistique descriptive, mais aussi sur le calcul des **probabilités**.



1- D**éfinitions**

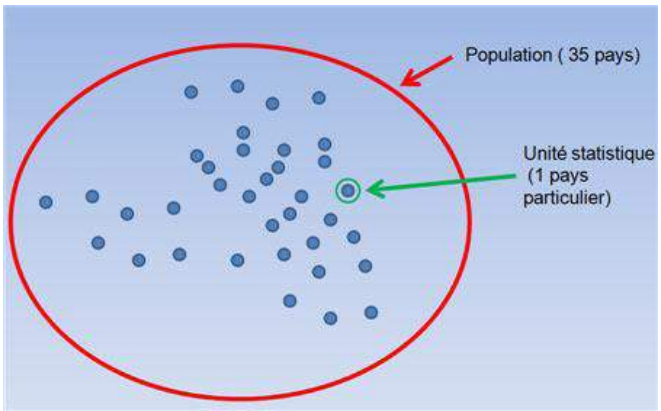
Au cours d'une enquête dans une classe 20 élèves, on pose les questions suivantes :
Combien avez-vous de frères et soeurs ?
Quelle est leur taille ?
Quel moyen de transport utilisez-vous pour venir à l' école ?

2- **Population, individu, échantillon**

Une Population est l'ensemble des éléments auxquels se rapportent lesdonnées étudiées. En statistique, le terme « population » s'applique à des ensemblesde toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprise d'un secteur donné…Dans une population donnée, chaque élément est appelé « individu» ou « unité statistique ».

3- **Population et unités statistiques**

En statistique, la **population** désigne un ensemble d'**unités statistiques**. Les unités statistiques sont les entités abstraites qui représentent des personnes, des populations d'animaux, poissons d'une rivière ou des objets. Les premières populations ayant fait l'objet d'un recensement ayant été des populations humaines (d'où le lien étroit entre statistique et démographie) le terme "individu" est parfois employé comme synonyme du terme "unité

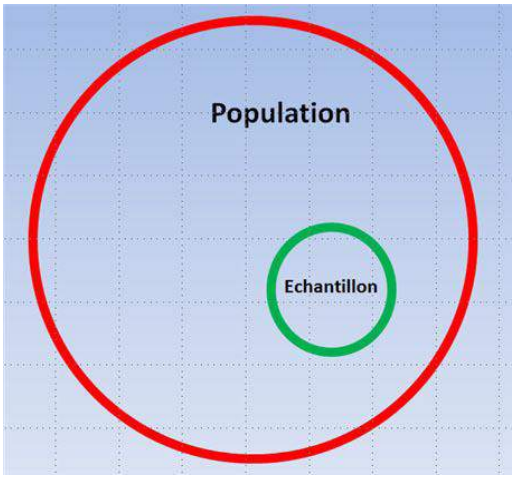


4- **Echantillons et sous-ensembles d'une population**

La taille de l'échantillon est le nombre d'événements qui le constituent. On dira qu'un échantillon est exhaustif lorsque sa taille est celle de la population.

Echantillon et population Dans une population donnée, chaque élément est appelé « individu » ou « unité statistique ».

. Le **diagramme d'EULER** ci-après décrit le lien entre l'échantillon et la population.

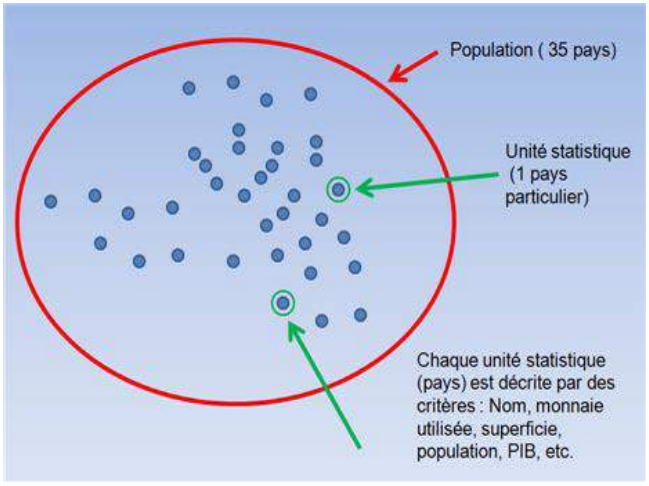


Le lien entre l'échantillon et la population

5- **Caractère : C'et** une propriété possédée par les unités statistique

Exemple de caractères
Le taux de glycémie, la vitesse de coagulation ; la production laitière ;

5.1 Critères de classification Nous avons vu dans l'exemple précédent que les unités statistiques d'une population pouvaient être regroupées suivant des **dimensions** ou **critères**. Ces critères sont choisis en fonction de ce qui intéresse le statisticien.



a- Variables quantitatives

Variables dont les valeurs sont numériques. C'est l'unique possibilité dans le cas de variablesaléatoires au sens strict.

On distingue deux types de variables quantitatives :

- **Variables discrètes**, dont les valeurs sont discrètes, si elle ne peut prendre que des certaines valeurs en nombre fini.

Exemple : nombre d'étudiants dans un amphi, durée d'une période pluviométrique humide.Nombre de crue par an

Dans ce cas la variable aléatoire x prend des valeurs discret x1, x2 ……..xn.

- ✓ **La fréquence relative** est notée*f*i de la classe i ; ***f*i=ni/N**

ni est le nombre de fois que x prend la valeuri

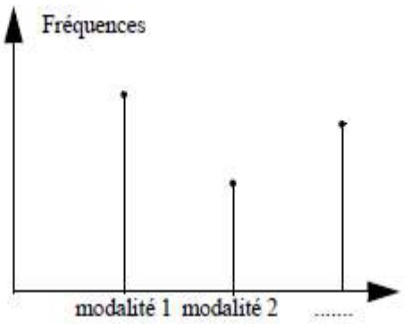
N est la somme des effectifs pour chaque valeur xi

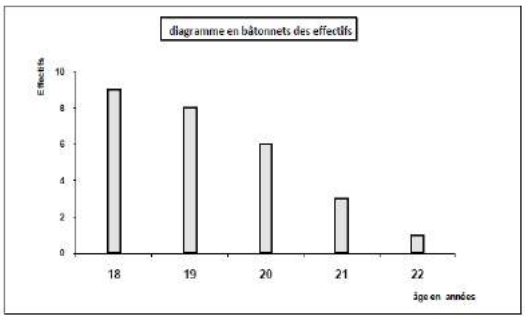
Exemple

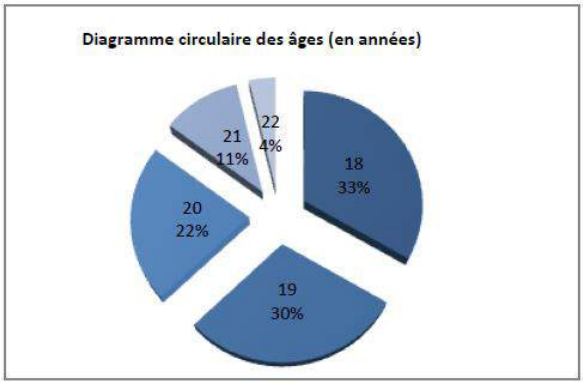
secteur	Effectif ni	fi
agriculture	200	200/650
Industrie	300	300/650
Commerce	150	150/650

- **Représentation graphique**

1- Diagramme en bâton (la valeur x selon les fréquences absolues ou relatives)







- **Variables continues**, pour lesquelles toutes les valeurs sont possibles, au moins sur un

Intervalle. Exemples : pluie journalière, débit, etc.).

le poids ou la taille.Les valeurs de la variable aléatoire x sont regroupées en classes [ci, ci+1].
Un centre de classe xi (moyenne arithmétique des deux extrémités)

L'amplitude de la classe i est ai=ci+1-ci

La fréquence relative est notée *f*i de la classe i

*f*i=ni/N

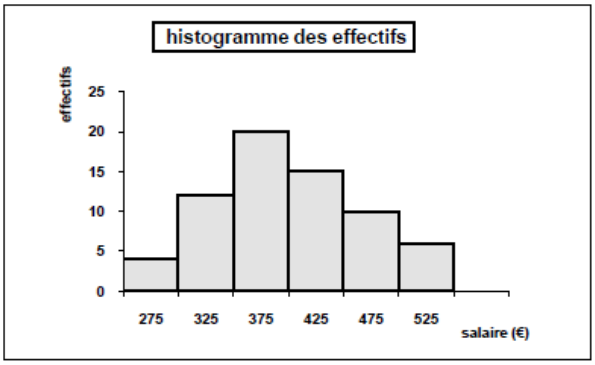
La fréquence cumulative F=$\sum_{j=1}^{i} fi$

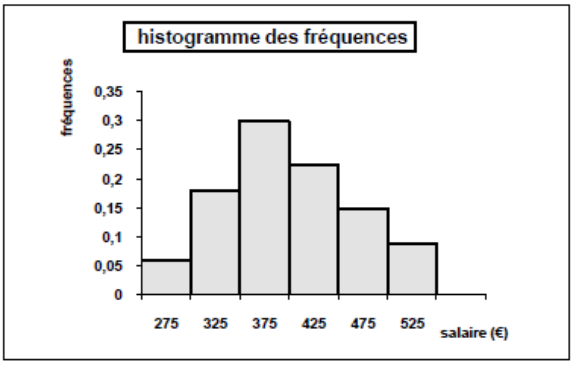
La règle de **STURGE** : $Nombre\,de\,classe = 1+(3{,}3\log n)$

La règle de **YULE :** $Nombre\,de\,classe = 2{,}5\sqrt[4]{n}$

classes	xi	ni	fi	Fi
ci	x1	n1	f1	F1
c1	x2	n2	f2	F2
c2	x2	n3	f3	F3
c3	x3	n4	f3	F3
.				
.				
.				
ck	xk	nk	fk	FK

a- Représentation graphique (histogramme)





La fréquence au non dépassement

la probabilité au non dépassement. Si X est une variable aléatoire

Continue, susceptible de prendre une valeur quelconque sur la droite

Réelle, et x une (le ces valeurs la probabilité que X soit inférieure à x

-la fréquence au non dépassement

la probabilité au non dépassement. Si X est une variable aléatoire

Continue, susceptible de prendre une valeur quelconque sur la droite

Réelle, et x une (le ces valeurs la probabilité que X soit inférieure à x

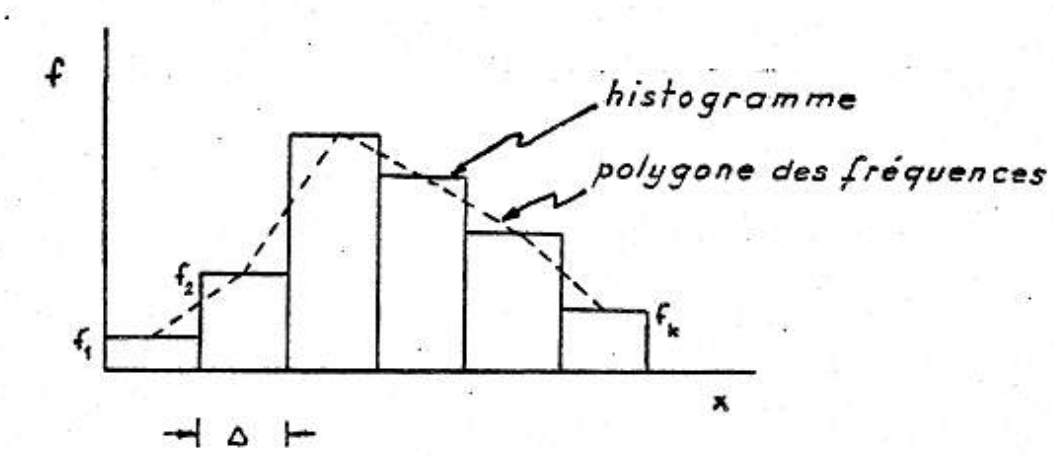
La f.d. s'appelle aussi <u>probabilité de non-dépassement</u> car la valeur de l'ordonnée, $F(x_1)$, pour une valeur x_1 de la variable, est la probabilité que la variable X soit égale ou inférieure à x_1, c'est-à-dire, $F(x_1) = P(X \leq x_1) = \int_a^{x_1} f(x)\, dx$

La fonction $\left[1 - F(x_1)\right]$ est la probabilité au dépassement en effet: $Pr\left[X > x_1\right] = 1 - Pr\left[X \leq x_1\right] = 1 - F(x_1)$

Avec:

f(x): Densité de probabilité

F(x)/Fonction de répartition



Si nous faisons l'intervalle Δ de plus en plus petit, lorsque $\Delta \to 0$ et $N \to \infty$, alors l'échantillon devient population et l'histogramme (ou le polygone des fréquences) devient une courbe continue f(x) qui s'appelle fonction de densité des probabilités (f.d.p.). L'intervalle est, dans ce cas, l'infiniment petit dx, et l'on a:

$$Pr\ [x \leq X \leq x + dx] = f(x).\, dx$$

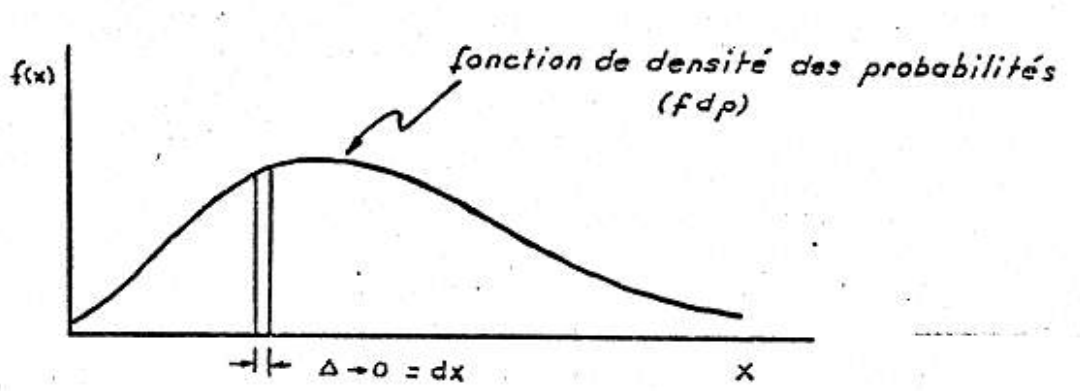


Figure 2

Application

Le résultat d'une enquête sur la taille des arbres d'espèce (x) pendent trois mois est reprise dans le tableau suivant:

Tailles (cm)	ficum (%)
[8,5; 9[	3
[9; 9,5[	10
[9,5; 9.3 [	22
[9.3; 9,5[	36
[9,5; 9.6 [	60
[9.6; 9.7 [	77
[9.7; 9.9 [	90
[9.9; 10,5[	97
[10,5; 10.2 [	100

-Trace un diagramme des fréquencescumulées.

-Détermine sur le graphique la médiane, les Troisième et sixième déciles ainsi que les Quartiles.

Bornes de classe	Fréquence absolue	Centre de classe	Fréquence relative
25.5<X<37.5	4	31.50	0.16
37.5<X<49.50	7	43.50	0.28
49.50<X<61.50	6	55.50	0.24
61.50<X<73.50	3	67.5	0.12
73.50<X<85.50	3		0.12
85.50<X<97.50	0		0
97.50<X<109.50	2		0.08

a- Variables qualitatives : Variables dont les valeurs ne sont pas numériques.

On en distingue deux types :

- **Variables ordinales**, dont les valeurs peuvent être ordonnées. Exemple : intensité d'une douleur qui peut aller de *absente* à *très intense*.
- **Variables catégorielles** ou **nominales**, dont les valeurs ne peuvent pas être ordonnées. Exemple : couleur des yeux.

6- **Modes de regroupement des unités statistiques**

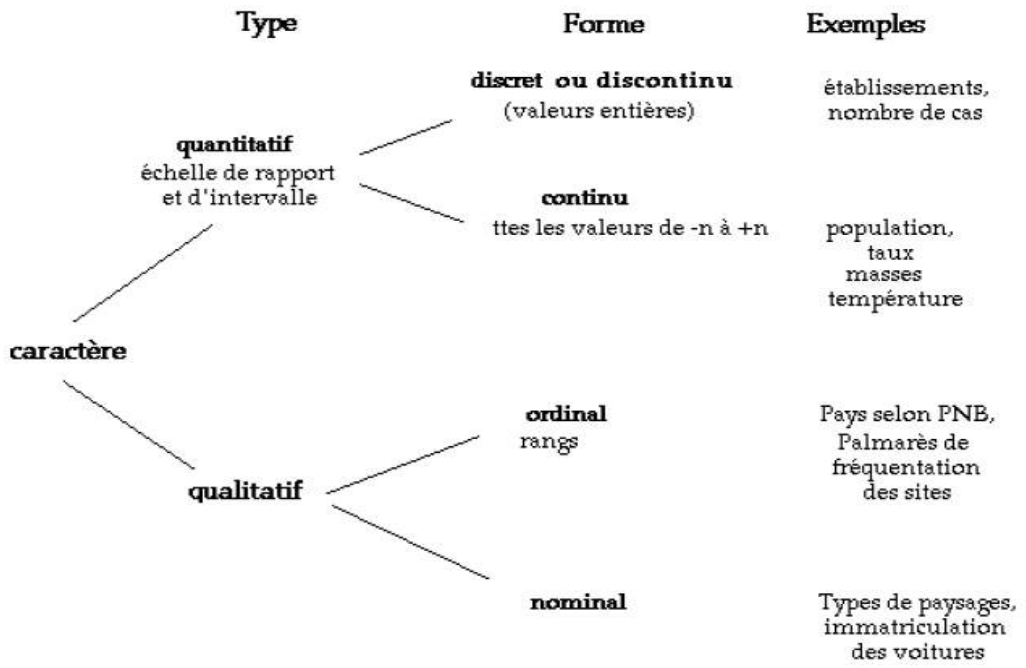
A - Série simple

B - Distribution par valeurs ou par modalités

C - Regroupement par catégories Lorsqu'il y a beaucoup de valeurs ou de modalités, on peut procéder à un regroupement par **catégories de valeurs** ou par **catégories de modalités**.

Application

Identifiez le type (et le sous type) des variables suivantes :	Réponses
a) Le nombre d'animaux par laboratoire ;	a) quantitatif discret
b) La niche écologique principale ;	b) qualitatif nominal
c) Le modèle de matériel utilisé ;	c) qualitatif nominal
d) La distance en kilomètre entre le prélèvement A et le prélèvement B	d) quantitatif continue
e) Être végétarien ou non	e) qualitatif ordinal



Quelques remarques sur les paramètres de position

- **Dans le cas d'une variable qualitative**, on se limite généralement à déterminer le mode (dans certains cas,on peut aussi déterminer la médiane). Il n'est pas possible de calculer la moyenne.

- **Pour des variables quantitatives**, on peut déterminer soit la moyenne, soit le mode, soit la médiane.

