Département de Géologie

Statistique appliquée

&

Introduction à la Géostatistique

Filière : Science de la Terre et de l'Univers

Module : Télédétection & Géostatistique

Semestre : S6

Année universitaire : 2020-2021

Compilation:

B. Ouchaou

Table des Matières

Pages

Généralités

I. Introduction 1
II. Terminologie de base 2

Statistique appliquée

I. Analyse unidimensionnelle 4
I.1. Paramètres de position 5
I.2. Paramètres de dispersion 5
I.3. Paramètres de forme 9
I.4. Test de Chi-deux 10
II. Analyse bidimensionnelle 12
II.1. Deux variables qualitatives (Chi-deux du tableau de contingence) 12
II.2. Une variable quantitative et une variable qualitative (test de Fisher) 13
II.3. Deux variables quantitatives 15
II.3.1. Covariance 15
II.3.2. Coefficient de corrélation 15
II.3.3. Droite de régression 16
III. Analyses multidimensionnelles 18
III.1. Généralités 18
III.2. Rappels sur les Matrices 19
III.2.1. Terminologie et opérations simples 19
III.2.2. Déterminant et trace 19
III.2.3.Matrices particulières 20
III.2.4. Valeurs propres et vecteurs propres 21
III.3. Analyse en Composantes principales (ACP) 23
III.3.1. Principes 23
III.3.2. Exemple simple pour assimiler la Méthode 25

Introduction à la Géostatistique

I. Généralités et Historique 27
II. Variographie 28
II.1. Eléments de vocabulaire 30
II.2. Covariogramme et variogramme 30
II.3. Variogramme expérimental 32
II.4. Variogramme théorique 32
III. Krigeage 36
III.1. Définitions 36
III.2. Modèles stationnaires 37
III.2.1. Krigeage simple 37
III.2.2. Krigeage ordinaire 38
III.3. Modèles non stationnaires 40
III.3.1. Généralités 40
III.3.2. Krigeage universel 40
III.4. Comparaison 41
III.5. Validation croisée 42
IV. Références 43

Généralités

I. Introduction

La Statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Elle peut-être définie comme étant l'ensemble des « méthodes ou techniques » permettant de traiter (analyser) des ensembles de données.

Les plus anciennes méthodes sont des méthodes de dénombrement (dénombrement des crues par le égyptiens, dénombrement des populations et des effectifs militaires). Dès la fin du 19ème siècle, la Statistique a connu un grand essor suite aux développements des autres Sciences, particulièrement la Théorie des Probabilités. Depuis les années 60 du 20ème siècle, l'outil informatique a facilité les calculs et les représentations graphiques. Depuis la fin du 20ème siècle, de nombreux logiciels ont été mis à disposition des non mathématiciens. Ce qui n'est pas sans danger dans la mesure où certains utilisateurs non avertis des précautions mathématiques à prendre appliquent des méthodes inadéquates et aboutissent à des résultats et/ou des interprétations aberrants. A titre d'exemple, dans plusieurs travaux, il y a confusion entre calcul et estimation. D'un autre côté, l'on remarque que dans plusieurs cas, au lieu d'adapter l'outil à l'objectif de l'analyse, l'on adapte l'analyse à l'outil à disposition.

On distingue deux branches.

- **Statistique descriptive** (Statistique exploratoire) qui consiste à présenter et analyser les données (représentations graphiques et calculs des caractéristiques numériques de l'échantillon).
- **Statistique inférentielle** (Statistique mathématique, Statistique inductive) dont l'objectif est de préciser un phénomène sur une population à partir d'un échantillon. Elle consiste à induire (inférer) du particulier au général par des approches probabilistes, en se basant sur les caractéristiques numériques de l'échantillon (dernière étape de la Statistique descriptive).

Les deux aspects se complètent plus qu'ils ne s'opposent. La Statistique descriptive précède, en général, la Statistique inférentielle.

La Statistique appliquée (*data analysis en Anglais*) correspond essentiellement à la Statistique descriptive avec insertion de quelques approches mathématiques (Coefficients, tests, intervalle de confiance).

Remarque : il y a souvent une confusion entre l'expression française « analyse des données » et l'expression anglaise « *data analysis* ». L'acception française, **« analyse des données »**, désigne une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles (multivariées) et descriptives, c'est-à-dire un sous-ensemble de la Statistique appliquée. L'acception anglaise, « ***data analysis*** », a un sens plus général et correspond à la «Statistique appliquée » en français.

Ce cours est divisé en deux parties : Statistique appliquée et Introduction à la Géostatistique, précédées de quelques éléments de Terminologie de base. En Statistique appliquée seront développées les Analyses unidimensionnelle et bidimensionnelle, parfois groupées sous le nom de Statistique appliquée élémentaire. Quant aux Analyses multivariées (« Analyse des données » selon l'acception française), un seul cas (ACP) sera abordé. En Géostatistique, on se limitera au krigeage linéaire univarié.

II. Terminologie de base

Les premières statistiques correctement élaborées ont été celles des recensements démographiques. Ainsi, le vocabulaire statistique (population, échantillon, individu) est influencé par celui de la démographie. On évoque ici un certain nombre de termes statistiques très courants et qu'il convient de bien connaître.

Variable : d'un point de vue mathématique, une variable est une application définie sur l'échantillon. En Statistique appliquée, c'est une observation, ou une caractéristique ou un caractère. Les caractères (variables) peuvent-êtres qualitatifs (catégoriels) ou quantitatifs (numériques).

- Variables qualitatives (catégorielles), qui ne se traduisent pas en chiffres (enroulement de coquille, couleur de sédiment, jeu de décrochement, groupes sanguins, sexe, etc.). On a souvent recours à un codage dans les tableaux numériques (exp. décrochement dextre $= 1$; décrochement senestre $= 2$). Cependant, le codage n'est pas une variable quantitative, la variable qu'il code demeure qualitative.

- Variables quantitatives (numériques), qui peuvent être traduites en nombres. Elles sont subdivisées en :

  - Variables continues : grandeurs mesurables (température, pression, teneurs, longueurs, surfaces, volumes, poids, densité, porosité, angles etc.).
  - Variables discontinues = discrètes : grandeurs comptables (nombre de strates, nombre de séismes, nombre de coulées volcaniques, nombre de filons, nombre de loges, nombre de sources etc.).

Individu (ou unité statistique) : on désigne ainsi tout élément étudié.

Données (statistiques) : le terme « données » est très utilisé en Statistique. Il désigne l'ensemble des observations de toutes les variables sur tous les individus. Les données sont, en général, stockées dans un fichier informatique sous forme de tableaux à deux entrées :

- individus en lignes ;
- variables en colonnes.

Attribut : une observation sur un individu (une cellule du tableau des données).