7 - Les statistiques de tendance centrale

1- Le mode

Définition

Le **mode (dominante)** d'une série est la valeur ou la modalité qui revient le plus fréquemment dans la série ou la distribution.

Exemple : Soit la série {8, 4, 4, 3, 4, 3, 8, 2,5} La valeur la plus fréquente de cette série est 4. Le mode est donc égal à 4. L'effectif associé à ce mode est 3.

1) Remarques à propos du mode

a) Une série peut avoir plusieurs modes Soit la série S = {4, 0, 1, 1, **2, 2, 2**, 3, 3, 4, **2**, 3, 4, 5, **2**, 1, 3, 3, 4, 5}, les "2" sont mis en gras et les "3" sont soulignés, car ce sont les valeurs qui reviennent le plus souvent : 5 fois chacune. Cette série a 2 modes, elle est **bimodale**. Ses deux modes sont : 2 et 3. L'effectif associé à chacun de ces modes est : 5. Bien entendu, on peut avoir des séries avec 3, 4, 5, etc. modes. Ce sont alors des **séries multimodales**.

b) Le mode n'existe pas forcément C'est le cas lorsque toutes les valeurs ont le même effectif comme dans l'exemple suivant : {8,6,5,7,3,1}. Dans ce cas, on peut aussi dire que toutes les valeurs sont modales.

c) Le mode n'est pas la valeur la plus élevée Il ne faut pas confondre le mode, qui est la valeur la plus fréquente, avec la valeur la plus élevée de la série. Dans la série {8, 6, 5, 7, 3,1}, il n'y a pas de mode, mais la valeur la plus élevée est 8. Il peut arriver que le mode soit aussi la valeur la plus élevée, mais ce n'est alors qu'une coïncidence.

Cas variable continue

La classe modèle est donc la classe [30 ; 40[

On utilise la méthode d'approximation

$M_0 = ei + a. \propto 1 /(\propto 1 - \propto 2)$

Où : ei : effectif de la classe modèle

α1 : la déférence entre la classe précédente et celle de la classe modèle

α2 : la déférence entre la classe modèle et celle de la classe suivante

a : largeur, amplitude des classes

$$Md = 4 + (\frac{0{,}500 - 0{,}48)}{0.45} 2 = 4{,}088$$

B - La moyenne arithmétique Le mot moyenne a pour origine le latin "médius», mot signifiant "qui est au milieu". "Médius" est aussi l'origine du mot "médiane". Pourtant, en

statistique, les deux mots conduisent à des définitions différentes. Ceci nous laisse supposer que la notion de milieu n'est pas toujours facile à définir.

1) **La moyenne arithmétique simple La moyenne arithmétique d'une série** ou **moyenne arithmétique simple** se calcule par une formule qui est donnée par l'expression

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2) **La moyenne arithmétique pondérée La moyenne arithmétique d'une distribution** ou **moyenne arithmétique pondérée** se calcule par une formule qui est donnée par l'expression :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

La somme varie cette fois de 1 à *k*, avec *k* qui représente le nombre de valeurs de la série. Dans le cas où aucune valeur n'est répétée $k=n$. Sinon $k<n$. Remarquons que la somme va de 1 à *k*, mais que cette somme est divisée par *n* et non par *k*. La notation *nj* représente les effectifs ou fréquences absolues des valeurs. Appliquons cette définition au calcul de la moyenne de la distribution :

x_i	n_i
0	1
1	3
2	5
3	5
4	4
5	2

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{(0\times1)+(1\times3)+(2\times5)+(3\times5)+(4\times4)+(5\times2)}{20} = \frac{0+3+10+15+16+10}{20} = \frac{54}{20} = 2,7$$

2) D’autres moyennes

a) La moyenne géométrique

C’est la moyenne applicable à des mesures de grandeurs dont la croissance est géométrique ou exponentielle.

La moyenne géométrique conserve le produit des xi: si on modifie les valeurs de deux observations tout en conservant leur produit, la moyenne géométrique sera inchangée.

La moyenne géométrique G de la série de valeurs

x, … , xi, … , xn

Supposées toutes positives (strictement), est définie ainsi

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \Rightarrow \quad \ln(G) = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Exemple

Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie suivante, ils aient été multipliés par 4 ; le coefficient multiplicateur moyen par décennie est égal à :

$$\sqrt{2 \cdot 4} = \sqrt{8} \approx 2{,}83$$

b) La moyenne harmonique

La *moyenne harmonique* est l'inverse de la moyenne arithmétique des inverses des valeurs. L'*inverse de la moyenne harmonique conserve ainsi la somme des inverses des* x_i: si on modifie les valeurs de deux observations tout en conservant la somme de leurs inverses, la moyenne harmonique sera inchangée.

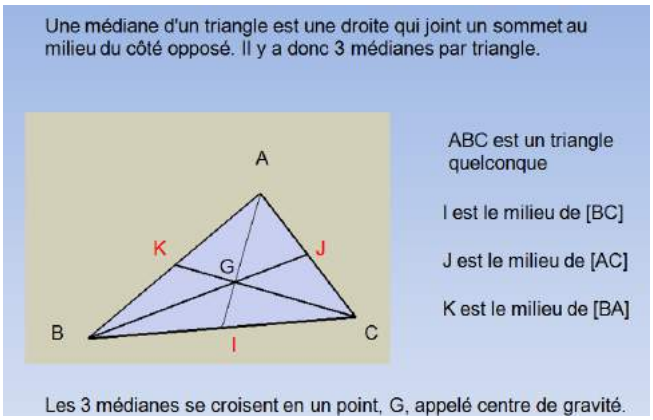
$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \text{ou} \quad H = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

Exemple

On achète des dollars une première fois pour 100€ au cours de 1,23 € le dollar, une seconde fois pour 100 € au cours de 0,97 € le dollar. Le cours moyen du dollar pour l'ensemble de ces deux opérations est égal à

$$\frac{200}{\frac{100}{1{,}23} + \frac{100}{0{,}97}} \approx 1{,}085\ €$$

Le mot « médiane » a pour origine le latin « médius», mot signifiant « qui est au milieu ». « Médius » (C'est la donnée qui permet de diviser une série ordonnée d'une façon croissante en 2 parties égales (50%, 50%).

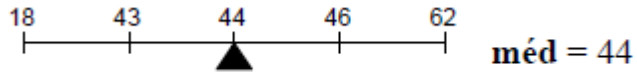


- Méthode

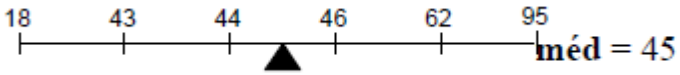
Soit une série statistique d'effectif total *n*, rangée par ordre croissant.

Pour déterminer son rang, il y a 2 cas

si *n* est impair : la médiane est la valeur de rang *n* + 1/2



si *n* est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre *n* + 1/2



6.1.3.4. Médiane, pour les données réparties par classes

Remarque

Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera **classe médiane**, la classe qui la contient (et permet donc d'en donner un encadrement).

La classe médiane est la première classe où la fréquence cumulée est supérieure à 0,500.

Application IV.3

classe	[0 ; 2[	[2 ; 4[	[4 ; 6[	[6 ; 8]
fréquence	10%	38%	45%	7%
fréquence cumulée	10%	48%	93%	100%

48% des valeurs sont strictement inférieures à 4

Et 93% des valeurs sont strictement inférieures à 6

La classe médiane est donc la classe [4 ; 6[

On peut donc en déduire l'encadrement suivant 4 ≤ méd < 6

$$Md = Bmd + (\frac{0{,}500 - Fmd_{-1})}{Fmd} Lmd$$

Où : Bmd : Borne inférieure de la classe médiane

Fmd-1 : Fréquence relative cumulée de la classe qui précède la classe médiane.

Fmd : Fréquence relative de la classe médiane.

Lmd : largeur, amplitude des classes

$$Md = 4 + (\frac{0{,}500 - 0{,}48)}{0.45} 2 = 4{,}088$$

$$\frac{Me - x_{i-1}}{x_i - x_{i-1}} = \frac{0,5 - F_{i-1}}{f_i} \quad \Rightarrow \quad Me = x_{i-1} + (x_i - x_{i-1}) \cdot \frac{0,5 - F_{i-1}}{f_i}$$

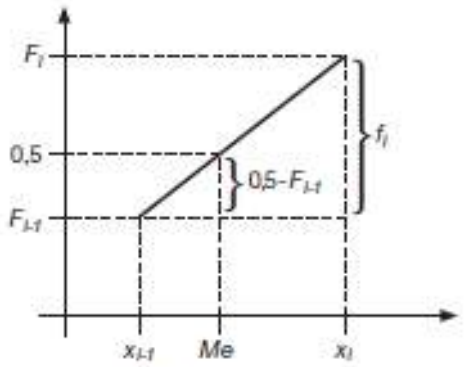


Figure 1.9 – Détermination graphique de la médiane pour une variable continue

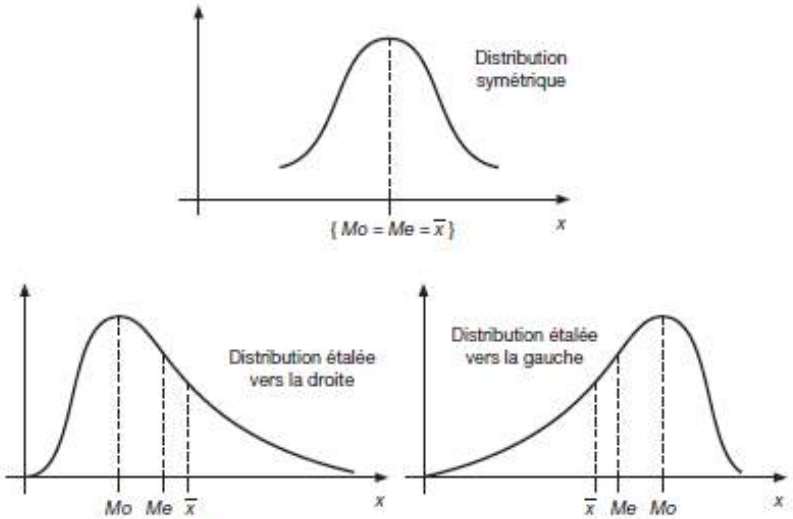


Figure 1.10 – Positions respectives du mode, de la médiane et de la moyenne

LES PARAMETRES DE DISPERSION: L'ECART MOYEN, L'ECART-TYPE ET LA VARIANCE

L'utilité des paramètres de position est d'indiquer d'une certaine manière autour de quelle valeur une série s'étend. Cependant, ce type de paramètre n'est pas suffisant pour caractériser une série.

1) **Intervalle de variation (ou « étendue ») L'intervalle de variation (IV)** ou l'étendue de la série est simplement une façon de résumer le minimum et le maximum de la série en un seul chiffre. On l'obtient ainsi : **Intervalle de variation de la série = valeur maximale – Valeur minimale**
2) **3) Rapport de variation**

   Le **rapport de variation** est simplement le rapport de la valeur maximale à la valeur minimale.

Dans le cas d'une variable quantitative, on appelle:
écart: la valeur absolue de la différence entre la moyenne et une valeur de la variable.

Exmp

Deux séries statistiques sont caractérisées par les données suivantes:
1 100 199
99 100 101
Si on calcule la moyenne de ces deux séries, on obtient:
X = *(1 + 100 + 199) / 3 = 100*
X = *(99 + 100 + 101) / 3 = 100*
Les moyennes sont les mêmes mais les séries sont très différentes. Dans la première série, la dispersion autourde la moyenne est très grande contrairement à la 2ème série !!!!
C'est pour cette raison que pour caractériser une série statistique, on doit définir des paramètres de dispersion enplus des paramètres de position. Ces paramètres de dispersion donnent des informations sur l' « étendue » de lasérie par rapport à la valeur centrale.

L'écart moyen, l'écart-type et la variance

Écart absolu moyen: la moyenne de la série des écarts des valeurs x_1,x_2…….$x_n$par rapport a une valeur central

$$e_x = \frac{1}{n}\sum_{i=1}^{n} Ixi - \bar{x}I$$

Écart absolu moyen pondéré Si les valeurs statistiques x_1,x_2…….x_n , ont des fréquences d'apparitions respectives (n1,n2…..nk)

$$e_x = \frac{1}{n}\sum_{i=1}^{n} ni.Ixi - \bar{x}I$$

$$e_x = \sum_{i=1}^{n} fi.Ixi - \bar{x}I$$

Écart absolu moyen par rapport à la médiane

$$e_x = \sum_{i=1}^{n} fi.Ixi - M_0I$$

Remarque
La variable contenue, On calcul les centre de classe

Ecart-type: la racine carrée de la moyenne des carres des écarts par apport a la moyenne.

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2}$$

.

écart-type pondéré : Si les valeurs statistiques x_1,x_2…….x_n , ont des fréquences d'apparitions respectives (n1,n2…..nk).

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ni(xi - \bar{x})^2}$$

*lorsqu'on emplois les fréquences relatives

$$\sigma_x = \sqrt{\sum_{i=1}^{n} fi(xi - \bar{x})^2}$$

La variance: est définie comme étant le carre de l'***écart-type***

$$V(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(xi - \overline{x})^2$$

La variance pondérée

$$V(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} ni(xi - \overline{x})^2$$

Ou

$$V(x) = \sigma^2 = \sum_{i=1}^{n} fi(xi - \overline{x})^2$$

Remarque

- Pour N >30 : on utilise N

- Pour N ≤ 30 : on utilise N-1

Le coefficient de variation :

Pour obtenir un nombre abstrait indépendante des unités de mesure et permettant de comparer les disparitions dans différents séries hydrologiques. On utilise le coefficient de variation qui est représenté par le rapport entre l écart-type et la moyenne : On définit aussi :

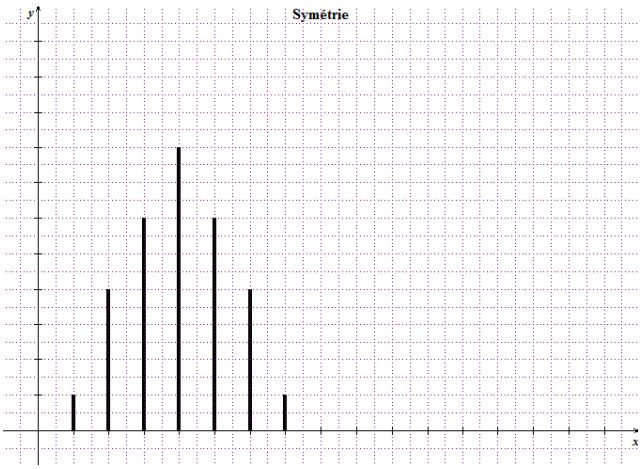
$Cv = \frac{\sigma}{\bar{x}}$ Qui compare donc la fluctuation à la valeur moyenne **(Bois Ph. & al., 2007)**.

Les paramètres de formes

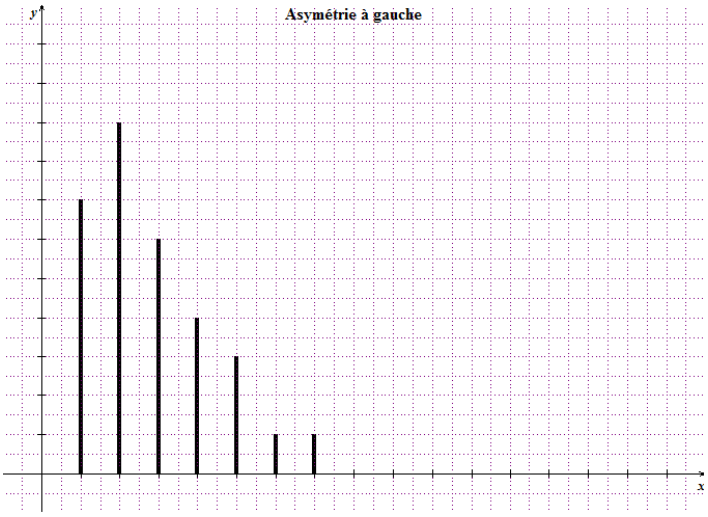
Les moments centrés

Pour comprendre la notion de symétrie et d'asymétrie, il faut faire appel aux représentations graphiques (ici, le diagramme en bâtons).

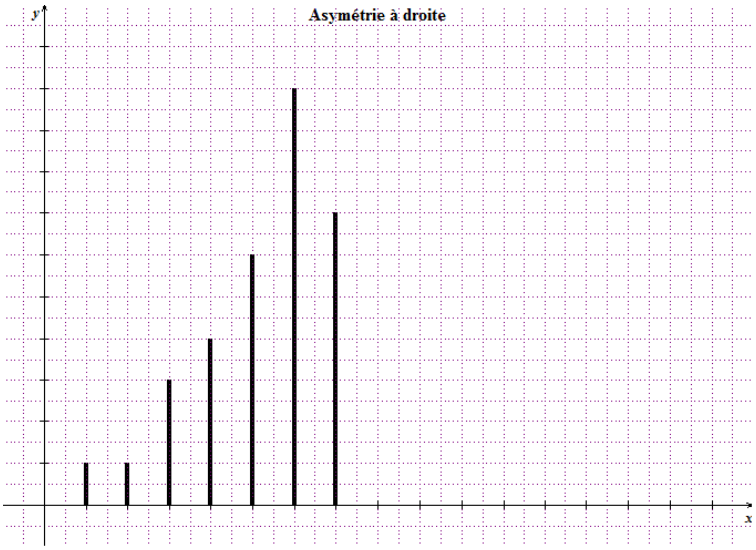
Une distribution de valeurs peut être symétrique, asymétrique à gauche, asymétrique à droite.



Dans ce cas on constate en général que la moyenne est égale à la médiane et aussi au mode.



Dans ce cas on constate généralement que la moyenne est supérieure à la médiane qui elle-même est supérieure au mode.



Dans ce cas on constate généralement que la moyenne est inférieure à la médiane qui elle-même est inférieure au mode.

Il peut être utile de quantifier l'asymétrie et non pas seulement de la constater.

C'est l'objet de ce qui suit.

Quand on connait les valeurs de la série statistique, on peut définir les moments centrés.

Le moment centré d'ordre p est :

$$Si\ p\ \in N, m_p = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i - \bar{x})^p$$

Nous connaissons le moment centré d'ordre 1 :

$$m_1 = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i - \bar{x}) = 0$$

Nous connaissons aussi le moment centré d'ordre 2 : $m_2 = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i - \bar{x})^2 = S^2$, c'est la variance.

Pour quantifier l'asymétrie, nous utiliserons le moment centré d'ordre 3 : $m_3 = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i - \bar{x})^3$.

Le coefficient d'asymétrie de Fischer

Il permet de déterminer le sens de l'asymétrie et de quantifier sa valeur.

$$\gamma_1 = \frac{m_3}{S^3}$$

"Gamma 1" est le quotient du moment centré d'ordre 3 par le cube de l'écart type.

Le signe de "gamma 1" est donc égal à celui du moment centré d'ordre 3 car l'écart -type est positif.