Population : ensemble concerné par une étude statistique. On parle aussi de champ de l'étude.

Enquête (statistique) : c'est l'opération consistant à observer (mesurer, compter ou questionner) l'ensemble des individus étudiés. On distingue :

- Recensement : enquête portant sur l'ensemble des individus de la population. On parle alors d'enquête exhaustive ;
- Sondage : enquête portant sur un sous ensemble de la population. On parle alors d'enquête non exhaustive. L'ensemble des individus concernés par l'étude est appelé « échantillon ».

Echantillon : dans une étude statistique, il est fréquent que l'on n'observe pas la population toute entière. Les observations du phénomène considéré sont donc réalisées sur une partie restreinte de la population, appelée échantillon. Donc, un échantillon est un sous-ensemble de la population sur lequel sont **effectivement** réalisées les observations. En général, on note « **N »** la taille de l'échantillon considéré.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon. Un échantillon est qualifié d'aléatoire lorsque chaque individu de la population a une probabilité connue et non nulle d'appartenir à l'échantillon. On peut distinguer :

- Echantillon global (N) : l'ensemble des individus analysés ;
- Echantillons fractionnés (n) : résultent d'une opération de tri et de séparations de groupes plus ou moins différents. On parle de **Décomposition** de l'échantillon.

Test : consiste à vérifier une information hypothétique. Il s'agit donc d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique. Il existe un très grand nombre de tests statistiques. Le tableau ci-dessous résume les plus utilisés en Sciences naturelles.

Echantillon(s)	Variable(s)	Hypothèse à vérifier	Exemples de Test
1	1	Valeurs extrêmes aberrantes ou non	Dixon Grubbs
1	1	Moyenne observée comparable à la moyenne théorique ou non	Student
1	1	Distribution observée comparable à la distribution normale (Gaussienne) ou non	David Khi-deux*
1	1	Distribution observée comparable à la distribution théorique ou non	Kolmogorov-Smirnov
1	1 (2 séries de mesures sur les mêmes individus)	Moyennes observées comparables ou non	Student (échantillons appariés)
1	2 qualitatives	Interdépendance entre les deux variables	Khi-deux du tableau de contingence*
1	2 qualitatives	Intensité de l'interdépendance entre les deux variables	V de Cramer
1	2 quantitatives	Interdépendance entre les deux variables	Alpha de Cronbach Corrélation de Pearson*
1	Plusieurs (sur les mêmes individus)	Les positions sont identiques ou non	Cochran (Q)
2	1	Les moyennes des deux échantillons sont comparables ou non	Student (échantillons non appariés)
2	1	Les deux échantillons suivent la même distribution ou non	Kolmogorov-Smirnov
2	1 (2 séries de mesures sur les mêmes individus)	Les positions sont identiques ou non	Wilcoxon
2 ou plus	1	Les variances sont similaires ou non	Levene Bartlett
2 ou plus	2 ou plus sur les mêmes individus	Les positions sont identiques ou non	Kruskal-Wallis
2 ou plus	2 (1 quantitative, une qualitative)	Homogénéité des variances	Cochran (C) Fisher (ANOVA)*

Exemples de tests statistiques : (*) exemples traités dans les enseignements de cette année.

Statistique appliquée

I. Analyse unidimensionnelle

Pour chaque individu, on étudie un seul caractère (variable, dimension). Il faut en premier lieu vérifier l'homogénéité de l'échantillon à étudier. Les représentations graphiques sont variées : courbes, histogrammes, diagrammes etc. (voir TD). Les graphiques des analyses unidimensionnelles sont conçus pour être des graphiques plans, sans épaisseur. Certains graphiques proposés par divers logiciels utilisant une épaisseur en perspective sont faux.

On peut utiliser soit les effectifs soit les pourcentages. Généralement, le choix se porte sur les effectifs lorsque l'échantillon est réduit ($n < 100$), sur les pourcentages lorsque l'échantillon est plus important ($n \geq 100$). Toutefois, cette limite n'est que conventionnelle.

Les paramètres numériques de l'échantillon sont de trois types : paramètres de position, paramètres de dispersion et paramètres de forme.

I.1. Paramètres de position

Ils permettent de concevoir de manière facile et concrète la valeur centrale de la variable étudiée.

*** Mode et classe modale (Mo)** : valeur la plus fréquente de la variable ou la classe la plus riche.

*** Médiane et classe médiane (Me)** : la valeur qui se situe au milieu (la moitié des observations lui est inférieure, l'autre moitié lui est supérieure). Dans le cas de l'étude par classe :

$$\text{Me} = \frac{\text{K}}{2} \qquad \text{ou} \qquad \text{Me} = \frac{\text{K}+1}{2}$$ avec K : nombre de classes.

*** Moyenne (μ),** ou moment d'ordre 1, exprime la valeur centrale d'une série de mesures.

Sa formule est : $\mu = \frac{\sum x_i}{N}$ (Moyenne simple) ou $\mu = \frac{\sum f_i x_i}{N}$ (Moyenne pondérée)

On peut également utiliser la Médiane (Me) et la somme des écarts à la médiane selon la formule :

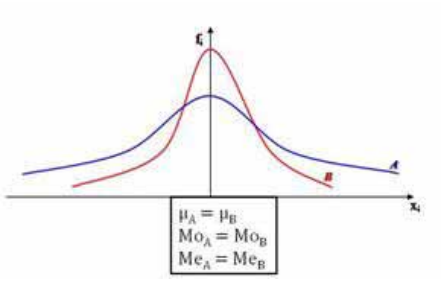
$$\mu = Me + \frac{\sum f_i (x_i - Me)}{N}$$

La signification de la moyenne peut-être faussée par quelques valeurs très grandes ou très petites par rapport à la plus part des observations, en particulier dans le cas de séries très dissymétriques**.** Ainsi, dans certains cas, pour minimiser l'effet des valeurs extrêmes, on ne tient pas compte de la plus petite valeur (x_1) et de la plus grande valeur (x_n). On parle alors de moyenne tronquée.

$$\mu = \frac{\sum x_i - (x_1 + x_n)}{N - 2} \qquad \text{ou} \qquad \mu = \frac{\sum f_i x_i - (x_1 + x_n)}{N - 2}$$

I.2. Paramètres de dispersion

Comme le montre l'illustration ci-contre, deux échantillons différents peuvent avoir les mêmes paramètres de position. Donc, les paramètres de position ne renseignent pas sur la répartition (dispersion) des attributs, d'où le recours à des paramètres de dispersion qui concernent la manière dont les différentes observations (valeurs d'une variable) se dispersent (se répartissent) dans l'échantillon.