Numérateur et dénominateur s'expriment avec une unité qui est le cube de l'unité de la variable. Le quotient est donc sans unité : c'est ce que nous voulions.

Gamma 1 est nul si la distribution est symétrique .

Si la distribution est **asymétrique à gauche**, **Gamma 1 est positif**.

Si la distribution est **asymétrique à droite**, **Gamma 1 est négatif**.

Le signe de Gamma 1 indique donc le sens de l'asymétrie.

De plus, si **Gamma 1 augmente en valeur absolue**, cela veut dire que **la distribution devient de plus en plus asymétrique**.

Le coefficient d'asymétrie :

Ce coefficient est le troisième moment centré autour de la moyenne. Comme son nom l'indique, ce coefficient mesure la symétrie de la distribution par rapport à la moyenne :

Le *coefficient d'asymétrie de Fisher*, noté ©1, est ainsi défini :

$$Sk = \frac{1}{\sigma^3} \frac{\sum (x_i - \mu)^3}{n}$$

Comme tout coefficient d'asymétrie, il est *nul* pour une distribution *symétrique*, négatif pour une distribution unimodale étalée vers la gauche, positif

pour une distribution unimodale étalée vers la droite (figure 1.12).

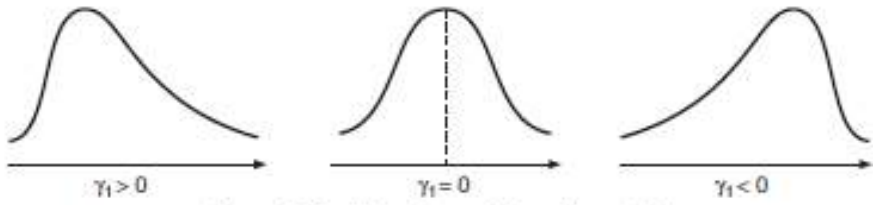


Figure 1.12 – Signe du coefficient d'asymétrie

Ce coefficient est nul pour une distribution parfaitement symétrique. Il est positif ou négatif en fonction de la position de la moyenne par rapport à la pointe de la distribution.

e- Le coefficient d'aplatissement :

3) L'aplatissement

Les coefficients d'aplatissement mesurent l'aplatissement d'une distribution ou l'importance des « queues » d'une distribution. Le *coefficient d'aplatissement de Fisher,* noté ©2 , est ainsi défini :

Ce coefficient est *nul* pour une *distribution normale* (chapitre 7), positif ou négatif selon que la distribution est plus ou moins aplatie que la distribution normale de même moyenne et de même écart-type.

Ces coefficients d'asymétrie et d'aplatissement sont invariants par changement d'origine et d'échelle, mais ils sont sensibles aux fluctuations d'échantillonnage puisqu'ils font intervenir des moments d'ordre élevé.

$$K = \frac{1}{\sigma^4} \frac{\sum (x_i - \mu)^4}{n} - 3$$

Avec : **(V.10)** m4 : moment centré d'ordre 4

m2 : le moment centré d'ordre 2 se confond avec la variance m2 = *S*2 **(Lang M. & Lavabre J., 2007).** Il peut être calculé aussi par :

$CK = Xi -)4Ni = 1(N-1)S4$ **(V.11)** Si ck est négatif, la distribution est plus aplatie que la distribution normale ; la distribution est dite platicurtique. Si ck est positif, la distribution est moins aplatie que la distribution normale ; la distribution est dite leptocurtique. Si ck = 0, l'aplatissement est le même que pour la loi normale et la courbe est dite mésocurtique. Les caractéristiques des stations hydrométriques sont représentées dans le tableau

Partons de l'exemple suivant : dans un carré de haricots, on a récolté 140 gousses (et non pas cosse qui est l'enveloppe du pois) et on a compté le nombre de grains dans chacune des gousses.
Voici le tableau de résultats.

Nombre de grains (variable discrète) X_i	1	2	3	4	5	6	7	8	9	10
Nombre de gousses effectifs n_i	2	6	9	18	32	38	20	7	6	2
Fréquences f_i	0,014	0,043	0,064	0,129	0,229	0,271	0,143	0,05	0,043	0,014

L'effectif total vaut: n = somme n_i = 140

La moyenne vaut:=5,51

Pour évaluer la dispersion autour de cette moyenne, l'idée qui vient spontanément à l'esprit est de déterminer les écarts entre cette moyenne et les diverses valeurs de la variable.
On appelle ***écart*** entre la valeur moyenne X et un nombre X, la valeur absolue de leur différence. On obtient pour la série étudiée:

Nombre de grains (variable discrète) X_i	1	2	3	4	5	6	7	8	9	10
Fréquences f_i	0,014	0,043	0,064	0,129	0,229	0,271	0,143	0,05	0,043	0,014
$\lvert X_i - \overline{X} \rvert$	4,51	3,51	2,51	1,51	0,51	0,49	1,49	2,49	3,49	4,49
$f_i \cdot \lvert X_i - \overline{X} \rvert$										

Pour les ***variables statistiques continues***, la valeur médiane *Me* est telle que ***F*(*Me*) = 50%**. On commence par chercher la *classe médiane* à l'aide des fréquences cumulées, la classe médiane [*xi* – 1 , *xi*[étant telle que *F*i –1 < 50% et ***Fi*** > 50%. La valeur de la médiane s'obtient ensuite par *interpolationlinéaire* en raison de l'hypothèse d'équirépartition à l'intérieur des classes. Cette détermination peut se faire par le calcul ou graphiquement

(*cf.* figure 1.9)

6.1.1. Le mode, ou valeur dominante, est la valeur la plus fréquente d'une distribution. Cette valeur se calcule toujours à partir d'un dénombrement des modalités du caractère. Il faut donc

distinguer le cas des caractères discrets et des caractères continus (voir notions de bases).

*** Caractère qualitatif et caractère discret** :

le mode est la modalité ou la valeur qui a **la fréquence simple la plus élevée** (ou l'effectif le plus élevé, ce qui revient au même).

*** Caractère quantitatif continu :** . Le mode est alors le centre de la classe modale, c'est à dire de la classe qui a **la fréquence moyenne la plus élevée**.

Application III. 1 : Cas de calcul des modes :

- Cas 1 : Données rangées : le mode est la valeur de la donnée qui apparaît le plus fréquemment (celle qui a le plus d'occurrences) :

140 ; 141 ; 144 ; 144 ; 148 ; 148 ; 152 ; 152 ;152 ; 154 ; 155 ; 158 ; 158 ; 161 ; 170 ; 172

Le mode est 152 car il possède le plus grand nombre d'occurrences (il est référencé 3 fois)

- Cas 2 : Données condensées : le mode est la valeur de la donnée qui possède la fréquence la plus élevée (relative ou absolue).

Modalités xi (age en années)	**14**	**16**	**18**	**21**	**22**	**24**	**25**	**Total**
Fréquences absolues	5	**12**	10	8	11	7	3	56
Fréquences relatives	0,089	**0,214**	0,179	0,143	0,196	0,125	0,054	1,000

Dans cette série statistique, le mode est égal à Mo = 16 ans

Dans le tableau des classes relatives à la longueur de la rectrice de Bonasa *umbellus* , la classe modale est [155mm-160mm[. Il est possible de calculer de façon plus précise le mode en appliquant la formule suivante :

$$Mo = bmo + (\frac{\Delta 1}{\Delta 1 + \Delta 2})Lmo$$

bmo : Borne inférieur de la classe modale
Lmo : largeur de la classe modale

Δ1 = différence entre l'effectif de la classe modale et l'effectif de la classe précédente.

Δ2 = différence entre l'effectif de la classe modale et l'effectif de la classe qui suit.

Δ1 = (17-9) = 8 ; **Δ2** = (17-16) = 1 ; **bmo** = 155 ; **Lmo** = 5

$$Mo = 155 + (\frac{8}{8+1})5 = 159 \text{ mm}$$

b) Les quantiles

Les*Quantiles* sont des *indicateurs de position*

Quartiles Les **quartiles** sont les **trois valeurs** qui partagent la population, dont les unités statistiques ont préalablement été classées par ordre croissant de valeurs (de la variable considérée), en **quatre sous populations** de même taille. On les désigne respectivement par Q1, Q2 et Q3.

.

Le *quantile d'ordre α* *(0<α<1)*noté

x_α est tel qu'une proportion α des individus ait une valeur du caractère X inférieure ou égale à x_α

Le quantile*x*0,5 est égal à la médiane.

On utilise couramment les quantiles d'ordre 1/4, 1/2 et 3/4. Ils sont ainsi notés et nommés :

*Q1= premier quartile = x*0,25

*Q*2= *deuxième quartile = médiane = x*0,5
*Q*3 = *troisième quartile = x*0,75
Les quartiles se déterminent, comme la médiane, à l'aide de la profondeur (variable discrète), ou à l'aide des fréquences cumulées (variable continue)

Dans le cas d'une variable statistique *continue,* on a *F(Q*1*)* = 0,25 et
*F(Q*3*)* = 0,75 et on calcule les quartiles par *interpolation linéaire,* en raison
de l'hypothèse d'équirépartition. Pour la distribution de l'ancienneté du chômage des femmes (*cf.*

figure 1.5) : $Q_1 = 1 + 2 \cdot \frac{25 - 16,8}{19} \approx 1,9 \text{ mois}$ $Q_3 = 12 + 12 \cdot \frac{75 - 68,7}{18,5} \approx 16,1 \text{ mois}$

On peut définir à partir des quartiles *Q*1 et *Q*3 le paramètre de tendance
centrale (*Q*1 + *Q*3)/2, égal à la médiane dans le cas d'une distribution symétrique,

3) Intervalle interquartile L'intervalle interquartile (IIQ) est la différence entre le troisième quartile et le premier quartile. Il s'écrit : *IIQ = Q3 - Q1*

L'intervalle interquartile sert à apprécier la dispersion de la série, de façon absolue, ou bien par comparaison avec une autre série (à condition que les valeurs de l'autre série soient exprimées dans la même unité).

Variance, écart-type et coefficient de variation Ces trois statistiques sont liées entre elles. Elles sont toutes les trois des indicateurs de la dispersion d'une série par rapport à sa valeur moyenne. Le plus simple est de commencer par l'étude de la variance.
1) La variance La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. De même que la moyenne, elle se résume à un seul chiffre qui s'obtient par un calcul que nous allons décomposer ci-après.
a) Définition

IV – COMPARAISON DE MOYENNES : ANOVA A UN FACTEUR

L'objectif de l'analyse de variance à 1 facteur est de tester l'égalité des moyennes théoriques d'une variable quantitative de différents groupes ou de différents niveaux du facteur considéré.

Ces populations sont en général des variantes (ou niveaux **k**) d'un ou plusieurs facteurs contrôlés de variation (facteurs **A**, **B**, ...).

Les observations sont sous la forme :

Groupe 1	.....	Groupe K
$x_{1,1}$	.....	$x_{K,1}$
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
x_{1,n_1}	.....	x_{K,n_K}
T_1		T_K
$m_1 = \frac{T_1}{n_1}$		$m_K = \frac{T_K}{n_K}$

$$N = \sum_{i=1}^{K} n_i \qquad N = \sum_{i=1}^{K} T_i$$

Conditions d'applications de l'ANOVA

- les populations étudiées suivent une distribution normale
- les variances des populations sont toutes égales (**HOMOSCEDASTICITE**)
- les échantillons E_i de tailles n_i sont prélevés aléatoirement et indépendamment dans les populations.

Procédure de calcul d'une ANOVA

- Déterminer si les échantillons varient de la même manière.
- Si nous démontrons l'homogénéité des variances, alors nous pouvons comparer les moyennes de ces échantillons.

Problèmes liés à l'égalité des variances

Test de l'homogénéité des variances

- $(H_0)_\sigma$: les variances sont homogènes
- $(H_1)_\sigma$: Au moins une des variances est différente des autres

→Utilisation d'un test de comparaison de plusieurs variances

Conclusion

- Si $(H_0)_\sigma$ est rejetée : il est théoriquement impossible de comparer des échantillons qui ne varient pas de la même manière.
- Si $(H_0)_\sigma$ n'est pas rejetée : par conséquent, il est possible de comparer les moyennes de tels échantillons

- ▪ Modèle:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1,\ldots I \text{ et } j = 1,\ldots, J$$

Variabilité totale

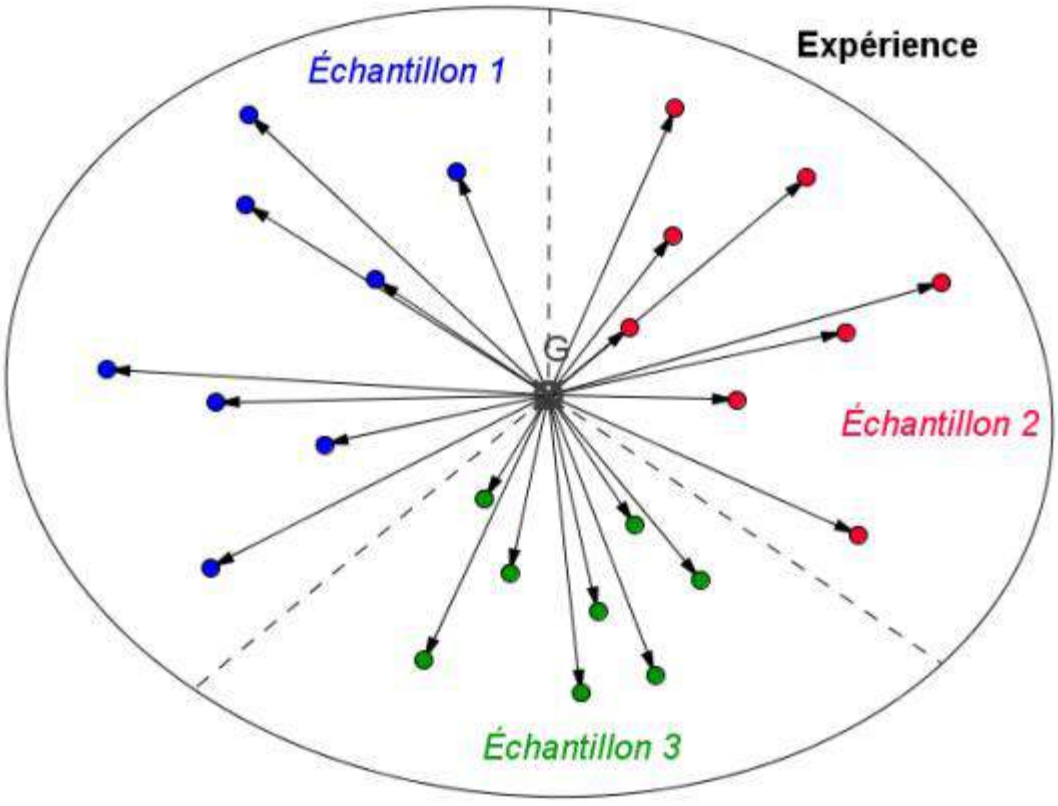


Figure 39 : Variabilité totale (tous les échantillons confondus)

- Variabilité totale au sein de l'expérience (quel que soit l'échantillon) : reflète les écarts de tous les individus par rapport à la moyenne générale (**G**) de l'expérience.
- Calcul de la Somme des Carrés des Écarts à la moyenne totale (SCE_T).
- Degrés de liberté (**DDL**) associés : **N-1**.

Variabilité factorielle

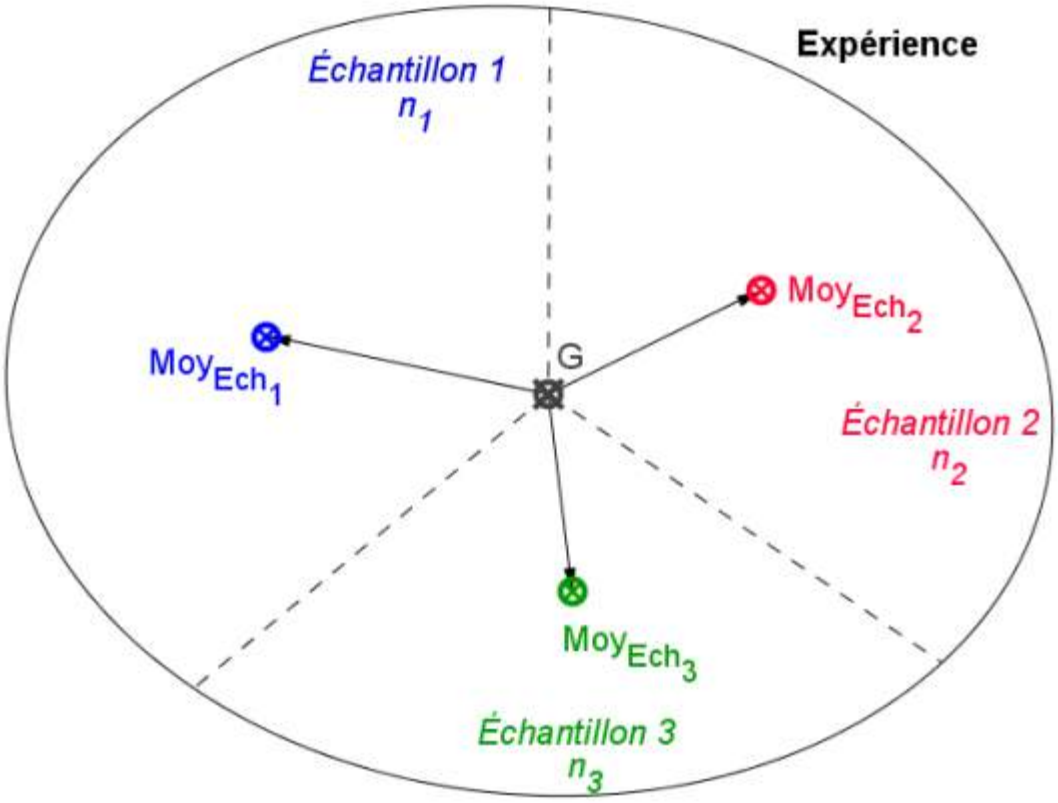


Figure 40 : Effet du facteur étudié sur les moyennes des échantillons par rapport à la moyenne générale

- Variabilité factorielle : reflète les écarts des moyennes des échantillons (supposées influencées par le facteur étudié) par rapport à la moyenne générale (**G**) de l'expérience.
- Calcul de la Somme des Carrés des Écarts à la moyenne factorielle (SCE_F).
- **DDL** associés : **k-1**.

Variabilité résiduelle

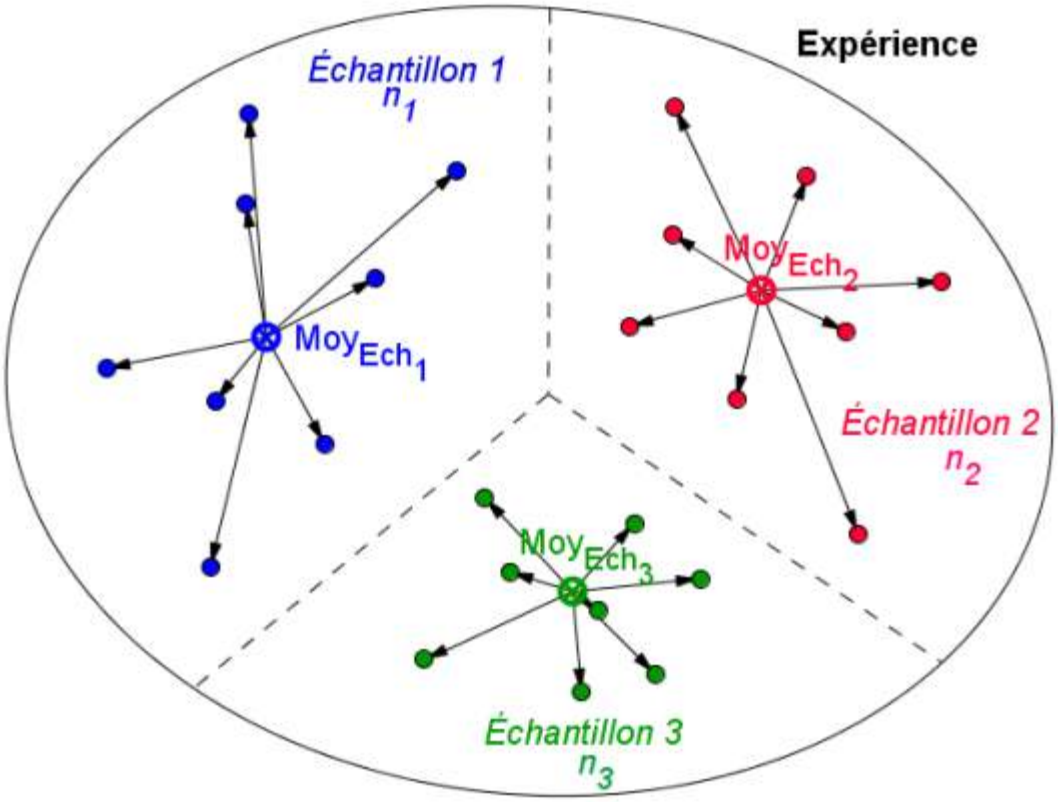


Figure 41 : Variabilité intragroupe (résiduelle)

- Variabilité résiduelle (liée à l'individu) : reflète l'importance des variations individuelles dans chaque échantillon.
- Calcul de la Somme des Carrés des Écarts à la moyenne résiduelle (SCE_R).
- **DDL** associés : **N-k**.

Bilan

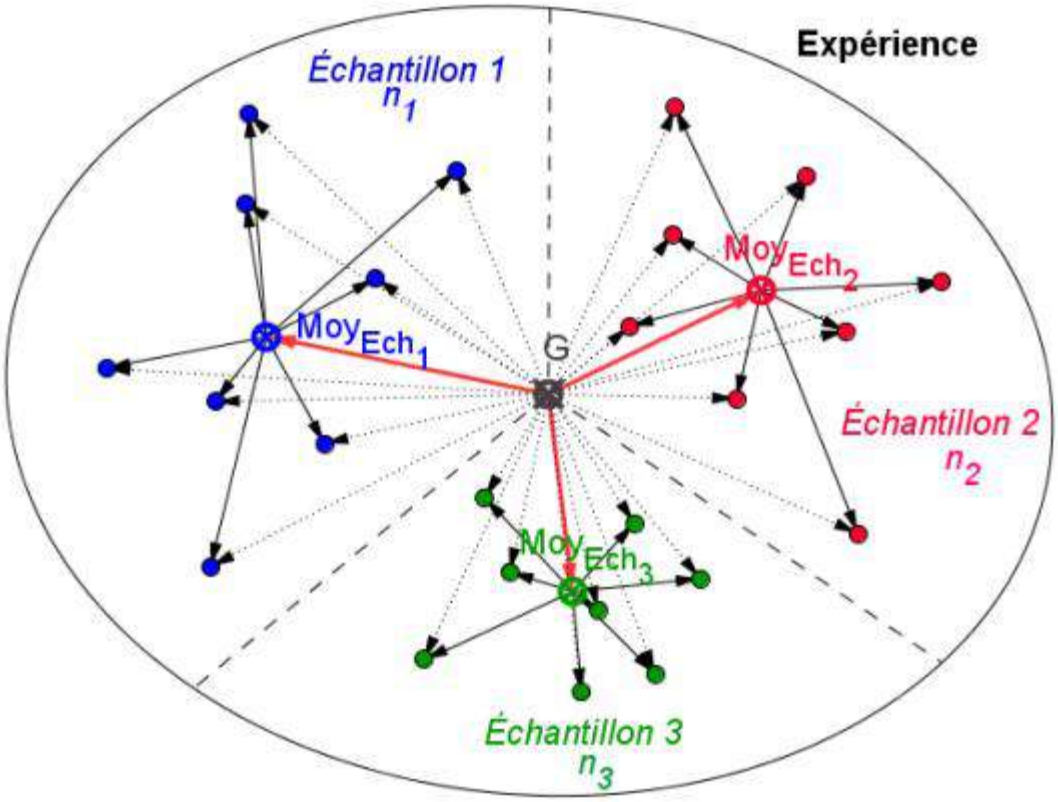


Figure 42 : Représentation combinée de toutes les sources de variabilités

Pour résumer :

- $SCET = SCEF + SCER$
- **DDL** associés : **N-1 = k-1 + N-k**.

$$\text{variance}_{\text{totale}} \equiv \text{variance}_{\text{entre les groupes}} + \text{variance}_{\text{à l'intérieur des groupes}}$$

variance$_{totale}$: différence entre valeurs et moyenne globale

$$SCE^*_{totale} = \sum_{i}^{n}(x_i - \bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$$

variance$_{\text{entre les groupes}}$: différence entre moyennes des groupes et moyenne globale

$$SCE_{inter\ groupe} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2$$

variance$_{\text{à l'intérieur des groupes}}$: différence entre valeurs et moyennes des groupes

$$SCE_{intra\ groupe} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$$

**SCE : Somme des Carrés des Ecarts à la moyenne*

La variance est la moyenne des carrés des écarts :

estimation de la variance$_{\text{totale}}$ = $s^2_{totale} = \frac{SCE_{totale}}{ddl^*_{total}} = \frac{SCE_{totale}}{n-1}$

estimation de la variance$_{\text{entre les groupes}}$ = $s^2_{intra} = \frac{SCE_{inter}}{ddl_{inter}} = \frac{SCE_{inter}}{k-1}$

estimation de la variance$_{\text{à l'intérieur des groupes}}$ = $s^2_{inter} = \frac{SCE_{intra}}{ddl_{intra}} = \frac{SCE_{intra}}{n-k}$

$$SCE_{totale} = SCE_{inter} + SCE_{intra}$$

**ddl : nombre de degrés de liberté*

- On comparera les variabilités factorielle $s_F^2 = \frac{SCE_F}{k-1}$ et résiduelle $s_R^2 = \frac{SCE_R}{N-k}$

Variabilités : comparaison variation factorielle – variation résiduelle

$$e_T = e_B + e_W$$
$$y_{ik} - \overline{y} = (\overline{y_i} - \overline{y}) + (y_{ik} - \overline{y_i})$$

$$SCE_T = SCE_B + SCE_W$$
$$\sum_{ik} (y_{ik} - \overline{y})^2 = \sum_{ik} (\overline{y_i} - \overline{y})^2 + \sum_{ik} (y_{ik} - \overline{y_i})^2$$

Exemple

	Machine 1	Machine 2	Machine 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\overline{X}_i$	$\overline{X}_1 = 49$	$\overline{X}_2 = 56$	$\overline{X}_3 = 51$	$\overline{\overline{X}} = 52$

$H_0 : \mu_1 = \mu_2 = \mu_3$

	Machine 1	Machine 2	Machine 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\bar{X}_i$	$\bar{X}_1 = 49$	$\bar{X}_2 = 56$	$\bar{X}_3 = 51$	$\bar{\bar{X}} = 52$
$\bar{X}_i - \bar{\bar{X}}$	-3	4	-1	$\Sigma(\bar{X}_i - \bar{\bar{X}}) = 0$
$(\bar{X}_i - \bar{\bar{X}})^2$	9	16	1	$\Sigma(\bar{X}_i - \bar{\bar{X}})^2 = 26$

$H_0 : \mu_1 = \mu_2 = \mu_3$

Variance expliquée

$$S^2_{\bar{X}} = \frac{26}{(3-1)} = 13$$

$$\sum_{j=1}^{n} (X_{1j} - \bar{X}_1)^2 = (47-49)^2 + (53-49)^2 + (49-49)^2 + (50-49)^2 + (46-49)^2 = 30$$

$$S_p^2 = \frac{(n-1)\sigma_{o1}^2 + (n-1)\sigma_{o2}^2 + \ldots + (n-1)\sigma_{ok}^2}{k(n-1)}$$

D'où $$S_p^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2}{k(n-1)}$$ mais aussi : $$S_p^2 = \frac{\sigma_{o1}^2 + \ldots + \sigma_{ok}^2}{k}$$

Ici : $$S_p^2 = \frac{30+50+14}{(4+4+4)} = 7.83$$

Facteur machine	Machine 1	Machine 2	Machine 3				
	47	55	54				
	53	54	50				
	49	58	51				
	50	61	51	k	3		
	46	52	49	n	5		
$\bar{X}_i$	49.00	56.00	51.00	$\bar{\bar{X}}$	52.00		
				S_x^2	13.00	Facteur A	Variance Expliquée inter échantillon
				nS_x^2	65.00		
σ_{oi}	7.5	12.5	3.5	S_p^2	7.83	Résiduelle (erreur)	Variance (commune) inexpliquée intra échantillon
$1/(n-1)\Sigma_j(X_{ij} - \bar{X}_i)^2$							

ANOVA à un facteur - Conclusion

Théorème d'analyse de la variance :

Variation	SCE	ddl	CM	F
Factorielle	$SCE_F=(k-1)\cdot s_F^2$	k-1	$CM_F=s_F^2$	$F=\frac{s_F^2}{s_R^2}$
Résiduelle	$SCE_R=(n-k)\cdot s_R^2$	n-k	$CM_R=s_R^2$	
Totale	$SCE_T=(n-1)\cdot s^2$	n-1	$CM_T=s^2$	

Le tableau d'Analyse de Variance est le suivant :

Origine	Σ des carrés (a)	ddl (b)	Variance (a)/(b)	F
Intergroupe	$\sum_{i=1}^{K}\frac{T_i^2}{n_i}-\frac{T_g^2}{N}$	K-1	S^2	$\frac{S^2}{s_e^2}$
Intragroupe	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}x_{ji}^2-\sum_{i=1}^{K}\frac{T_i^2}{n_i}$	N-K	s_e^2	
Totale	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}x_{ji}^2-\frac{T_g^2}{N}$	N-1		

Tableau 32 : Tableau d'ANOVA

$$SCE_T = SCE_F + SCE_R$$

Sous l'hypothèse H_0 :

- F suit une loi de Snédécor à $\nu_1 = k-1$ et $\nu_2 = n-k$ ddl
- (test unilatéral : le rapport n'est pas obligatoirement supérieur à 1)

Choix du risque

- Risque de première espèce α (erreur commise lorsqu'on rejette H_0 à tort).

Décision

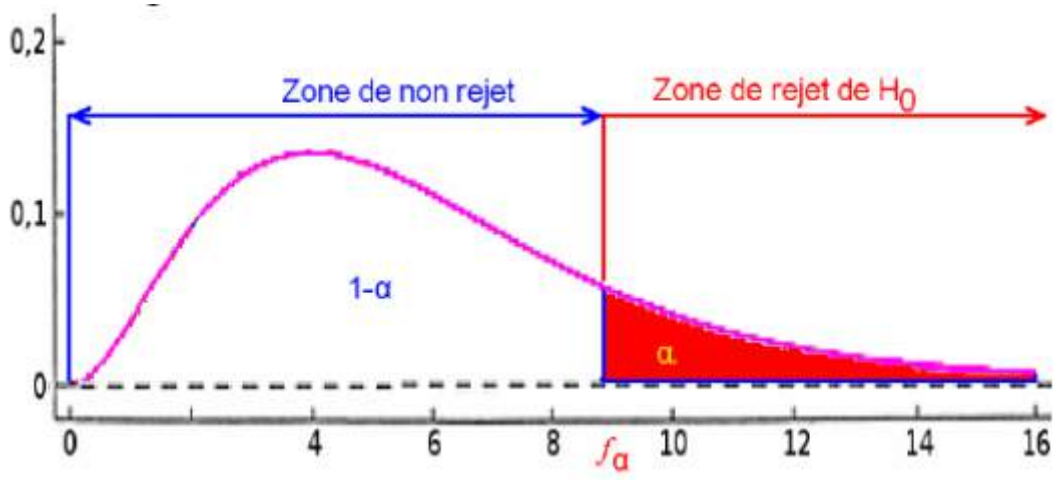
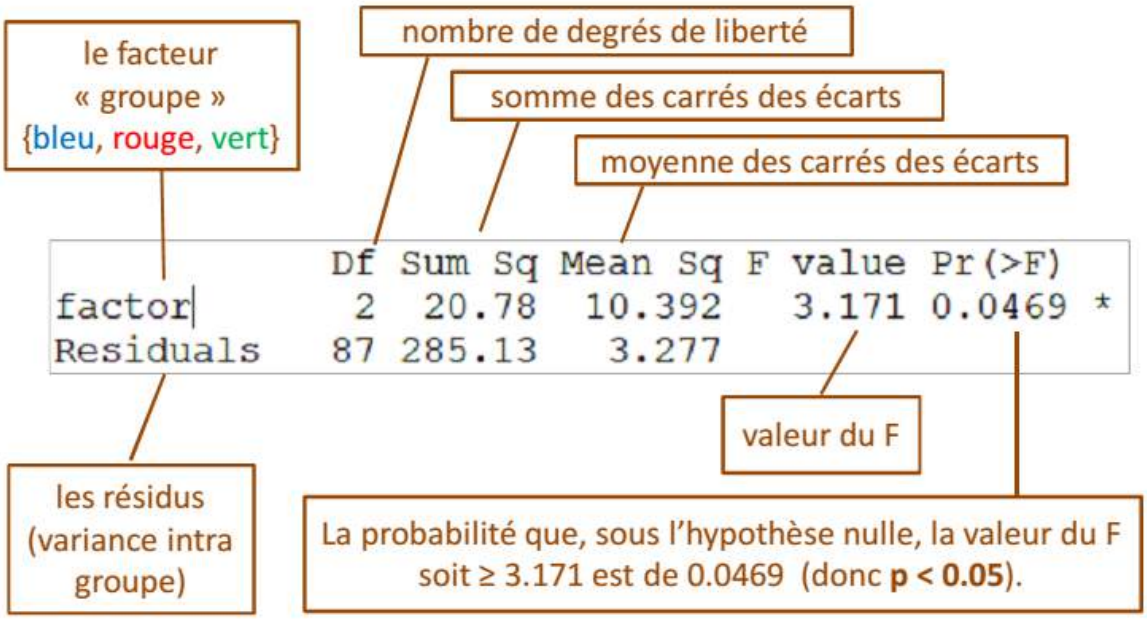


Figure 46 : Zones de rejet de l'hypothèse nulle pour une distribution de Snédécor et un test unilatéral

- Si $F > f_\alpha$ => **rejet de H_0 au risque α** :
    - La variance factorielle est significativement supérieure à la variance résiduelle : les moyennes diffèrent significativement entre-elles.

  → on attribue une influence significative au facteur étudié.
    - Recherche du degré de signification p (recherche du risque α le plus petit possible pour conclure au rejet de H_0)
- Sinon rien ne permet de dire que les moyennes des populations ne sont pas égales => H_0 **n'est pas rejetée**.

ANOVA des données de notre exemple :



Conclusion : rejet de l'hypothèse nulle *(au risque α)*, trop improbable.
Les moyennes des trois groupes sont significativement différentes.

ANOVA à deux facteurs - Introduction

Définition

- Étude simultanée d'un facteur **A** à **p** modalités et d'un facteur **B** à **q** modalités.
- Pour chaque couple de modalités **(A, B)** :
  - On a un échantillon E_{ij} ($i \in [1;p]$ et $j \in [1;q]$).
  - Tous les E_{ij} sont de mêmes tailles n .

AttentionConditions d'applications de l'ANOVA

- les populations étudiées suivent une distribution normale
- les variances des populations sont toutes égales (**HOMOSCEDASTICITE**)
- les échantillons E_i de tailles n_i sont prélevés aléatoirement et indépendamment dans les populations.

Procédure de calcul d'une ANOVA

- Déterminer si les échantillons varient de la même manière.
- Si nous démontrons l'homogénéité des variances, alors nous pouvons comparer les moyennes de ces échantillons.

Problèmes liés à l'égalité des variances

Test de l'homogénéité des variances

- $(H_0)_\sigma$: les variances sont homogènes
- $(H_1)_\sigma$: Au moins une des variances est différente des autres

→ utilisation d'un test de comparaison de plusieurs variances

Conclusion

- Si $(H_0)_\sigma$ est rejetée : il est théoriquement impossible de comparer des échantillons qui ne varient pas de la même manière.
- Si $(H_0)_\sigma$ n'est pas rejetée : par conséquent, il est possible de comparer les moyennes de tels échantillons

ANOVA à 2 facteurs (2 modalités) - 2 ways ANOVA (2 levels)

$$SCE_{totale} = SCE_{inter} + SCE_{intra}$$

$$SCE_{Totale} = SCE_{Factorielle} + SCE_{Résiduelle}$$

$$SCE_{Factorielle} = SCE_{Facteur\ A} + SCE_{Facteur\ B} + SCE_{Interaction\ AB}$$

Fiche 16 – Comparaison de deux variances : Test F

IV – COMPARAISON DE MOYENNES : ANOVA A UN FACTEUR

Les données et objectifs :

Cette technique s'applique à des tableaux décrivant pour chaque individu une variable quantitative Y en fonction d'un facteur. On appelle facteur une variable qualitative prenant plusieurs modalités donton étudie l'influence sur la variable Y. Par exemple, le facteur peut être la variété, le dosage d'unapport nutritif, le type d'engrais, un traitement …

Le tableau:

Y	Facteur	Facteur est une colonne déclarée en facteur
y_{11}	A	
y_{12}	A	
$y_{2\,k}$	B	k représente la répétition de la mesure pour la modalité 2

L'objectif est d'évaluer si le facteur influence significativement la variable Y.
Pour tester l'hypothèse nulle H_0 "toutes les moyennes sont égales", on a le plus souvent recours àl'analyse de variance (ANOVA) développée par Fischer (1890-1962).

Le modèle linéaire pour un facteur

En présence d'un seul facteur, on considère que la variable Y suit pour chaque modalité i une loi normale $\mathbf{N(\mu_i,\sigma^2)}$

Objectif : L'objectif du test est de montrer l'existence de différences significatives entre les moyennes.