*** Etendue** : écart entre la plus grande et la plus petite des observations d'une variable (Max – Min).

* Intervalle interquartiles

<u>Rappel sur les quantiles</u>

Les quantiles sont les valeurs qui divisent un jeu de données en intervalles contenant le même nombre de données. On distingue :

Quartiles (25%, 50% et 75%)	**4 groupes de 25%**	$Me = Q_2$
Déciles (10%, 20% …….. 90%)	10 groupes de 10%	$Me = D_5$
Vingtiles (5%, 10% ……95%)	20 groupes de 5%	$Me = V_{10}$
Centiles (1%, 2% …… 99%)	100 groupes de 1%	$Me = C_{50}$

Certains auteurs utilisent les Quintiles (20%, 40%, 60% et 80%, soit 5 groupes de 20 %). L'inconvénient est que la médiane ne coïncide avec aucun quintile.

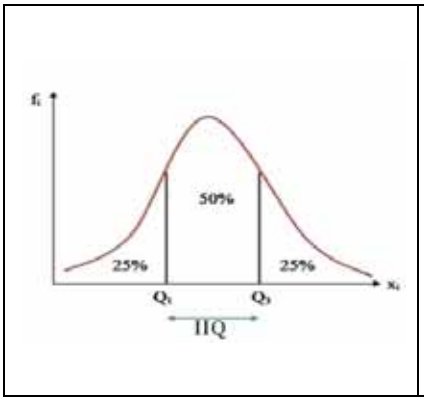
Les quantiles les plus utilisés en Géologie sont les quartiles.

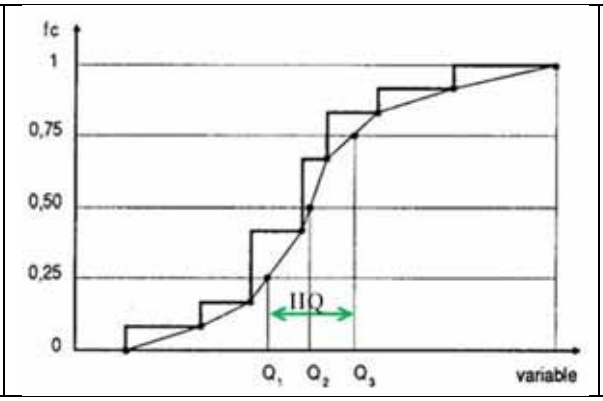
Premier quartile	$Q_1 = x_{1/4}$	25%
Deuxième quartile	$Q_2 = x_{2/4}$ = médiane	50%
Troisième quartile	$Q_3 = x_{3/4}$	75%

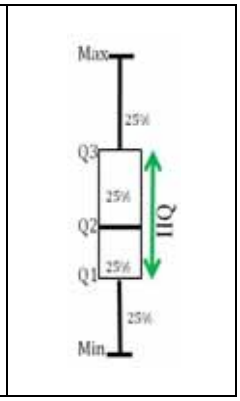
<u>L'intervalle interquartiles</u> mesure l'écart qui existe entre les deux valeurs qui bornent la partie centrale de la distribution concernant 50% des fréquences. Donc, il correspond à l'écart entre le troisième quartile et le premier quartile selon la formule :

$$IIQ = Q_3 - Q_1$$

Il permet d'éliminer l'influence des valeurs qui se trouvent aux extrémités de la distribution et dont certaines peuvent être aberrantes. L'intervalle interquartiles est utile pour certains graphiques notamment la courbe cumulative et le diagramme en boite (ou boite à moustaches *(boxplot* en anglais*)*).







*** Variance :** c'est la moyenne des **carrés des écarts** à la moyenne ou le moment centré d'ordre 2. Son but est d'avoir une idée sur la variabilité d'une variable au sein de l'échantillon étudié. La variance dépend de l'étendue de l'échantillon et des écarts à la moyenne mais ne dépend pas de la moyenne elle même. La variance est toujours positive ou nulle puisque c'est une moyenne des **carrés** des écarts à la moyenne. Elle ne peut être nulle que si tous les écarts sont nuls, c'est-à-dire si toutes les observations sont égales. Sa formule est :

$$V = m_2 = \frac{\sum(x_i - \mu)^2}{N} \quad ou \quad V = \frac{\sum f_i(x_i - \mu)^2}{N} \quad ou\, V = \frac{SC_t}{N}$$

Avec SC_t = somme des carrés des écarts à la moyenne.

*** Ecart type** : son intérêt est qu'il s'exprime en même unité que la variable, ce qui n'est pas le cas de la variance. Sa formule est : $\sigma = \sqrt{V}$

*** Coefficient de variation :** c'est un coefficient sans unité. Il s'exprime en % et permet de comparer les dispersions de variables en unités de mesures différentes. Sa formule est :

$$C.V. = 100\frac{\sigma}{\mu}.$$

*** Marge d'erreur :** un écart type (σ) d'une valeur quelconque n'a pas la même signification pour un échantillon de quelques individus et un échantillon de plusieurs milliers d'individu. La marge d'erreur permet de tenir compte de la taille de l'échantillon (N) selon la formule :

$$ME = t_\alpha \frac{\sigma}{\sqrt{N}}$$

Avec t_α le coefficient de fonction du niveau de confiance, à lire sur la table de Student (page suivante) lorsque le degré de liberté est très grand (∞ sur la table de Student).

Les niveaux de confiances (1 – α) les plus utilisés en Géologie sont :

- 95%, donc α = 0,050 (t_α = 1,96)
- 99%, donc α = 0,010 (t_α = 2,5758, généralement majoré à 2,58).

Table de Student

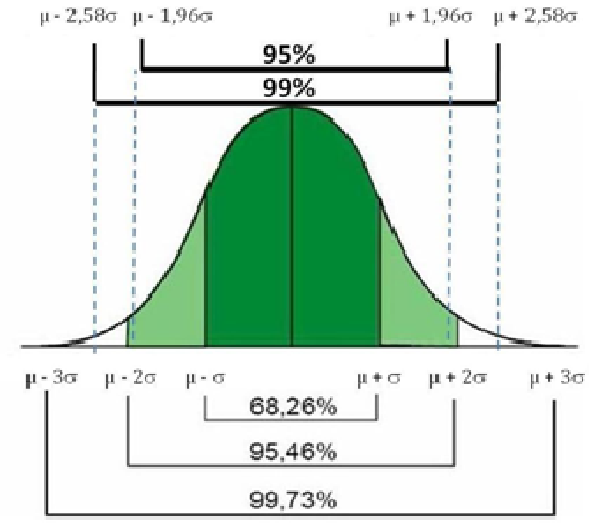
ν \ α	0,900	0,500	0,300	0,200	0,100	0,050	0,020	0,010	0,001
1	0,1584	1,0000	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6193
2	0,1421	0,8165	1,3862	1,8856	2,9200	4,3027	6,9646	9,9248	31,5991
3	0,1366	0,7649	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,9240
4	0,1338	0,7407	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,1322	0,7267	1,1558	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0,1311	0,7176	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,1303	0,7111	1,1192	1,4149	1,8946	2,3646	2,9980	3,4995	5,4079
8	0,1297	0,7064	1,1081	1,3968	1,8595	2,3060	2,8965	3,3554	5,0413
9	0,1293	0,7027	1,0997	1,3830	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,1289	0,6998	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,1286	0,6974	1,0877	1,3634	1,7959	2,2010	2,7181	3,1058	4,4370
12	0,1283	0,6955	1,0832	1,3562	1,7823	2,1788	2,6810	3,0545	4,3178
13	0,1281	0,6938	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,1280	0,6924	1,0763	1,3450	1,7613	2,1448	2,6245	2,9768	4,1405
15	0,1278	0,6912	1,0735	1,3406	1,7531	2,1314	2,6025	2,9467	4,0728
16	0,1277	0,6901	1,0711	1,3368	1,7459	2,1199	2,5835	2,9208	4,0150
17	0,1276	0,6892	1,0690	1,3334	1,7396	2,1098	2,5669	2,8982	3,9651
18	0,1274	0,6884	1,0672	1,3304	1,7341	2,1009	2,5524	2,8784	3,9216
19	0,1274	0,6876	1,0655	1,3277	1,7291	2,0930	2,5395	2,8609	3,8834
20	0,1273	0,6870	1,0640	1,3253	1,7247	2,0860	2,5280	2,8453	3,8495
21	0,1272	0,6864	1,0627	1,3232	1,7207	2,0796	2,5176	2,8314	3,8193
22	0,1271	0,6858	1,0614	1,3212	1,7171	2,0739	2,5083	2,8188	3,7921
23	0,1271	0,6853	1,0603	1,3195	1,7139	2,0687	2,4999	2,8073	3,7676
24	0,1270	0,6848	1,0593	1,3178	1,7109	2,0639	2,4922	2,7969	3,7454
25	0,1269	0,6844	1,0584	1,3163	1,7081	2,0595	2,4851	2,7874	3,7251
26	0,1269	0,6840	1,0575	1,3150	1,7056	2,0555	2,4786	2,7787	3,7066
27	0,1268	0,6837	1,0567	1,3137	1,7033	2,0518	2,4727	2,7707	3,6896
28	0,1268	0,6834	1,0560	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739
29	0,1268	0,6830	1,0553	1,3114	1,6991	2,0452	2,4620	2,7564	3,6594
30	0,1267	0,6828	1,0547	1,3104	1,6973	2,0423	2,4573	2,7500	3,6460
40	0,1265	0,6807	1,0500	1,3031	1,6839	2,0211	2,4233	2,7045	3,5510
60	0,1262	0,6786	1,0455	1,2958	1,6706	2,0003	2,3901	2,6603	3,4602
80	0,1261	0,6776	1,0432	1,2922	1,6641	1,9901	2,3739	2,6387	3,4163
120	0,1259	0,6765	1,0409	1,2886	1,6577	1,9799	2,3578	2,6174	3,3735
∞	0,1257	0,6745	1,0364	1,2816	1,6449	1,9600	2,3263	2,5758	3,2905

*** Intervalle de confiance :** il mesure le degré de précision que l'on a sur les estimations issues de l'échantillon. Autrement dit, il permet d'avoir une idée sur la moyenne de la population à laquelle appartient l'échantillon étudié. On passe donc de la statistique descriptive à la Statistique inférentielle.

La formule générale est :

$$I.C. = \left[\mu - t_{\propto}\frac{\sigma}{\sqrt{N}};\ \mu + t_{\propto}\frac{\sigma}{\sqrt{N}}\right] = \left[\mu \pm t_{\propto}\frac{\sigma}{\sqrt{N}}\right] = [\mu \pm ME]$$

On dit que le produit ($t_{\alpha}\frac{\sigma}{\sqrt{N}}$) représente la marge d'erreur (ME) entre les données d'un sondage (échantillon) et les données du champ d'étude (population entière).



Les intervalles de confiances les plus utilisés en Géologie en particulier, en Sciences naturelles en général, sont :

$$I.C \text{ à } 95\% = \mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}$$

$$I.C \text{ à } 99\% = \mu \pm 2{,}58\frac{\sigma}{\sqrt{N}}$$

L'intervalle $[\mu \pm \frac{\sigma}{\sqrt{N}}]$ est l'intervalle de confiance de la moyenne de la population à 68,26%.

L'intervalle $[\mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}]$ est l'intervalle de confiance de la moyenne de la population à 95%.

L'intervalle $[\mu \pm 2\frac{\sigma}{\sqrt{N}}]$ est l'intervalle de confiance de la moyenne de la population à 95,46%.

L'intervalle $[\mu \pm 2{,}58\frac{\sigma}{\sqrt{N}}]$ est l'intervalle de confiance de la moyenne de la population à 99%.

L'intervalle $[\mu \pm 3\frac{\sigma}{\sqrt{N}}]$ est l'intervalle de confiance de la moyenne de la population à 99,73%.

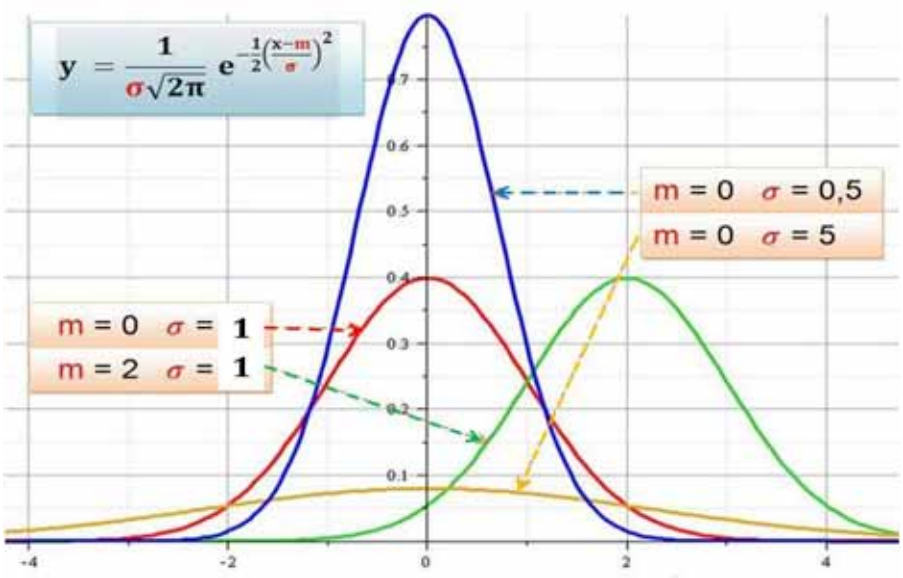
Exemples de lecture pour l'I.C. à 95% ($t_\alpha = 1{,}96$).