Hypothèse nulle : H_0 « les moyennes sont toutes égales » contre H « les moyennes ne sont pas toutes égales ».

Il est important de comprendre que l'ANOVA n'est pas un test permettant de « classer » des moyennes. On étudiera certains tests dits « tests de comparaison multiples » permettant de répondre à ce problème au paragraphe III.

Principe du test :

L'écart total e_T se décompose en un écart expliqué par le modèle, e_B et un écart résiduel e_W, soit :

$$\begin{array}{ccccc} e_T & = & e_B & + & e_W \\ y_{ik} - \overline{y} & = & (\overline{y_i} - \overline{y}) & + & (y_{ik} - \overline{y_i}) \end{array}$$

On utilise l'écriture anglosaxonne avec :
B pour between groups (entre groupes)W pour within group (intra groupe)

On montre que la somme des carrés des écarts se décompose en une somme intra et inter groupes

$$\begin{array}{ccccc} SCE_T & = & SCE_B & + & SCE_W \\ \sum_{ik} (y_{ik} - \overline{y})^2 & = & \sum_{ik} (\overline{y_i} - \overline{y})^2 & + & \sum_{ik} (y_{ik} - \overline{y_i})^2 \end{array}$$

En notant SCE_T la somme des carrés des écarts totaux, SCE_B la somme des carrés des écarts intergroupeset SCE_W la somme des carrés des écarts intra-groupe.
On obtient les différentes variances, ou carrés moyens, en divisant les sommes de carrés d'écart par leurs degrés de liberté. Total: $n-1$ Inter : $q-1$ Intra : $n-$ q
On vérifie que $q-1+n-$ q = $n-1$. Les degrés de liberté se décomposent de manière additive commeles sommes de carrés d'écarts. On obtient alors les carrés moyens (mean squares) ou variances par les formules suivantes :

$$CM_T = \frac{SCE_T}{n-1} \qquad CM_B = \frac{SCE_B}{q-1} \qquad CM_W = \frac{SCE_W}{n-q}$$

avec n l'effectif total et $q$le nombre de modalités.

On montre alors que la statistique

$F = \frac{CM_B}{CM_W}$ suit la loi de Fisher à (q-*1;n-q*) ddl sous H_0.

Test : On teste H_0 « les moyennes sont toutes égales » contre H « les moyennes ne sont pas toutes Egales

- si $F < F_{1-\alpha}(q\text{-}1, n\text{-}q)$, on accepte H_0
- sinon on rejette H_0 avec un risque de première espèce égal à α (ou p).

V – COMPARAISONS DE MOYENNES : ANOVA à 2 FACTEURS

Objectif et données : On étudie maintenant une variable quantitative ***Y*** en fonction de deux facteurs, le rendement en fonction de la variété et de l'engrais utilisé par exemple. L'objectif est alors

de tester la signification de l'effet moyen de chaque facteur et de leur intéraction.

Le tableau de donnée se présente sous R sous la forme :

Y	Facteur I	Facteur II	
y_{111}	1	1	
y_{235}	2	3	5ème répétition de la combinaison 2-3
y_{ijk}	i	j	k ème répétition de la combinaison $i-j$

Test d'ANOVA

Le principe du test est similaire. On décompose la somme des carrés des écarts en fonction du facteur
étudié et on construit une statistique qui suit la loi de Fisher.
Il est ainsi possible de tester si une interaction est significative, si l'effet moyen d'un facteur est significatif …

Tableau d'analyse de variance

Source	ddl	SCE	CM	F	p value
facteur 1	$q_1 - 1$	SCE_{F1}	$\frac{SCE_{F1}}{q_1 - 1}$	$\frac{CM_{F1}}{CM_{res}}$	
facteur 2	$q_2 - 1$	SCE_{F2}	$\frac{SCE_{F2}}{q_2 - 1}$	$\frac{CM_{F2}}{CM_{res}}$	
Interaction	$(q_1 - 1)(q_2 - 1)$	$SCE_{F1\,F2}$	$\frac{SCE_{F1\,F2}}{(q_1 - 1)(q_2 - 1)}$	$\frac{CM_{F1\,F2}}{CM_{res}}$	
résiduelles	$n - q_1 - q_2 - 1$	SCE_{res}	$\frac{SCE_{res}}{n - q_1 - q_2 - 1}$		

Exemple 1 : On souhaite étudier le rendement d'une céréale en fonction de l'engrais et de la nature du terrain. On cultive p=2 types de terrain T1 et T2 et q=3 types d'engrais E1, E2 et E3. Chacune des combinaisons T*E est répétée 4 fois. (Fichier **rendement.txt)**

Terrain	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T1	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2	T2
Engrais	E1	E1	E1	E1	E2	E2	E2	E2	E3	E3	E3	E3	E1	E1	E1	E1	E2	E2	E2	E2	E3	E3	E3	E3
Rendt	61	76	47	77	34	30	67	62	85	104	74	75	77	47	54	77	46	50	29	53	69	70	75	85

Exemple 2

Présentation des données :

- Plantation d'arbres dans 3 forêts
- Comparaison de la hauteur des arbres

Forêt 1	Forêt 2	Forêt 3
23,3	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24,0
25,0	22,5	24,0
26,2	23,5	24,5

Présentation des données :

- **Les forêts** : Variable qualitative contenant trois modalités, appelée facteur (à effets fixes).
- **Hauteur des arbres** : Réponse, notée Y.

L'analyse de variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités sur les moyennes d'une variable quantitative Y.

Les échantillons sont de même taille => expérience équilibrée.

- Moyenne de chaque échantillon :

$$\overline{y_i} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}, \quad i = 1,\ldots, I.$$

- Variance de chaque échantillon :

$$s^2{}_i(y) = \frac{1}{J}\sum_{j=1}^{J} (y_{ij} - \overline{y_i})^2, \quad i = 1,\ldots, I.$$

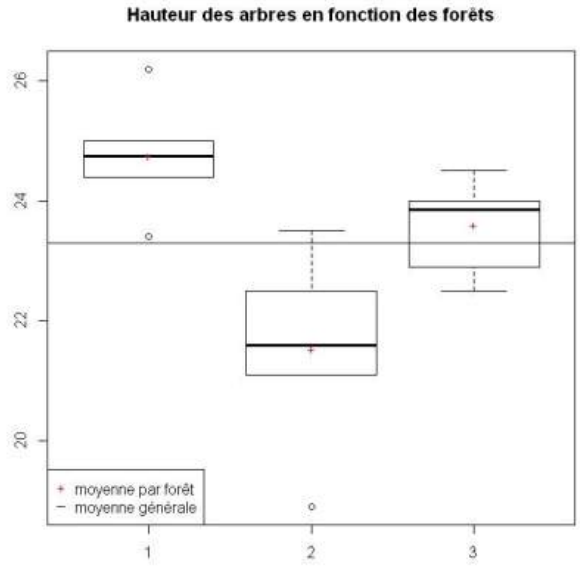
Application à l'exemple :

$\overline{y_1} = 24{,}75$ $\overline{y_2} = 21{,}53$ $\overline{y_3} = 23{,}6$

$s_1 = 0{,}83$ $s_2 = 2{,}49$ $s_3 = 0{,}57$

Nombre d'observations : n = I*J = 6*3=18

Application à l'exemple :



- **Modèle**:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1,\ldots I \text{ et } j = 1,\ldots, J$$

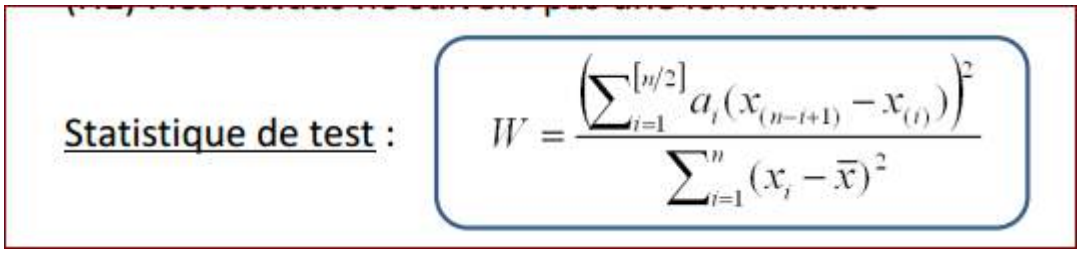
Test de comparaison des moyennes :
Hypothèse nulle (H0) :
Contre (H1) : Les ne sont pas tous égaux.
=> Utilisation de **l'analyse de la variance à un facteur.**

Les trois conditions pour l'ANOVA:

1. Les p échantillons comparés sont**indépendants**.
2. La variable quantitative étudiée suit une**loi normale** dans les p populations comparées.
3. Les p populations comparées ont même variance :**Homogénéité des variances** ou homoscédasticité.

3. Normalité :

Test de **Shapiro-Wilk** sur l'ensemble des résidus
(H0) : les résidus suivent une loi normale
(H1) : les résidus ne suivent pas une loi normale



TEST Statistique

L'inférence statistique est la partie des statistiques qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations, mais extrapole les constatations faites à un ensemble plus vaste et permet de tester des hypothèses sur cet ensemble ainsi que de prendre des décisions.

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

Soient H0 et H1 deux hypothèses (H0 est appelée hypothèse nulle, H1 hypothèse alternative), dont une et une seule qui est vraie. La décision consiste à retenir H0 ou H1

Pour un test bilatéral, nous pouvons émettre les hypothèses suivantes :

- Hypothèse nulle, H0 : pA = pB
- Hypothèse alternative, H1 : pA ≠ pB.

➢ Pour un test unilatéral, les hypothèses deviennent :

- Hypothèse nulle, H0 : pA = pB
- Hypothèse alternative, H1 : pA > pB ou pA < pB

LES TESTS PARAMETRIQUES

Un test est dit paramétrique si son objet est de tester une hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire qui suit la loi normale ou ayant un effectif important (n > 30).

Le test de Student

Ce test permet de comparer :

Les tests d'hypothèses vont permettre aux statisticiens de comparer des échantillons entre eux ou encore de comparer un échantillon avec une population de référence...

Types de tests d'hypothèses:

Dans le cadre de ces travaux pratiques, nous envisagerons trois types de tests d'hypothèses.

1. test de comparaison d'une moyenne d'un échantillon par rapport à une population
2. test de comparaison de 2 échantillons tirés de 2 populations indépendantes
3. test de comparaison de 2 échantillons tirés de 2 populations pairées

Les tests 2 et 3 sont des cas particuliers qui peuvent être plus simplement traités par l'ANOVA.

- une moyenne d'un échantillon à une valeur donnée
- les moyennes de deux échantillons indépendants
- les moyennes de deux échantillons appariés.

L'emploi de ce test reste subordonné en général à deux conditions d'application importantes qui sont la normalité et le caractère aléatoire et simple des échantillons.

Calcul : Soit X une variable aléatoire distribuée selon une loi normale, la variable aléatoire définie ci-dessus suit une loi de Student avec n - 1 degrés de liberté.

$$t_{obs} = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

où μ_0 est la moyenne de la population spécifiée par H_0, est la moyenne de l'échantillon, S^2 est la variance de l'échantillon et n est la taille de l'échantillon On compare la valeur calculée de t (t_{obs}) avec la valeur critique appropriée de t avec n - 1 degrés de liberté. On rejette H_0 si la valeur absolue de t_{obs} est supérieure à cette valeur critique.

Pour un risque d'erreur α choisi, n-1 degrés de liberté et un test bilatéral (H1 : $m \neq \mu_0$), on calcul :

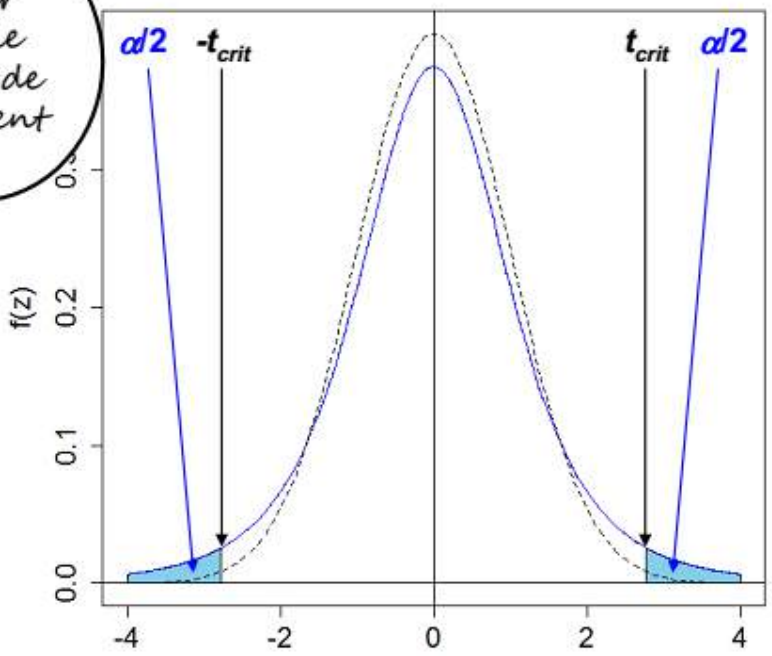
$$t_{obs} = \frac{m - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

Voir table du t de Student

$$t_{crit} = t_{n-1,1-\frac{\alpha}{2}}$$

Et on rejette H0 si :

$$|t_{obs}| > t_{crit}$$



4.2 Test de la moyenne (ou Test de Student)

Un contrôle anti-dopage a été effectué sur 16 sportifs. On a mesuré la variable X de moyenne μ, qui est le taux (dans le sang) d'une certaine substance interdite. Voici les données obtenues :

0.35	0.4	0.65	0.27	0.14	0.59	0.73	0.13
0.24	0.48	0.12	0.70	0.21	0.13	0.74	0.18

La variable X est supposée gausienne et de variance $\sigma^2 = 0.04$. On veut tester, au niveau de signification nominal 5% l'hypothèse selon laquelle le taux moyen dans le sang de la population des sportifs est égal 0.4.

On pose des hypothèses de test unilatérales :

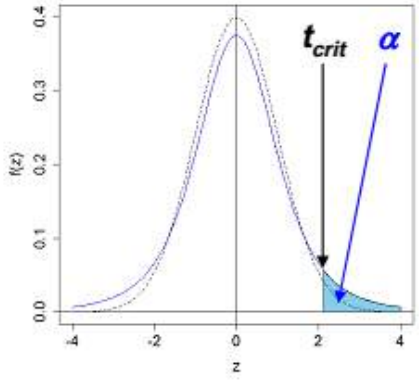
$$H_0 : \mu = \mu_0 = 0.4 \text{ contre } H_1 : \mu > 0.4$$

Pour un risque d'erreur α choisi, *n*-1 degrés de liberté et un test unilatéral droit ou gauche (H1 : $m > \mu_0$ ou $m < \mu_0$), on calcul :

$$t_{obs} = \frac{m - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

$$t_{crit} = t_{n-1,1-\alpha}$$

Voir table du t de Student



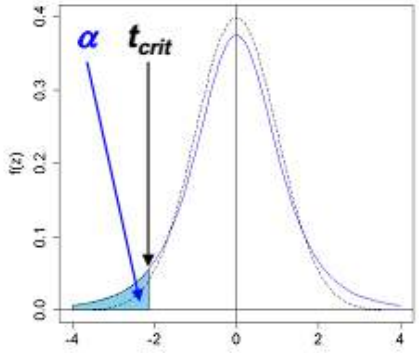
Et on rejette H0 si :

- Cas unilatéral droit (H1 : $m > \mu_0$) :

$$t_{obs} > t_{crit}$$

- Cas unilatéral gauche (H1 : $m < \mu_0$) :

$$t_{obs} < -t_{crit}$$

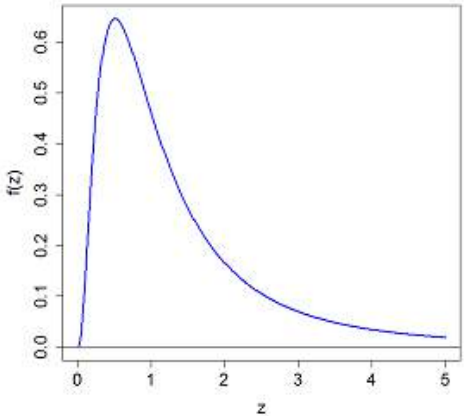


Pourquoi comparer deux variances ?

Le test *F*-Fisher de comparaison de deux variances observées s'effectue toujours sur deux échantillons et est souvent utilisé pour :

- Vérifier la condition de non différence des variances entre deux échantillons encore appelée homoscédasticité (cf. test *t*-Student de comparaison de deux moyennes observées)
- L'objectif du test ***F* de Fisher** est de comparer le rapport des deux variances, qui est supposé égal à 1 sous H0, avec la loi de Fisher

Ronald Fisher (1890-1962)



Cette fois-ci, le critère fonction des observations est le rapport des variances estimées pour les deux échantillons de taille n_1 et n_2 dont on veux démontrer qu'il est différent de 1 :

- H0 :

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = 1$$

- H1 :
  - Cas bilatéral : $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \neq 1$
  - Cas unilatéral droit : $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} > 1$
  - Cas unilatéral gauche : $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < 1$

En général, le test F de Fisher est utilisé de manière bilatérale

Attention, ce test de comparaison de variances entre deux échantillons n'est correct que lorsque l'une de ces deux conditions est observée :

- Les 2 échantillons sont de taille suffisante ($n_1 > 30$ et $n_2 > 30$)
- Au moins un des 2 échantillons est de petite taille ($n_1 \leq 30$ ou $n_2 \leq 30$) mais la distribution de la variable étudiée suit une loi Normale au sein de l'échantillon

H1 bilatérale

Pour un risque d'erreur α choisi, n_1-1 et n_2-1 degrés de liberté et un test bilatéral, on calcul :

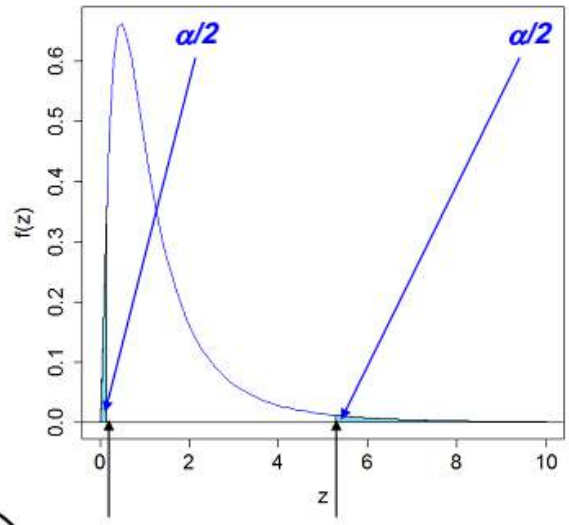
$$F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

$$F_{critINF} = F_{n_1-1,n_2-1,\frac{\alpha}{2}}$$

$$F_{critSUP} = F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$$

Et on rejette H0 si :

$$F_{obs} > F_{critSUP} \; ou \; F_{obs} < F_{critINF}$$



Attention, la loi de Fisher n'est pas symétrique et $F_{critINF} \neq -F_{critSUP}$

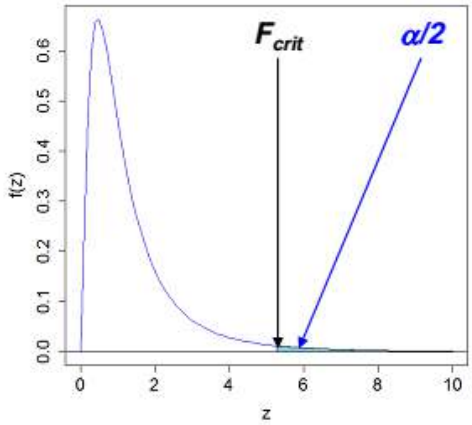
UNIVERSITÉ de Picardie Jules Verne

À noter que :

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} < F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$$

Équivaut à :

$$\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} > F_{n_2-1,n_1-1,1-\frac{\alpha}{2}}$$



Par conséquent, on calcul en pratique le rapport F_{obs} en faisant en sorte de calculer le rapport de la variance la plus grande (numérateur) sur la variance la plus petite (dénominateur) et on rejette H0 si :

$$F_{obs} > F_{n_{numérateur}-1,n_{dénominateur}-1,1-\frac{\alpha}{2}}$$

Voir table du F de Fisher

Table *F* de Fisher

Pour un risque d'erreur α de 5%, on croise les degrés de liberté du numérateur et du dénominateur pour trouver la valeur du *F* critique noté F_{crit} :

— ν_1 : nombre de degrés de liberté pour la plus faible des deux variances;
— ν_2 : nombre de degrés de liberté pour la plus forte des deux variances.
Pour $\alpha = 0{,}05$:

ν_2 ↓ / ν_1 →	1	2	3	4	5	6	8	12	24	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,01	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

- La moyenne des notes de statistique en PACES est de 45/ 100 (μ_0)

- Chez 18 étudiants on trouve une moyenne (m) de 54/ 100 avec un écart type (s) de 9
  - Peut on conclure que la moyenne m de cet échantillon est supérieur (test unilatéral) à la moyenne habituelle /théorique μ_0 ?