* L'intervalle $[\mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}]$ est un intervalle de valeur qui a 95% de chance de contenir la vraie valeur de la moyenne de la population.
* L'intervalle $[\mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}]$ représente la fourchette de valeurs à l'intérieur de laquelle nous sommes certains à 95% de trouvez la vraie moyenne de la population.
* Il y a 95 chances sur 100 (19 sur 20) que la moyenne de la population soit comprise dans l'intervalle $[\mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}]$.
* Si nous tirons d'autres échantillons de la même population, dans 95 cas sur100 (19 cas sur 20), la moyenne se trouvera dans l'intervalle $[\mu \pm 1{,}96\frac{\sigma}{\sqrt{N}}]$.

Rappel : les paramètres de dispersion concernent les variables et non pas les individus.

I.3. Paramètres de forme

Les paramètres de forme donnent l'allure de la répartition des individus d'un échantillon en comparaison avec une distribution normale ou gaussienne (m = 0 et σ = 1 de la figure ci-dessous). On distingue généralement deux catégories : les mesures d'asymétrie (*Skewness* en anglais) et les mesures d'aplatissement (*Kurtosis* en anglais).



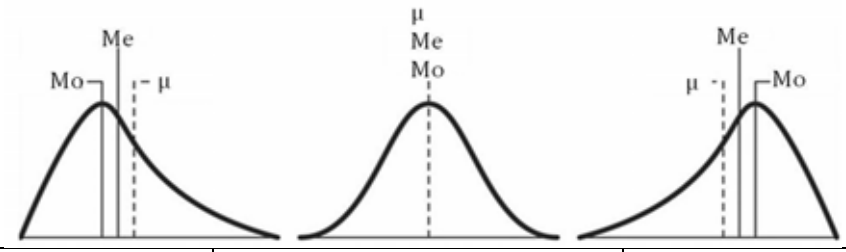
I.3.1. Mesures d'asymétrie :

Les coefficients d'asymétrie sont des grandeurs sans unité. On distingue :

Coefficient de Pearson : $b_1 = \frac{\mu - Mo}{\sigma}$

Coefficient de Fisher : $\gamma_1 = \frac{m_3}{\sigma^3}$ avec $m_3 = \frac{1}{N}\sum fi(x_i - \mu)^3$

C'est le rapport entre le moment centré d'ordre 3 (m_3) et le cube de l'écart-type. Le coefficient de Fisher a toujours le même signe que m_3.

Asymétrie positive	Symétrie	Asymétrie négative
		
μ > Me > Mo	μ = Me = Mo	μ < Me < Mo
Distribution asymétrique à gauche (étalée à droite)	Distribution symétrique	Distribution asymétrique à droite (étalée à gauche)
$b_1 > 0$	$b_1 = 0$	$b_1 < 0$
$\gamma_1 > 0$	$\gamma_1 = 0$	$\gamma_1 < 0$

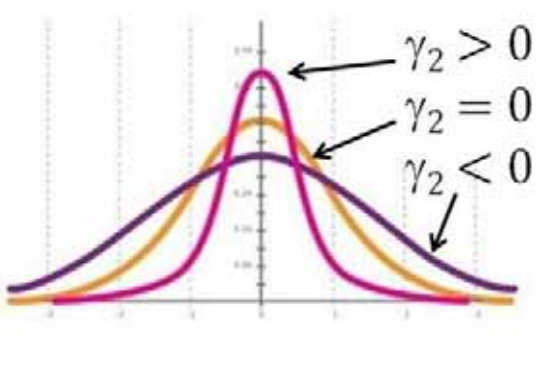
On considère que l'échantillon suit une loi normale à 95 % lorsque la valeur de son asymétrie (γ_1) est comprise entre -2 et +2.

I.3.2. Mesures d'aplatissement : elles sont basées sur le moment centré d'ordre 4 (m_4)

$$m_4 = \frac{1}{N}\sum fi(x_i - \mu)^4$$

Coefficient d'aplatissement de Pearson (*Kurtosis non normalisé*) souvent noté b_2 $\quad b_2 = \frac{m_4}{\sigma^4}$

Coefficient d'aplatissement de Fisher (*Kurtosis normalisé*) souvent noté γ_2 $\quad \gamma_2 = \frac{m_4}{\sigma^4} - 3$



Tant que ces coefficients sont fort tant que la courbe sera effilée et donc grande homogénéité de l'échantillon.

$\gamma_2 > 0$ ($b_2 > 3$) : distribution leptokurtique
$\gamma_2 = 0$ ($b_2 = 3$) : distribution mesokurtique (normale)
$\gamma_2 < 0$ ($b_2 < 3$) : distribution platykurtique

Remarque étymologique : en grec
Kurtos = courbe ou arrondi ;
Lepto = mince ;
Platy = large.

On considère que l'échantillon suit une loi normale à 95 % lorsque la valeur de son coefficient d'aplatissement (γ_2) est comprise entre -2 et + 2.

I.4. Test de Chi-deux (χ^2) : l'objectif est de vérifier la distribution des individus dans l'échantillon en comparant le Chi-deux calculé (χ_c^2) au Chi-deux théorique (χ_t^2). Autrement dit, la question qui se pose est de savoir à partir de quand peut-on dire que les variations observées sont dues au hasard, et à partir de quand peut-on estimer qu'elles sont dues à un ou plusieurs facteur(s) autre que le hasard ?

- Si $\chi_c^2 < \chi_t^2$: la distribution est due au hasard.
- Si $\chi_c^2 > \chi_t^2$: il y a un ou plusieurs facteur(s), autre que le hasard, qui contrôle(nt) la distribution.

Le χ_t^2 est à lire dans le tableau du Chi-deux (page suivante) pour un degré de liberté (dl = N -1).

La formule permettant de calculer le χ_c^2 est : $\chi_c^2 = \sum \frac{(F_o - F_t)^2}{F_t}$

Avec: F_o les fréquences observées et F_t les fréquences théoriques.

Le calcul de la fréquence théorique est basé sur la densité de probabilité de la loi normale selon la formule :

$$F_t = 100 * Ec * \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Avec : Ec = étendue de la classe

LOI DU KHI-DEUX AVEC k DEGRÉS DE LIBERTÉ
QUANTILES D'ORDRE $1-\alpha$

k	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.94	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.81	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Remarques.

- L'effectif minimal admis en Statistique pour le calcul du chi-deux est de 5 individus par classe. Par conséquent, dans certains cas, il faut procéder à des groupements de classes pour que tous les effectifs utilisés soient ≥ 5.
- Evitez la confusion entre fréquences et effectifs.

II. Analyse bidimensionnelle

L'objectif essentiel de la Statistique descriptive bidimensionnelle est de mettre en évidence une éventuelle variation simultanée des deux variables. Sur chaque individu, on étudie deux variables. Le but est d'avoir une idée sur l'**interdépendance** des caractères (exp. : la taille et le poids sont interdépendants ; la taille et la couleur des yeux ne le sont pas).

Tous les paramètres de l'étude unidimensionnelle peuvent être utilisés séparément pour chaque variable, mais certains d'entre eux n'ont pas une grande importance dans l'analyse bidimensionnelle. En outre, les paramètres de l'analyse bidimensionnelle dépendent de la nature des deux variables étudiées. Trois cas seront distingués :

- Deux variables qualitatives ;
- Une variable quantitative et une variable qualitative ;
- Deux variables quantitatives.

II.1. Deux variables qualitatives

Le plus souvent, les données sont présentées dans un tableau a double entrée (tableau croisé), appelé tableau de contingence qui indique les effectifs du croisement entre les deux variables qualitatives. L'analyse statistique consiste à chercher si les effectifs observés sont assez proches ou assez éloignés des effectifs théoriques. Le test qui permet de répondre à cette question est le test du χ^2 du tableau de contingence. On peut dire donc que le test du χ^2 permet de déterminer la probabilité que les lignes et les colonnes d'un tableau de contingence sont indépendantes.

On commence par le calcul des marges du tableau (les totaux des lignes et des colonnes). La deuxième étape correspond aux calculs des pourcentages théoriques. On suppose que les deux variables sont indépendantes. On calcul les % des colonnes sans tenir compte des lignes et les % des lignes sans tenir compte des colonnes. Le % théorique de chaque cellule est le produit des % du croisement qu'elle représente. Une fois les % théoriques des cellules calculés on en déduit les effectifs théoriques. Le $\boldsymbol{\chi}_p^2$ de **chaque cellule**, dit Chi-deux partiel est :

$$\boldsymbol{\chi}_{\mathrm{p}}^2 = \frac{(\mathrm{N_o} - \mathrm{N_t})^2}{\mathrm{N_t}} \text{ avec } \mathrm{N_o} \text{ : effectif observé et } \mathrm{N_t} \text{ : effectif théorique.}$$

Le $\boldsymbol{\chi}_c^2$ **du tableau** est la somme des $\boldsymbol{\chi}_p^2$ de ses cellules $\boldsymbol{\chi}_c^2 = \sum \boldsymbol{\chi}_{\mathrm{p}}^2 = \sum \frac{(\mathrm{N_o} - \mathrm{N_t})^2}{\mathrm{N_t}}$

On compare le $(\boldsymbol{\chi}_c^2)$ avec le $(\boldsymbol{\chi}_t^2)$ sachant que le degré de liberté (dl) est égale à (Nombre de ligne – 1)*(Nombre de colonne – 1) :

- si $\boldsymbol{\chi}_c^2 < \boldsymbol{\chi}_t^2$: la distribution est due au hasard ;
- si $\boldsymbol{\chi}_c^2 > \boldsymbol{\chi}_t^2$: il y a un ou plusieurs facteur(s), autre que le hasard, qui contrôle(nt) la distribution.

Remarque : c'est un test extrêmement sensible aux effectifs.

Exemple : nous cherchons à comparer les abondances relatives de deux métaux. Nous avons effectué 350 sondages en trois secteurs miniers. Le tableau ci-dessous résume les nombres de sondages où l'un ou l'autre métal est dominant.

	Secteur 1	Secteur 2	Secteur 3
Métal 1 dominant	64	62	22
Métal 2 dominant	114	74	14

1. Tracez le diagramme des profils colonnes.
2. Démontrez si la répartition est aléatoire ou s'il y a un ou plusieurs facteur(s), autre que le hasard, qui contrôle(nt) cette répartition.

II.2. Une variable quantitative et une variable qualitative

Lorsque nous avons à comparer une variable numérique (quantitative) pour un nombre de groupes (catégories ou classes) ≥ 2, nous utilisons la technique d'analyse statistique connue sous le nom d'ANOVA, c'est-à-dire « Analyse de la variance », soit « *Analysis of variance* » en anglais.

Eléments de vocabulaire pour l'ANOVA

F : valeur calculée pour le test de Fisher	N : effectif total
V_t : variance totale	n_c : effectifs des classes
V_b : variance interclasses (*Between groupes variance*)	K : nombre de classes
V_w : variance intra-classes (*Within groupes variance*)	μ_g : moyenne générale
SC : somme des carrés des écarts à la moyenne	μ_c : moyennes des classes
CM : carré moyen des écarts, qui se réfère au degré de liberté	dl : degré de liberté

Variance interclasses (V_b) : elle reflète la variabilité observée entre les moyennes des différentes classes et la moyenne générale. On l'appelle également « variance expliquée ». Donc elle mesure la variabilité entre les différents groupes (fluctuations entre les groupes).

- Si elle est faible : pas de grandes différences entre les différents groupes.
- Si elle est forte : grande différence entre les différents groupes.

Variance intra-classes (V_w) : c'est le résidu de la variance « variance résiduelle ». Elle reflète les fluctuations individuelles (variance mesurée à l'intérieur des groupes).

- Si elle est faible : chaque groupe est constitué d'individus semblables, donc grande homogénéité.
- Si elle est grande : les groupes rassemblent en leur sein des individus assez différents, donc grande hétérogénéité.

L'ANOVA permet de décomposer la variance totale en variance interclasses (intergroupes) et variance intra-classes (intra-groupes). Cependant, L'ANOVA ne divise pas la variance en parties additives. Par contre, elle permet de diviser la somme des carrés des écarts à la moyenne en parties additives.

$$(V_t \ddagger V_b + V_W) \text{ alors que } (SC_t = SC_b + SC_w).$$

L'un des Tests qui permettent de dire s'il y a une différence statistiquement significative ou non entre les différents groupes (classes) est le test de Fisher (F) ou test d'égalité des variances. C'est le seul qui sera traité dans ce cours. Le tableau ci-dessous résume les éléments nécessaires pour le calculer.