- Choix du test et vérification des conditions d'utilisation

  - L'échantillon est petit (n<30) mais la moyenne μ suit une loi normale dans la population dont est issue l'échantillon (test unilatéral)

  - La statistique t= $\frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$ suit une loi normale de 18-1 = 17 dll

- Définir H0 et H1
  - H0 : $\mu \leq \mu_0$
  - H1 : $\mu > \mu_0$

- Fixer le risque alpha et définir la règle de décision
  - On s'intéresse seulement à savoir si $\mu > \mu_0$ donc situation unilatéral
  - Valeur seuil en situation bilatérale t (2.5% , 17 ddl) = 2.11
  - Valeur seuil en situation unilatérale t (5% , 17 ddl) ≥ 1.74
  - Si H1 avait été $\mu < \mu_0$ avec H0 : $\mu \geq \mu_0$
    - Valeur seuil en situation unilatérale t (5% , 17 ddl) ≤ 1.74
  - Si t = $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} \geq 1.74$ on rejette H0, on accepte H1
  - Si t = $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} < 1.74$ on ne rejette pas H0

  - Calcul de la statistique
    - t= $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{54-45}{\frac{9}{\sqrt{18}}} = \frac{9}{\frac{9}{4.24}} = 4.24$
    - 4.24 (t calculé) ≥ 1.74 (t seuil) on rejette H0, on accepte H1 au risque de 5%
    - Au regard des résultats la moyenne de la population représentée par ces étudiants est supérieure à la moyenne habituelle
    - p < 0,0005 sur la table t de Student unilatéral

ddl \ α	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781	
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587	
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437	
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318	
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221	
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140	
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073	
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015	
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965	
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922	

Comparaison d'une moyenne observée à une moyenne théorique

- **Petits échantillons n<30**
  - Si la distribution de la variable aléatoire suit une **loi normale** et la **variance est inconnue**

- **Test T de Student**
  - La comparaison de la moyenne m observée sur n cas à une valeur théorique μ_0 est basée sur le rapport : $\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}}$
  - S désigne l'écart type estimé sur l'échantillon

- Si |t| est inférieur à la valeur lue dans la table de t (loi de Student) pour ddl = n-1 et le risque 5%, la différence n'est pas significative (ie l'échantillon provient bien de la population au regard de la moyenne);

- Si |t| est supérieur ou égal à la valeur lue dans la table de t (loi de Student) pour ddl = n-1 et le risque 5%, la différence est significative

- Valeur |t| trouvée dans la table de Student pour le risque fixé = degré de signification

- La moyenne des notes de statistique en PACES est de 45/ 100 (μ_0)

- Chez 18 étudiants on trouve une moyenne (m) de 54/ 100 avec un écart type (s) de 9
  - Peut on conclure que la moyenne m de cet échantillon est supérieur (test unilatéral) à la moyenne habituelle /théorique μ_0 ?
  - La moyenne m des notes suit une loi normale

- Choix du test et vérification des conditions d'utilisation

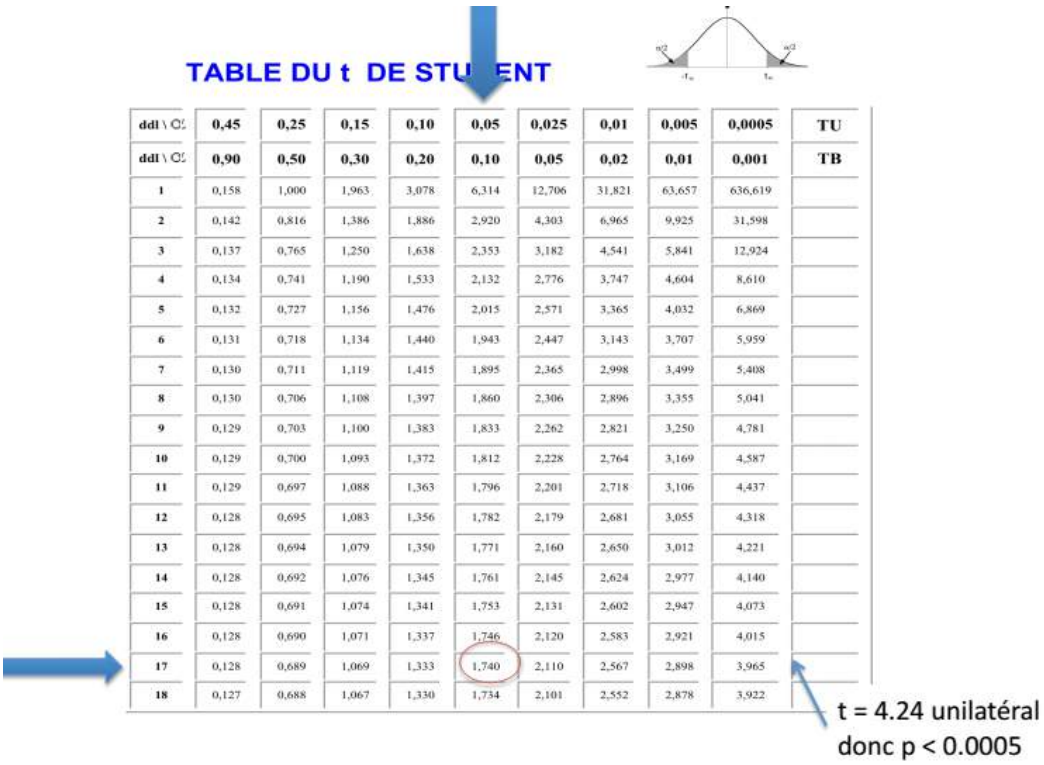
  - L'échantillon est petit (n<30) mais la moyenne μ suit une loi normale dans la population dont est issue l'échantillon (test unilatéral)

  - La statistique t= $\frac{m-\mu_0}{\frac{s}{\sqrt{n}}}$ suit une loi normale de 18-1 = 17 dll

- Définir H0 et H1
  - H0 : $\mu \leq \mu_0$
  - H1 : $\mu > \mu_0$

- Fixer le risque alpha et définir la règle de décision
  - On s'intéresse seulement à savoir si $\mu > \mu_0$ donc situation unilatéral
  - Valeur seuil en situation bilatérale t (2.5% , 17 ddl) = 2.11
  - Valeur seuil en situation unilatérale t (5% , 17 ddl) ≥ 1.74
  - Si H1 avait été $\mu < \mu_0$ avec H0 : $\mu \geq \mu_0$
    - Valeur seuil en situation unilatérale t (5% . 17 ddl) ≤ 1.74
  - Si t = $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} \geq 1.74$ on rejette H0, on accepte H1
  - Si t = $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} < 1.74$ on ne rejette pas H0

  - Calcul de la statistique
    - t= $\frac{m - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{54-45}{\frac{9}{\sqrt{18}}} = \frac{9}{\frac{9}{4.24}} = 4.24$
    - 4.24 (t calculé) ≥ 1.74 (t seuil) on rejette H0, on accepte H1 au risque de 5%
    - Au regard des résultats la moyenne de la population représentée par ces étudiants est supérieure à la moyenne habituelle
    - p < 0,0005 sur la table t de Student unilatéral



ddl \ α	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781	
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587	
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437	
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318	
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221	
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140	
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073	
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015	
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965	
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922	

Comparaison de 2 moyennes observées sur 2 échantillons indépendants

- Les données sont non appariées
  - Moyenne sur 2 échantillons indépendants

- Taille de l'échantillon
  - Grands échantillons ≥ 30
  - Petits échantillons <30

- La comparaison entre 2 moyennes observées mA et mB sur nA et nB individus respectivement (promotion PACES de 2 facultés A et B)

- Test t de Student dans le cas des petits échantillons ie nA et/ou nB < 30 (au moins l'un est petit)
  - $t = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}} = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$
- S^2 estimation de la variance supposée commune ie **test de l'égalité des variances au préalable**
  - $S^2 = \frac{(nA-1)\times S\hat{A}^2 + (nB-1)\times S\hat{B}^2}{nA+nB-2} = \frac{\sum(x-mA)^2 - \sum(x-mB)^2}{nA+nB-2}$
- Si |t| est inférieur à la valeur lue dans la table de t pour un ddl = $nA + nB - 2$
- **2 hypothèses normalité et égalité de variance** = test t robuste en première approximation
  - Exemple identique au précèdent mais effectif de n_A = 9 et n_B =18
  - Exemple on compare la moyenne des notes de statistiques de PACES entre 2 facultés A et B
    - On tire au sort n_A = 9 et n_B = 18 dans chacune des promotions et on compare la moyenne
    - mA = 45/100 (d'écart type $s = 3$)
    - mB = 49/100 (d'écart type $s = 5$)

- Est-ce que la moyenne est différente entre les deux facultés (test bilatéral) ?

- Choix du test et vérification des conditions d'utilisation
  - 2 échantillons de n < 30
  - Deux variables continues, loi normale, on suppose l'égalité de variance
  - **On doit le faire le test d'égalité des variances test F au préalable (cf chap I)**
  - Moyennes mA et mB des notes suivent une loi normale
  - Statistique $t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$ suit une loi de Student à (9 + 18 -2) = 25 ddl

- Définir H0 et H1
  - H0 : $\mu A = \mu B$
  - H1 : $\mu A \neq \mu B$

- Fixer le risque alpha et définir la règle de décision
  - Risque bilatéral 5%
  - Rejet de H0 si
    - $t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$ t (5%, ddl = 25) = 2.06
    - Rejet si |t| > 2.06

- Calcul de la statistique
  - $S^2 = \frac{(nA-1)\times SA^2 + (nB-1)\times SB^2}{nA+nB-2} = \frac{\sum(x-mA)^2 - \sum(x-mB)^2}{nA+nB-2} = \frac{(9-1)\times 3^2 + (18-1)\times 5^2}{9+18-2} = \frac{72+425}{25} = 19.88$
  - $t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}} = \frac{45-49}{\sqrt{\frac{19.88}{9} + \frac{19.88}{18}}} = -2.20$
  - $|t| \geq 2.06$

  - Conclusion on rejette H0 au risque alpha bilatéral de 5%
  - La moyenne des notes de faculté B est supérieure à celle de A ie elle est différente
  - p compris entre 0.05 et 0,02 (table de Student bilateral)

TABLE DU t DE STUDENT

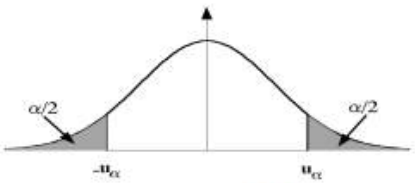
ddl \ Cα	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ Cα	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781	
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587	
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437	
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318	
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221	
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140	
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073	
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015	
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965	

p compris entre 0.05 et 0.02 (table du t de Student bilateral avec 2.20 et 25 ddl

22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792	
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767	
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745	
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725	

Loi normale centrée réduite

Table de l'écart réduit



La table donne la probabilité α pour que l'écart réduit dépa[illegible] en valeur absolue, une valeur donnée u, c'est-à-dire la probabilité de ne pas trouver z dans l'interva[illegible]-u ; u] centré sur 0. Chacune des 2 aires hachurées correspondent à une probabilité égales $\alpha/2$. La p[illegible]abilité d'observer z dans l'intervalle]-u ; u[est évidemment 1- α.

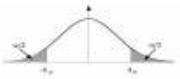
Test bilatéral : lire α Test unilatéral à droite ou à g[illegible]che : diviser α par 2

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170		1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

- Calcul de la statistique
  - $S2 = \frac{(nA-1)\times SA^2+(nB-1)\times SB^2}{nA+nB-2} = \frac{\sum(x-mA)^2-\sum(x-mB)^2}{nA+nB-2} = \frac{(9-1)\times 3^2+(18-1)\times 5^2}{9+18-2} = \frac{72+425}{25} = 19.88$
  - $t = \frac{m_A - m_B}{\sqrt{\frac{S2}{n_A}+\frac{S2}{n_B}}} = \frac{45-49}{\sqrt{\frac{19.88}{9}+\frac{19.88}{18}}} = -2.20$
  - $|t| \geq 2.06$

  - Conclusion on rejette H0 au risque alpha bilatéral de 5%
  - La moyenne des notes de faculté B est supérieure à celle de A ie elle est différente
  - p compris entre 0.05 et 0,02 (table de Student bilateral)

TABLE DU t DE STUDENT



ddl \ C%	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ C%	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781	
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587	
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437	
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318	
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221	
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140	
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073	
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015	
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965	

p compris entre 0.05 et 0.02 (table du t de Student bilateral avec 2.20 et 25 ddl

22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792	
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767	
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745	
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725	

Comparaison de 2 moyennes observées sur échantillons appariés

- Les échantillons ne sont pas indépendants
  - 2 examinateurs corrigent les copies de 100 étudiants (ie même échantillon de copie)
  - Les tests précédents ne sont plus valables car ils présupposent l'indépendance des échantillons = ils corrigent les copies des mêmes étudiants
- Les échantillons appariés avec des effectifs < 30
  - Pour comparer les moyennes de deux séries appariées de faible effectif, on forme pour chaque paire la différence des deux mesures et on compare la moyenne des différences à 0 par le rapport
  - $t = \frac{\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{m}{\frac{s}{\sqrt{n}}} = \frac{m-0}{\frac{s}{\sqrt{n}}}$
  - m et S désignent la moyenne et l'écart type estimés sur l'échantillon des n différences di
  - $S^2 = \frac{\sum_{i=1}^{n} di^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}}{n-1}$
    - Si |t| est inférieur à la valeur lue dans la table de t pour le nombre de degrés de liberté (n-1) et le risque 5 % les moyennes ne diffèrent pas significativement au seuil de 5 %
    - Si |t| est supérieur les moyennes différent significativement et le risque indiqué par la table pour la valeur |t| trouvée fixe le degré de signification
    - Formule applicable si la différence est distribuée selon une loi normale

- On mesure l'effet du stress lié au examen PACES sur la glycémie de 9 étudiants
  - La glycémie suit une loi normale dans la population dont sont issues les étudiants
  - Pour chaque étudiant 2 mesures sont faites avant et après les examens
- Résultats

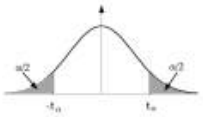
étudiants	1	2	3	4	5	6	7	8	9
avant	5.5	4.3	6.5	4.5	5.2	4.3	5.0	5.4	5.2
après	5.4	6.7	6.5	6.0	5.2	5.0	4.8	4.7	4.5

- Le stress du à l'examen modifie t il la glycémie des étudiants ?

- Choix du test statistique et vérification des conditions d'utilisation
  - Données appariées, un seul échantillon, loi normale
  - Test de la différence à zéro
  - $t = \frac{\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{m}{\frac{s}{\sqrt{n}}} = \frac{m-0}{\frac{s}{\sqrt{n}}}$ suit une loi de Student à 9-1 = 8 ddl

- Définir les hypothèses
  - H0 : $\mu = 0$
  - H1 : $\mu \neq 0$
- Fixer le risque alpha et définir la règle de décision
  - Alpha 5 % bilatéral
  - Zone de rejet de H0 : 2.306 sur la table de Student à 8 ddl

TABLE DU t DE STUDENT



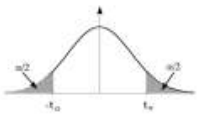
ddl \ C	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ C	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	

- Calcul de la statistique

étudiants	1	2	3	4	5	6	7	8	9	
avant	5.5	4.3	6.5	4.5	5.2	4.3	5.0	5.4	5.2	
après	5.4	6.7	6.5	6.0	5.2	5.0	4.8	4.7	4.5	
di	-0.1	2.4	0	1.5	0	0.7	-0.2	-0.7	-0.7	2.9
di^2	0.01	5.76	0	2.25	0	0.49	0.04	0.49	0.49	9.53

- $S^2 = \frac{\sum_{i=1}^{n} di^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}}{n-1} = \frac{9.53 - \frac{8.41}{9}}{8} = 1.0744$
- m = 2.9 / 9 = 0.32
- $t = \frac{0.32}{\frac{1.036}{\sqrt{9}}} = 0.93$
- p compris entre 0.50 et 0.30

TABLE DU t DE STUDENT



ddl \ α	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924	
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610	
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869	
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959	
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408	
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041	

- $t = \frac{0.32}{\frac{1.036}{\sqrt{9}}} = 0.93$

- Appliquer les règles de décisions
  - $t < 2.306$, on ne rejette pas H0 : la différence de glycémie n'est pas différent de zéro

- Conclusion le stress lié aux examens ne semble pas agir sur la glycémie

Régression simple et multiple : principes et exemples d'application

1.1. Régression linéaire simple

Un exemple simple d'ajustement par les moindres carrés est donné par l'analyse bivariée de variables quantitatives qui peut se simplifier par le calcul des variances et de la covariance des deux variables X et Y retenues.

La variance répond à la formule suivante :

$$VarX = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

où : n, nombre d'individus

x_i, valeur de la variable x pour l'individu i

$\bar{x}$, moyenne arithmétique de la variable x

La covariance considère les variations communes des deux variables selon la formule :

$$CovXY = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})$$

où : n, nombre d'individus

x_i, valeur de la variable x pour l'individu i

$\bar{x}$, moyenne arithmétique de la variable x

y_i, valeur de la variable x pour l'individu i

$\bar{y}$, moyenne arithmétique de la variable y

Enfin, le coefficient de corrélation est donné par la formule :

$$Ceof.cor = \frac{CovXY}{\sqrt{VarX} * \sqrt{VarY}}$$

Le coefficient de corrélation correspond au cosinus de l'angle formé entre deux droites de régression se croisant aux coordonnées des moyennes arithmétiques des deux variables observées (centre de gravité supposé). On définit donc deux droites répondant chacune à une équation affine :

$$X' = a1Y + b1$$

Et

$$Y' = a2X + b2$$

X' et Y' étant les valeurs estimées à partir des valeurs observées X et Y.

Dans le cas de l'analyse bivariée, les coefficients des équations sont facilement donnés

par :

$$a1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$b1 = \bar{y} - a1\bar{x}$$

$$b2 = \bar{x} - a2\bar{y}$$

Prenons comme exemple la matrice théorique suivante (table A1) :

id	X	Y	X'	Y'	X-moyX	Y-moyY	(X-moyX)²	(Y-moyY)²	(X-moyX)(Y-moyY)
1	2	18	1.847222222	13.95157895	-4.777777778	8.333333333	22.82716049	69.44444444	-39.81481481
2	3	15	3.622222222	13.05473684	-3.777777778	5.333333333	14.27160494	28.44444444	-20.14814815
3	4	12	5.397222222	12.15789474	-2.777777778	2.333333333	7.716049383	5.444444444	-6.481481481
4	5	9	7.172222222	11.26105263	-1.777777778	-0.666666667	3.160493827	0.444444444	1.185185185
5	6	6	8.947222222	10.36421053	-0.777777778	-3.666666667	0.604938272	13.44444444	2.851851852
6	8	5	9.538888889	8.570526316	1.222222222	-4.666666667	1.49382716	21.77777778	-5.703703704
7	10	6	8.947222222	6.776842105	3.222222222	-3.666666667	10.38271605	13.44444444	-11.81481481
8	11	7	8.355555556	5.88	4.222222222	-2.666666667	17.82716049	7.111111111	-11.25925926
9	12	9	7.172222222	4.983157895	5.222222222	-0.666666667	27.27160494	0.444444444	-3.481481481

Table A1 : Exemple théorique

Le coefficient de corrélation est de *-0.72844463*, les équations sont :

$Y' = -0.8968X + 15.745$ (en jaune)

et

$X' = -0.5917Y + 12.497$ (en magenta)

id	X	Y	X'	Y'	X-moyX	Y-moyY	(X-moyX)²	(Y-moyY)²	(X-moyX)(Y-moyY)
1	2	18	1.847222222	13.95157895	-4.777777778	8.333333333	22.82716049	69.44444444	-39.81481481
2	3	15	3.622222222	13.05473684	-3.777777778	5.333333333	14.27160494	28.44444444	-20.14814815
3	4	12	5.397222222	12.15789474	-2.777777778	2.333333333	7.716049383	5.444444444	-6.481481481
4	5	9	7.172222222	11.26105263	-1.777777778	-0.666666667	3.160493827	0.444444444	1.185185185
5	6	6	8.947222222	10.36421053	-0.777777778	-3.666666667	0.604938272	13.44444444	2.851851852
6	8	5	9.538888889	8.570526316	1.222222222	-4.666666667	1.49382716	21.77777778	-5.703703704
7	10	6	8.947222222	6.776842105	3.222222222	-3.666666667	10.38271605	13.44444444	-11.81481481
8	11	7	8.355555556	5.88	4.222222222	-2.666666667	17.82716049	7.111111111	-11.25925926
9	12	9	7.172222222	4.983157895	5.222222222	-0.666666667	27.27160494	0.444444444	-3.481481481

Table A1 : Exemple théorique

Le coefficient de corrélation est de *-0.72844463*, les équations sont :

$Y' = -0.8968X + 15.745$ (en jaune)

et

$X' = -0.5917Y + 12.497$ (en magenta)

id	X	Y	X'	Y'	X-moyX	Y-moyY	(X-moyX)²	(Y-moyY)²	(X-moyX)(Y-moyY)
1	2	18	1.847222222	13.95157895	-4.777777778	8.333333333	22.82716049	69.44444444	-39.81481481
2	3	15	3.622222222	13.05473684	-3.777777778	5.333333333	14.27160494	28.44444444	-20.14814815
3	4	12	5.397222222	12.15789474	-2.777777778	2.333333333	7.716049383	5.444444444	-6.481481481
4	5	9	7.172222222	11.26105263	-1.777777778	-0.666666667	3.160493827	0.444444444	1.185185185
5	6	6	8.947222222	10.36421053	-0.777777778	-3.666666667	0.604938272	13.44444444	2.851851852
6	8	5	9.538888889	8.570526316	1.222222222	-4.666666667	1.49382716	21.77777778	-5.703703704
7	10	6	8.947222222	6.776842105	3.222222222	-3.666666667	10.38271605	13.44444444	-11.81481481
8	11	7	8.355555556	5.88	4.222222222	-2.666666667	17.82716049	7.111111111	-11.25925926
9	12	9	7.172222222	4.983157895	5.222222222	-0.666666667	27.27160494	0.444444444	-3.481481481

Table A1 : Exemple théorique

Le coefficient de corrélation est de *-0.72844463*, les équations sont :

$Y' = -0.8968X + 15.745$ (en jaune)

et

$X' = -0.5917Y + 12.497$ (en magenta)

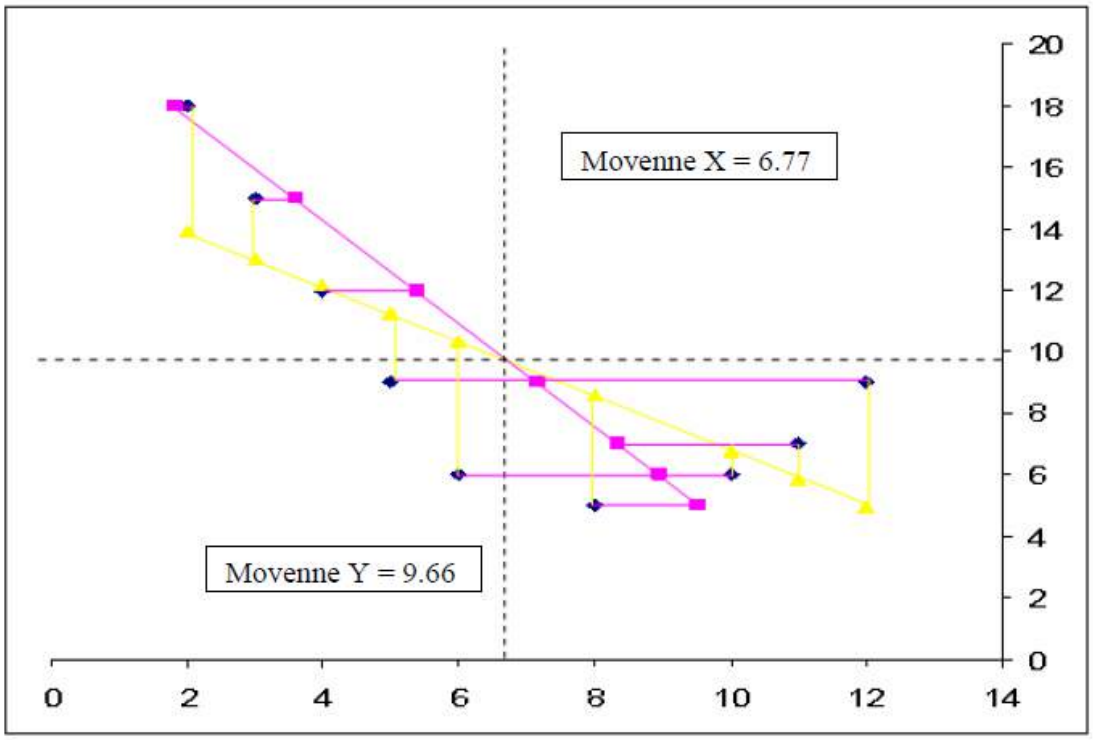


Figure A8 : Les deux droites de régression et le coefficient de corrélation

1.2. Régression linéaire multiple

L'exemple développé à partir de deux variables permet de comprendre la logique de la théorie de la régression mais il ne peut être généralisé de la sorte aux régressions multiples. Le système à deux équations à deux inconnus présenté se résolvait facilement comme on l'a vu. Les équations se compliquent avec plusieurs régresseurs, deux méthodes distinctes permettent de résoudre les équations. La première repose sur la connaissance des coefficients de corrélation linéaire simple de toutes les paires de variables entre elles, de la moyenne arithmétique et des écarts-types de toutes les variables. La seconde repose sur des calculs matriciels.

1.2.1. Les étapes de calcul fondé les variables descriptives

Soit un ensemble de p variable où la *p-ième* variable est la variable indépendante. Toutes les variables sont au préalable centrées-réduites. Soit r_{12}, r_{13}… r_{pp} les coefficients de corrélations linéaires des paires de variables et s_1, s_2, …, s_p les écarts-types.

Prenons un exemple avec $p = 4$ soit 3 variables dépendantes. Dans un premier temps on calcule les coefficients de régression linéaire a'_1, a'_2, a'_3 en résolvant un système de *p-1* équations à *p-1* inconnues :

$$r_{1p} = a'_1 + r_{12}a'_2 + r_{13}a'_3$$

$$r_{2p} = a'_2 + r_{21}a'_1 + r_{23}a'_3$$

$$r_{3p} = a'_3 + r_{31}a'_1 + r_{32}a'_3$$

Pour résoudre ce système on procède par substitutions successives :

$$a'_1 = r_{1p} - r_{12}a'_2 + r_{13}a'_3$$

d'où

$$r_{2p} = a'_2 + (r_{21} * (r_{1p} - r_{12}a'_2 + r_{13}a'_3)) + r_{23}a'_3$$

$$a'_2 = r_{2p} - r_{21}a'_1 + r_{23}a'_3$$

$$a'_3 = r_{3p} - r_{31}a'_2 + r_{32}a'_3$$

Connaissant désormais les coefficients de régression on détermine ceux des variables brutes :

$$a_1 = a_1' \frac{S_y}{S_{x1}} \; ; \; a2 = a_2' \frac{S_y}{S_{x2}} \text{ et } a3 = a_3' \frac{S_y}{S_{x3}}$$

Enfin, la constante d'ajustement est donnée en résolvant l'équation pour la coordonnée à l'origine :

$$\varepsilon = \bar{y} - a_1 \bar{x}_1 - a_2 \bar{x}_2 - a_3 \bar{x}_3$$

Le coefficient de détermination multiple est donné par :

$$R^2 = \sum_{j=1}^{p-1} a_j' r_{jp}$$

Prenons garde au fait que ce coefficient – dont les a'_{p-1} constituent en quelque sorte la contribution – croît avec le nombre de variable. Par conséquent, ce comportement déterministe lié aux propriétés des variables aléatoires doit être compenser, on calcule alors le coefficient ajusté :

$$R^2 ajusté = 1 - \frac{(n-1)}{n-(p-1)-1}(1-R^2)$$

Où : n : nombre d'individus

On peut également résoudre le système d'équations en prenant comme principe l'justement par les moindres carrés (Chadule) :

$$\sum_{i=1}^{n} \varepsilon_i^2 \min$$

Où : ε : variance résiduelle

Les coefficients a_j sont alors extraits des équations :

$$Cov_{p,1} = a_1 Var_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Cov_{1,p-1}$$

$$Cov_{p,2} = a_1 Cov_{2,1} + a_2 Var_2 + \ldots + a_{p-1} Cov_{2,p-1}$$

...

$$Cov_{p,p-1} = a_1 Covp - 1,_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Var_{p-1}$$

Les *p-1* coefficients sont ensuite obtenus par résolution du système. Avec deux variables explicatives X_1 et X_2 et une variable à expliquer Y on a par exemple :

$$a_1 = \frac{\left(Var_{X_2} * Cov_{YX_1}\right) - \left(Cov_{YX_2} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_1} - \left(r_{YX_2} * r_{X_1X_2}\right)\right)}{\sigma_{X_1} * \left(1 - r^2_{X_1X_2}\right)}$$

$$a_2 = \frac{\left(Var_{X_1} * Cov_{Y_{21}}\right) - \left(Cov_{YX_1} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_2} - \left(r_{YX_1} * r_{X_1X_2}\right)\right)}{\sigma_{X_{21}} * \left(1 - r^2_{X_1X_2}\right)}$$

Les coefficients a_j sont alors extraits des équations :

$$Cov_{p,1} = a_1 Var_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Cov_{1,p-1}$$

$$Cov_{p,2} = a_1 Cov_{2,1} + a_2 Var_2 + \ldots + a_{p-1} Cov_{2,p-1}$$

...

$$Cov_{p,p-1} = a_1 Covp - 1,_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Var_{p-1}$$

Les *p-1* coefficients sont ensuite obtenus par résolution du système. Avec deux variables explicatives X_1 et X_2 et une variable à expliquer Y on a par exemple :

$$a_1 = \frac{\left(Var_{X_2} * Cov_{YX_1}\right) - \left(Cov_{YX_2} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_1} - \left(r_{YX_2} * r_{X_1X_2}\right)\right)}{\sigma_{X_1} * \left(1 - r^2_{X_1X_2}\right)}$$

$$a_2 = \frac{\left(Var_{X_1} * Cov_{Y_{21}}\right) - \left(Cov_{YX_1} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_2} - \left(r_{YX_1} * r_{X_1X_2}\right)\right)}{\sigma_{X_{21}} * \left(1 - r^2_{X_1X_2}\right)}$$

Les coefficients a_j sont alors extraits des équations :

$$Cov_{p,1} = a_1 Var_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Cov_{1,p-1}$$

$$Cov_{p,2} = a_1 Cov_{2,1} + a_2 Var_2 + \ldots + a_{p-1} Cov_{2,p-1}$$

...

$$Cov_{p,p-1} = a_1 Covp - 1,_1 + a_2 Cov_{1,2} + \ldots + a_{p-1} Var_{p-1}$$

Les *p-1* coefficients sont ensuite obtenus par résolution du système. Avec deux variables explicatives X_1 et X_2 et une variable à expliquer Y on a par exemple :

$$a_1 = \frac{\left(Var_{X_2} * Cov_{YX_1}\right) - \left(Cov_{YX_2} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_1} - \left(r_{YX_2} * r_{X_1X_2}\right)\right)}{\sigma_{X_1} * \left(1 - r^2_{X_1X_2}\right)}$$

$$a_2 = \frac{\left(Var_{X_1} * Cov_{Y_{21}}\right) - \left(Cov_{YX_1} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_2} - \left(r_{YX_1} * r_{X_1X_2}\right)\right)}{\sigma_{X_{21}} * \left(1 - r^2_{X_1X_2}\right)}$$

Le coefficient de corrélation multiple est alors donnée par :

$$R_{Y,X_1X_2} = \sqrt{\frac{\left(r^2_{YX_1} + r^2_{YX2} - 2\left(r_{YX_1} * r_{YX_2} * r_{X_1X_2}\right)\right)}{1 - r^2_{X_1X_2}}} = r_{YY'}$$

1.2.2. La notation matricielle

L'équation de type :

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_2 + \beta_2 1 \vec{x}_2 + \varepsilon$$

est donnée sous forma matricielle par :

$$y = X\beta + \varepsilon$$

Où :

$$y = \begin{matrix} y_1 \\ y_2 \\ \dots \\ y_{n-1} \\ y_n \end{matrix}, X = \begin{matrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \dots & \dots & \dots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{matrix}, \beta = \begin{matrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{matrix}, \varepsilon = \begin{matrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{matrix}$$

Il s'agit dès lors de calculer le vecteur des estimateurs $\hat{\beta}$ défini par l'égalité suivante :

$$\hat{\beta} = (X * X')^{-1} X' y$$

En notation matricielle X' signifie la matrice X transposée et X^{-1} la matrice inverse.

Dans l'exemple qui suit nous réalisons une régression multiple pour expliquer la hauteur de neige en fonction de l'altitude, de la rugosité, de la pente, de l'orientation, de la latitude et de la longitude (table A2).