Variation	SC	dl	CM	F
Totale	$SC_t = \sum (x_i - \mu_g)^2$	N - 1	$\frac{SC}{dl}$	$\frac{CM_b}{CM_w} = \frac{\frac{SC_b}{K-1}}{\frac{SC_W}{N-K}}$
Interclasse (expliquée)	$SC_b = \sum n_c(\mu_c - \mu_g)^2$	K - 1		
Intra-classe (résiduelle)	$SC_w = \sum\sum(x_i - \mu_c)^2$	N - K		

Remarques

Dans la pratique, puisque la somme des carrés des écarts est additive, nous pouvons déduire SC_w sans passer par la formule $\sum\sum(x_i - \mu_c)^2$

$$SC_t = SC_b + SC_w \Rightarrow \mathbf{SC_w = SC_t - SC_b}$$

Le dl intra-classe est additif

$dl_w = (n_{c1} -1) + (n_{c2} -1) + ... + (n_{ck} - 1) = (n_{c1} + n_{c2} + ...+ n_{ck}) - K = N - K$

Une fois la valeur de F connue, on la compare à la valeur critique sur la table de Fisher au croisement de la colonne (K-1) et la ligne (N-K). La table de Fisher la plus utilisée est celle du 95ème centile (soit $\alpha = 0,05 = 5\%$).

Table de Fisher-Snedecor, α = 5 % (95e centile)

ν_2 (dén.)	ν_1 (numérateur)																			
	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	80	100	200	500	1 000
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	248.02	250.10	251.14	251.77	252.20	252.72	253.04	253.68	254.06	254.19
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46	19.47	19.48	19.48	19.48	19.49	19.49	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75	5.72	5.70	5.69	5.67	5.66	5.65	5.64	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50	4.46	4.44	4.43	4.41	4.41	4.39	4.37	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81	3.77	3.75	3.74	3.72	3.71	3.69	3.68	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.34	3.32	3.30	3.29	3.27	3.25	3.24	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08	3.04	3.02	3.01	2.99	2.97	2.95	2.94	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.83	2.80	2.79	2.77	2.76	2.73	2.72	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.64	2.62	2.60	2.59	2.56	2.55	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.97	1.95	1.92	1.91	1.88	1.86	1.85
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.79	1.76	1.74	1.71	1.70	1.66	1.64	1.63
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74	1.69	1.66	1.64	1.61	1.59	1.55	1.53	1.52
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.78	1.69	1.63	1.60	1.58	1.54	1.52	1.48	1.46	1.45
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.75	1.65	1.59	1.56	1.53	1.50	1.48	1.44	1.41	1.40
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.72	1.62	1.57	1.53	1.50	1.47	1.45	1.40	1.37	1.36
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.70	1.60	1.54	1.51	1.48	1.45	1.43	1.38	1.35	1.34
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.69	1.59	1.53	1.49	1.46	1.43	1.41	1.36	1.33	1.31
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.68	1.57	1.52	1.48	1.45	1.41	1.39	1.34	1.31	1.30
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.62	1.52	1.46	1.41	1.39	1.35	1.32	1.26	1.22	1.21
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.61	1.50	1.43	1.39	1.36	1.32	1.30	1.23	1.19	1.17
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.59	1.48	1.42	1.38	1.35	1.30	1.28	1.21	1.16	1.14
1 000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.58	1.47	1.41	1.36	1.33	1.29	1.26	1.19	1.13	1.11
2 000	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.58	1.46	1.40	1.36	1.32	1.28	1.25	1.18	1.12	1.09

Si F est inférieur à la valeur critique, les variances de la variable numérique ne sont pas statistiquement différentes en comparant les différentes classes.

Si F est supérieur à la valeur critique, cela signifie que les variances de la variable numérique sont trop différentes pour que les différentes classes puissent être considérées comme homogènes.

II.3. Deux variables quantitatives

Il est possible de représenter l'ensemble des données sur un graphique plan : « Diagramme de dispersion » ou « Nuage de points ». Les paramètres sont : Covariance ; Coefficient de corrélation et Régression linéaire.

II.3.1. Covariance (Cov)

La covariance est la généralisation bidimensionnelle de la variance. C'est un indice symétrique :

$$\mathrm{Cov}_{(x,y)} = \mathrm{Cov}_{(y,x)}.$$

Elle est basée sur l'écart entre la Somme des produits et le Produit des sommes.

$$\mathrm{Cov}_{(x,y)} = \frac{1}{N}\left(\text{Somme des produits} - \frac{\text{Produit des sommes}}{N}\right)$$

ou

$$Cov_{(x,y)} = \frac{1}{N}\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{N}\right)$$

Elle peut prendre toute valeur réelle (négative, nulle ou positive ; petite ou grande). Toutefois ; $Cov^2_{(x,y)} \leq Var_{(x)} * Var_{(y)})$. C'est ce qu'on appelle l'Inégalité de Cauchy-Schwarz.

II.3.2. Coefficient de corrélation (r)

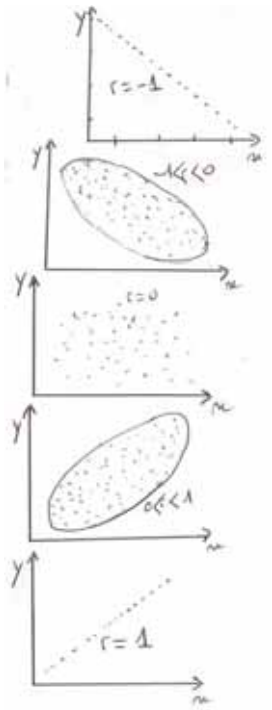
Comme la variance, la covariance n'a pas de signification concrète. Dans le cas de la variance, on doit passer à l'écart-type pour avoir un indicateur interprétable ; dans celui de la covariance, il faudra passer au coefficient de corrélation. On l'appelle également Coefficient de Pearson. Il est indépendant des unités de mesure. Sa formule est :

$$\boldsymbol{r_{(x,y)} = \frac{Cov_{(x,y)}}{\sigma_x \sigma_y}}$$

Sa valeur est comprise entre **-1 et 1**. Son signe indique le sens de la liaison. Sa valeur absolue indique l'intensité de la liaison. Tant que la valeur absolue tend vers 1, tant que la corrélation est forte. Tant que la valeur absolue tend vers 0, tant que la corrélation est faible.

On considère que la corrélation entre deux variables est forte quand la valeur absolue de r est supérieure à 0,5.

Corrélation	négative	positive
Faible	-0,5< r < 0	0 < r < 0,5
Forte	-1 < r < -0,5	0,5 < r < 1



r = - 1 : corrélation négative très forte. Les points sont alignés.

-1 < r < 0 : corrélation négative. Les points sont organisés en un nuage orienté. Au fur et à mesure que les valeurs de la variable prise en abscisse augmentent, celles de la variable prise en ordonnées diminuent.

r = 0 : pas de corrélation. Les deux variables ne sont pas interdépendantes. Les points constituent un nuage diffus, non orienté.

0 < r < 1 : corrélation positive. Les points sont organisés en un nuage orienté. Au fur et à mesure que les valeurs de la variable prise en abscisse augmentent, celles de la variable prise en ordonnées augmentent aussi.

r = 1 : corrélation positive très forte. Les points sont alignés.