H_NEIGE	vecteur	altitude	rugosite	pente	orient.	lat	long.
95	1	2768	252	22	324	8760219	438465.0625
150	1	4108	333	29	308	8760195	438474.0625
4	1	4045	62	5	249	8760168	438480.0625
0	1	4572	85	8	14	8760135	438489.0625
0	1	4614	115	10	63	8760105	438495.0625
80	1	4321	176	16	130	8760072	438498.0625
95	1	3886	72	6	199	8760039	438504.0625
20	1	4206	57	5	32	8760012	438507.0625
90	1	4192	266	23	197	8759985	438513.0625
10	1	4051	69	6	113	8759955	438519.0625
10	1	3746	62	5	149	8759922	438519.0625
50	1	3789	42	3	218	8759895	438525.0625
45	1	3771	44	4	53	8759865	438531.0625
60	1	3796	48	4	101	8759838	438534.0625
55	1	3885	77	7	332	8759811	438537.0625
3	1	4295	113	10	18	8759787	438540.0625
33	1	4467	147	13	50	8759760	438546.0625

H_NEIGE	vecteur	altitude	rugosite	pente	orient.	lat	long.
95	1	2768	252	22	324	8760219	438465.0625
150	1	4108	333	29	308	8760195	438474.0625
4	1	4045	62	5	249	8760168	438480.0625
0	1	4572	85	8	14	8760135	438489.0625
0	1	4614	115	10	63	8760105	438495.0625
80	1	4321	176	16	130	8760072	438498.0625
95	1	3886	72	6	199	8760039	438504.0625
20	1	4206	57	5	32	8760012	438507.0625
90	1	4192	266	23	197	8759985	438513.0625
10	1	4051	69	6	113	8759955	438519.0625
10	1	3746	62	5	149	8759922	438519.0625
50	1	3789	42	3	218	8759895	438525.0625
45	1	3771	44	4	53	8759865	438531.0625
60	1	3796	48	4	101	8759838	438534.0625
55	1	3885	77	7	332	8759811	438537.0625
3	1	4295	113	10	18	8759787	438540.0625
33	1	4467	147	13	50	8759760	438546.0625

Analyse en Composantes Principales (ACP)

Ajustement du nuage des individus

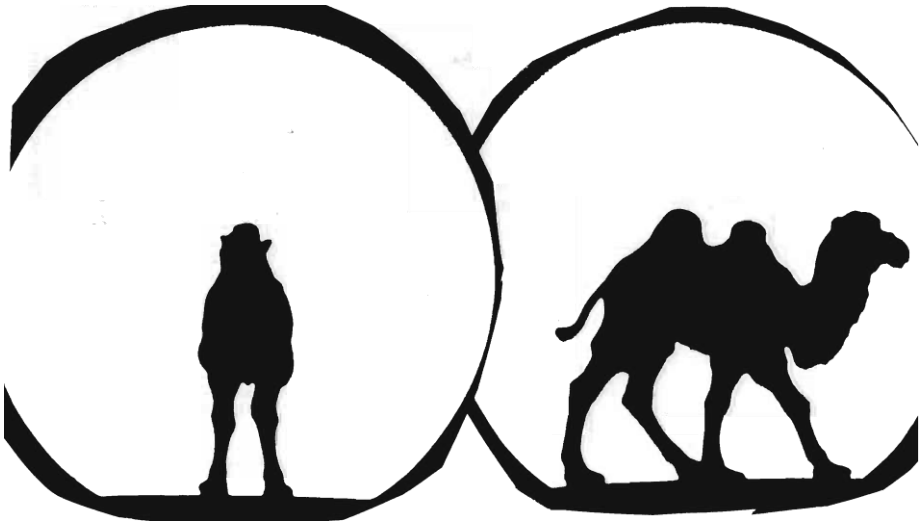


Figure: Quel animal? (*illustration JP Fénelon*)

Objectifs de l'ACP :

1-Descriptif - exploratoire : visualisation de données par graphiques simples

2-Synthèse - résumé de grands tableaux individus × variables

Exemples

- Analyse sensorielle : note du descripteur *k* pour le produit *i*

- Ecologie : concentration du polluant *k* dans la rivière *i*

- Economie : valeur de l'indicateur *k* pour l'année *i*

- Génétique : expression du gène *k* pour le patient *i*

- Biologie : mesure *k* pour l'animal *i*

- Marketing : valeur d'indice de satisfaction *k* pour la marque *i*

- Sociologie : temps passé à l'activité *k* par les individus de la

Les données température

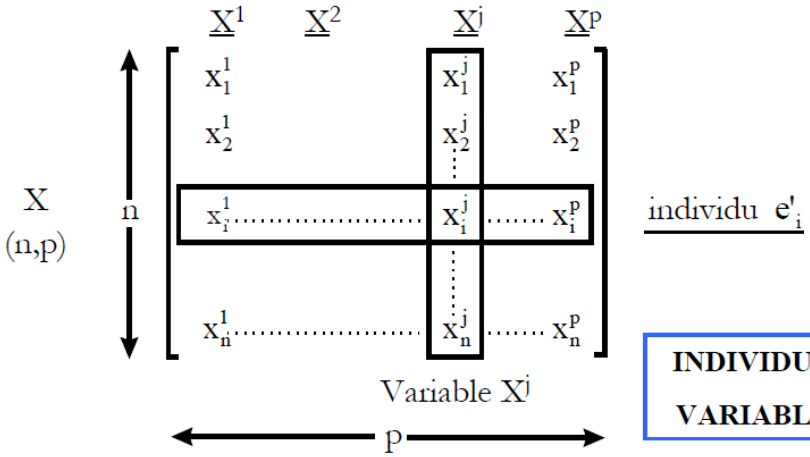
- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

I. L'ANALYSE EN COMPOSANTES PRINCIPALES
LE PROBLÈME

1. LES DONNÉES

p variables quantitatives observées sur n individus.



INDIVIDU = Élément de R^p

VARIABLE = Élément de R^n

Pour la variable k, on note :

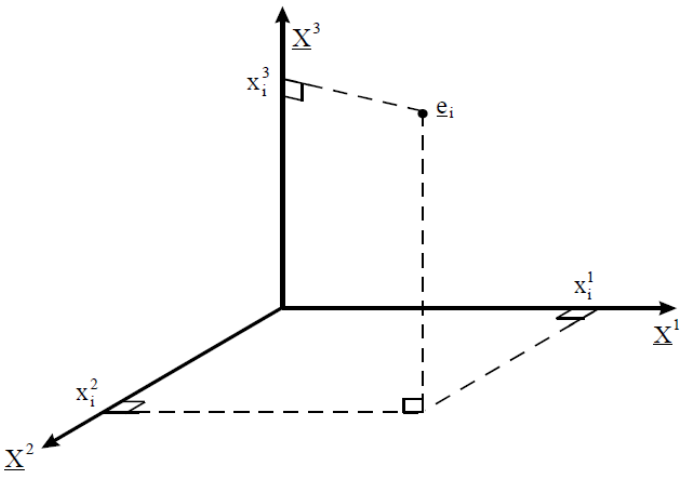
$$\text{la moyenne : } \bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$$

$$\text{l'écart-type : } s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$$

On cherche à représenter le nuage des individus.

A chaque individu noté e_i, on peut associer un point dans R^p = espace des individus.

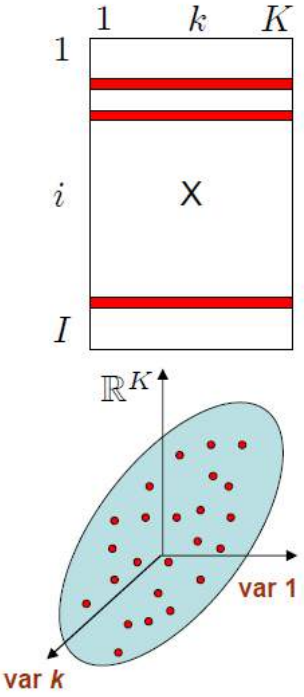
A chaque variable du tableau X est associé un axe de R^p.



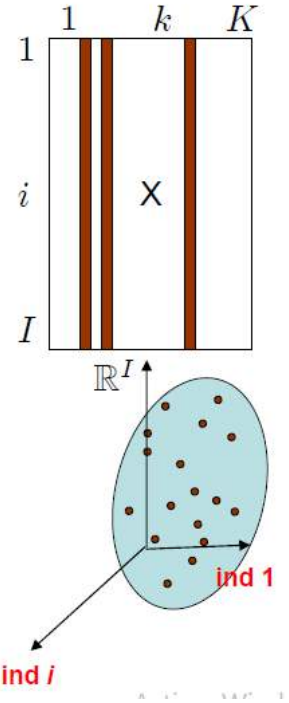
Impossible à visualiser dès que p > 3.

Deux nuages de points

Etude des individus



Etude des variables



On notera :

$\mathbf{X} = (x_{ij})_{n \times p}$ la matrice des données brutes où $x_{ij} \in \mathbb{R}$ est la valeur du $i^{\text{ème}}$ individu sur la $j^{\text{ème}}$ variable.

$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \in \mathbb{R}^p$ la description du $i^{\text{ème}}$ individu (ligne de $\mathbf{X}$)

$\mathbf{x}^j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \in \mathbb{R}^n$ la description de la $j^{\text{ème}}$ variable (colonne de $\mathbf{X}$).

Analyse en Composantes Principales (ACP)

Le nuage des individus N_I

1 individu = 1 ligne du tableau) 1 point dans un espace à K dim Le nuage des

Notion de ressemblance : distance (au carré) entre individus i et i' :

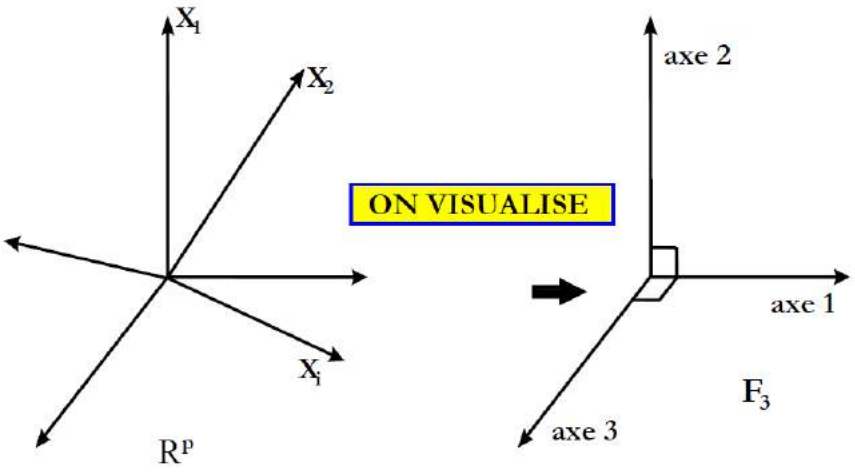
$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \qquad \text{(merci Pythagore)}$$

Etude des individus $\equiv$ Etude de la forme du nuage N^I

Etude des individus _ Etude de la forme du nuage N_I

Centrage – réduction des données

_ Centrer les données ne modifie pas la forme du nuage) toujours centrer



axes principaux

Positionnement des individus – Principe de l'ACP (1) – Notion d'inertie

Principe : Construire un système de représentation de dimension réduite (q << p) qui préserve les **distances** entre les individus. On peut la voir comme une compression avec perte (contrôlée) de l'information.

Distance euclidienne entre 2 individus (i, i')

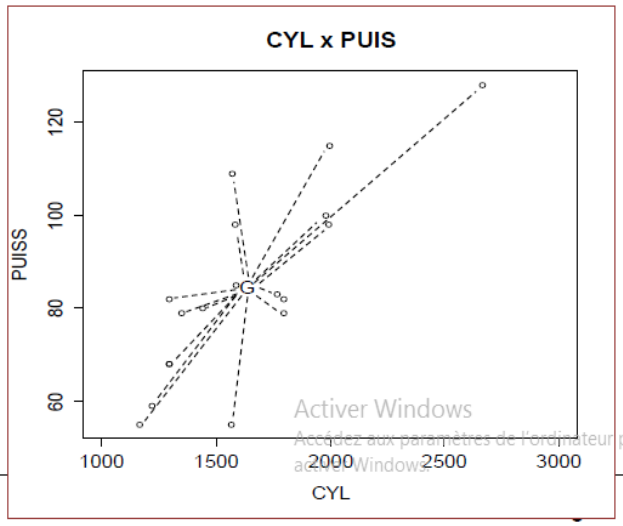
$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

Un critère global : distance entre l'ensemble des individus pris 2 à 2, **inertie** du nuage de points dans l'espace originel. Elle traduit la quantité d'information disponible.

$$I_p = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(i,i')$$

Autre écriture de l'inertie : écart par rapport au barycentre G (vecteur constitué des moyennes des p variables)

$$I_p = \frac{1}{n} \sum_{i=1}^{n} d^2(i,G)$$



L'inertie indique la dispersion autour du barycentre, c'est une variance multidimensionnelle (calculée sur p dimensions)

réservation des proximités ans le repère réduit

(1) Les proximités entre individus sont préservées si on prend un nombre q de composantes suffisamment représentatives (en terme de % d'inertie exprimé

(2) Si on prend les « p » facteurs, on retrouve les distances dans le repère originel

Distances dans le repère originel (variables centrées et réduites)

$$d^2(1,2) = (-1.2814 - (-1.1273))^2 + (-1.4953 - (-1.2933))^2 = 0.06455$$
$$d^2(2,6) = 1.14415$$
$$d^2(1,6) = 1.72529$$

	Modele	CYL	PUISS
1	Toyota Corolla	-1.2814	-1.4953
2	Citroen GS Club	-1.1273	-1.2933
3	Simca 1300	-0.9292	-0.8389
4	Lada 1300	-0.9292	-0.8389
5	Lancia Beta	-0.9209	-0.1319
6	Alfasud TI	-0.7751	-0.2834
7	Rancho	-0.5219	-0.2329
8	Renault 16 TL	-0.1835	-1.4953
9	Alfetta 1.66	-0.1697	1.2316
10	Fiat 132	-0.1284	0.6761
11	Audi 100	-0.1202	0.0196
12	Mazda 9295	0.3779	-0.0814
13	Peugeot 504	0.4522	-0.2834
14	Princess 1800	0.4577	-0.1319
15	Opel Rekord	0.9558	0.7771
16	Taunus 2000	0.9943	0.6761
17	Datsun 200L	1.0081	1.5346
18	Renault 30	2.8408	2.1911

Données centrées et réduites

	Modele	F1 (89.83%)	F2 (10.17%)
1	Toyota Corolla	1.9635	0.1513
2	Citroen GS Club	1.7117	0.1174
3	Simca 1300	1.2502	-0.0639
4	Lada 1300	1.2502	-0.0639
5	Lancia Beta	0.7444	-0.5580
6	Alfasud TI	0.7484	-0.3477
7	Rancho	0.5337	-0.2044
8	Renault 16 TL	1.1871	0.9276
9	Alfetta 1.66	-0.7509	-0.9909
10	Fiat 132	-0.3873	-0.5689
11	Audi 100	0.0711	-0.0989
12	Mazda 9295	-0.2097	0.3248
13	Peugeot 504	-0.1194	0.5201
14	Princess 1800	-0.2304	0.4169
15	Opel Rekord	-1.2254	0.1263
16	Taunus 2000	-1.1812	0.2250
17	Datsun 200L	-1.7980	-0.3723
18	Renault 30	-3.5581	0.4594

Coordonnées dans le repère factoriel

Si on ne tient compte que de la 1ère composante ($\lambda_1 = 89.83\%$), les dist sont approximées. On constate né que les proximités sont assez bien respectées (globalement).

$$d^2_{\{F_1\}}(1,2) = (1.9335 - 1.711 = 0.06340$$
$$d^2_{\{F_1\}}(2,6) = 0.92783$$
$$d^2_{\{F_1\}}(1,6) = 1.147632$$

Si on tient compte des 2 compos retrouve les distances exactes en individus.

$$d^2_{\{F_1,F_2\}}(1,2) = (1.9635 - 1.7117)^2 + (0.1 = 0.06455$$
$$d^2_{\{F_1,F_2\}}(2,6) = 1.14415$$
$$d^2_{\{F_1,F_2\}}(1,6) = 1.72529$$

Une des questions clés de l'AC définir le nombre de composa à retenir pour obtenir une approximation suffisamment

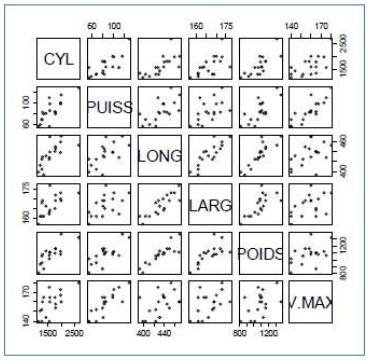
Relations entres variables – Principe de l'ACP (2) – Matrice des corrélations

Le coefficient de corrélation mesure la liaison (linéaire) entre deux variables X_j et X_m

$$r_{jm} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m)}{s_j \times s_m}$$

Matrice des corrélations R sur les données « autos »

CORR	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS		1	0.641	0.521	0.765	0.844
LONG			1	0.849	0.868	0.476
LARG				1	0.717	0.473
POIDS					1	0.478
V.MAX						1



Elle traduit numériquement ce que l'on peut observer dans les graphiques croisés des variables

On peut essayer de la réorganiser manuellement pour mieux faire apparaître les « blocs » de variables mais....

	POIDS	CYL	PUISS	LONG	LARG	V.MAX
POIDS	1.000	0.789	0.765	0.868	0.717	0.478
CYL		1.000	0.797	0.701	0.630	0.665
PUISS			1.000	0.641	0.521	0.844
LONG				1.000	0.849	0.476
LARG					1.000	0.473
V.MAX						1.000

(1) Ce ne sera jamais parfait

(2) La manipulation est inextricable dès que le nombre de variables est élevé

Relations entres variables – Principe de l'ACP (2) – Construction des composantes

Construire la première composante F_1 qui permet de maximiser le carré de sa corrélation avec les variables de la base de données

$$\lambda_1 = \sum_{j=1}^{p} r_j^2(F_1)$$

Habituellement, Inertie totale = Somme des variances des variables
Lorsque les données sont réduites (ACP normée), Inertie totale = Trace(R) = p

$$I_p = p$$

⇒ Part d'inertie expliquée par F_1 = $\dfrac{\lambda_1}{p}$

De nouveau, on observe la décomposition de l'information en composantes non corrélées (orthogonales)

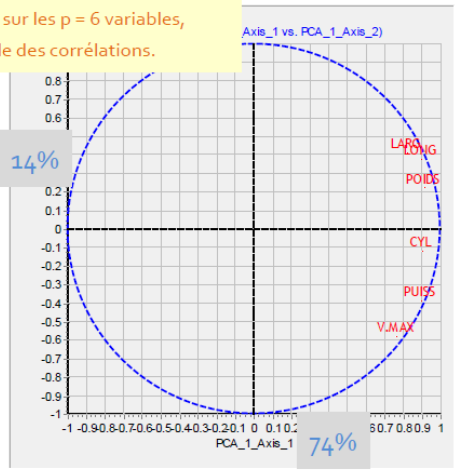
$$\sum_{k=1}^{p} \lambda_k = p$$

Exemple de traitement pour les p = 6 variables de la base de données

Axis	Eigen value	Proportion (%)	Cumulative (%)
1	4.421	73.68%	73.68%
2	0.856	14.27%	87.95%
3	0.373	6.22%	94.17%
4	0.214	3.57%	97.73%
5	0.093	1.55%	99.28%
6	0.043	0.72%	100.00%
Tot.	6	-	-

Relations entres variables – Principe de l'ACP (2) – Approximation des corrélations

ACP sur les p = 6 variables, cercle des corrélations.



Liaison de la variable « poids » avec le 1er axe

$$r_{poids}(F_1) = 0.905 \quad \text{et} \quad r^2_{poids}(F_1) = 0.819$$

La représentation de la variable n'est pas complète, on a besoin d'un second facteur F_2

$$r_{poids}(F_2) = 0.225 \quad \text{et} \quad r^2_{poids}(F_2) = 0.050$$

Si on exploite tous les « p » facteurs

$$\sum_{k=1}^{p} r^2_{poids}(F_k) = 0.819 + 0.050 + \cdots = 1$$

L'ACP produit aussi une approximation dans l'espace des variables (**approximation des corrélations**)
[Ex. si on ne prend en compte que « q = 1 » facteur]

Approximation assez bonne parce que POIDS et CYL sont bien représentées sur le 1er facteur

$$\begin{cases} r_{poids,cyl} = 0.789 \\ r_{poids,cyl}(F_1) = \sum_{k=1}^{q=1} r_{poids}(F_k) \times r_{cyl}(F_k) = 0.90519 \times 0.89346 = 0.809 \end{cases}$$

$$\begin{cases} r_{poids,v.max} = 0.478 \\ r_{poids,v.max}(F_1) = 0.90519 \times 0.75471 = 0.683 \end{cases}$$

Approximation mauvaise parce que V.MAX est mal représentée sur le