Coefficient de détermination. C'est le carré du coefficient de corrélation (R^2). Il est toujours compris entre 0 et 1. Il indique l'intensité de l'interdépendance entre les deux variables indépendamment du sens de cette liaison. Tant que R^2 tend vers 1, tant que l'interdépendance des deux variables est forte. Tant que R^2 tend vers 0, tant que l'interdépendance des deux variables est faible.

II.3.3. Droite de régression (Régression linéaire)

Une corrélation entre les deux variables (numériques) de l'analyse bidimensionnelle signifie que la variable prise en ordonnées (y) varie en fonction de la variable prise en abscisses (x). On dit que « x » est une variable causale (elle cause la variation de y) ou indépendante. La variable « y » est dite variable dépendante (elle dépend de x). Du point de vue Mathématique, on dit que « y » est fonction de « x ». Cette fonction s'appelle la régression de « y » sur « x ». Son équation est :

$$y = ax + b.$$

Cette équation est celle d'une droite. D'où le nom Droite de régression. C'est la droite théorique la plus proche de tous les points du nuage.

a : est la pente de la droite, ou le taux de variation du caractère dépendant (ordonnées). Sa formule est :

$$a = \frac{Cov_{(x,y)}}{\sigma_x^2} = \frac{Cov_{(x,y)}}{Var_x}$$

b : est la valeur à l'origine (valeur théorique de y pour x = 0). Pour la calculer, on suppose que le centre du nuage de points (barycentre) est occupé par les moyennes des deux variables. Donc :

$\mu_y = a\mu_x + b$ et par conséquent $b = \mu_y - a\mu_x$

Lorsque la valeur absolue de a ($|a|$) est égale à 1, on parle d'Isométrie. Si la $|a|$ est différente de 1, on parle d'Allométrie.

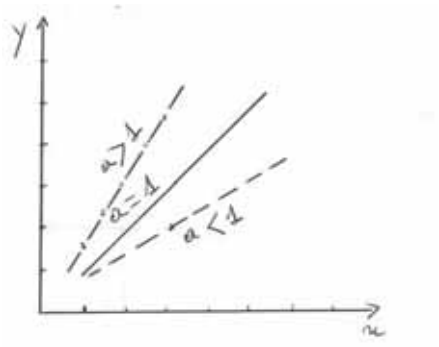
Dans le cas de Corrélation positive, le taux de variation (a) s'appelle « taux d'accroissement ». Il permet de comparer la vitesse d'accroissement des deux variables. Trois cas sont différenciés :

<u>a = 1 : Isométrie</u>, les deux variables s'accroissent à la même vitesse.

<u>a ≠ 1 : Allométrie</u>, les deux variables s'accroissent à des vitesses différentes.

<u>a < 1 : allométrie négative</u> = allométrie minorante. La variable dépendante (ordonnées) s'accroit moins vite que la variable indépendante (abscisse).

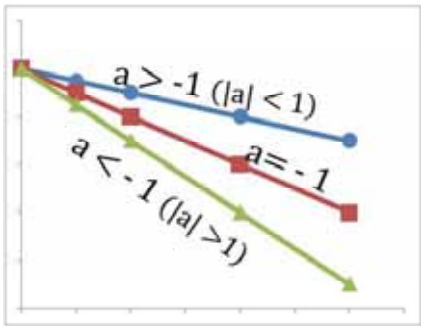
<u>a > 1 : allométrie positive</u> = allométrie majorante. La variable dépendante s'accroit plus vite que la variable indépendante.



Dans le cas de Corrélation négative, le taux de variation (a) s'appelle « taux de décroissement ». Il permet de comparer la vitesse de décroissement de la variable dépendante (y) avec la vitesse d'accroissement de la variable indépendante (x). Là aussi, trois cas sont différenciés.

<u>a = - 1 : Isométrie</u>, la vitesse de décroissement de la variable prise en ordonnée est égale à la vitesse d'accroissent de la variable prise en abscisses.

<u>a ≠ - 1 : Allométrie</u>, la vitesse de décroissement de la variable prise en ordonnée est différente de la vitesse d'accroissent de la variable prise en abscisses.



<u>a < -1 : allométrie positive</u> = allométrie majorante ($|a| > 1$). Le décroissement de la variable dépendante (ordonnées) est plus rapide que l'accroissement la variable indépendante (abscisse).

<u>-1 < a <0 : allométrie négative</u> = allométrie minorante($|a| < 1$). Le décroissement de la variable dépendante (ordonnées) est moins rapide que l'accroissement de la variable indépendante (abscisse).

III. Analyses multidimensionnelles

III.1. Généralités

On désigne par Analyses Multidimensionnelles (*Multivariate Analysis*) l'ensemble des méthodes de la Statistique descriptive (exploratoire) permettant de traiter simultanément un nombre quelconque de variables. Elles consistent à chercher des facteurs en nombre réduit et résumant le mieux possible les données considérées. L'objectif est de revenir à un espace de dimension réduite en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possibles des données initiales. Pour qu'une variable soit intégrée dans l'analyse, sa distribution doit montrer une certaine variance.

Les bases théoriques de ces méthodes sont anciennes et sont principalement issues de « psychomètres » américains : Spearman (1904) pour l'Analyse en Facteurs ; Hotteling (1935) pour l'Analyse en Composantes Principales et l'Analyse Canonique ; Hirschfeld (1935) pour l'Analyse des Correspondances. Leur emploi ne s'est généralisé qu'avec la diffusion des moyens de calcul dans le courant des années 1960. Actuellement, leur utilisation est intimement liée à l'outil informatique. On distingue deux grandes familles :

Méthodes de classification :

* Classification ascendante hiérarchique ;
* Algorithmes de réallocation dynamique ;
* Cartes de Kohonen (réseaux de neurones).

Méthodes factorielles, elles sont très variées. Les plus utilisées sont :

*ACP : Analyse en Composantes Principales (variables quantitatives) ;
*AFD : Analyse Factorielle Discriminante (1 variable qualitative, n variables quantitatives) ;
* ACM : Analyse des Correspondances Multiples (variables qualitatives).

Remarques.

- l'Analyse Factorielle des Correspondances simples (AFC) portant sur 2 variables qualitatives est parfois considérée comme une analyse multidimensionnelle.
- Pour qu'une structure factorielle soit stable, elle doit avoir été vérifiée sur un minimum de cas. La règle veut qu'il y ait un minimum de 5 cas par variable.
- Il faut faire particulièrement attention à la tendance qu'ont certains chercheurs à donner aux facteurs des noms qui font du sens et qui impressionnent mais qui ne reflètent pas ce qui a été mesuré.

Dans cet enseignement, on se limitera à un seul exemple (ACP). Auparavant, nous reviendrons sur quelques rappels relatifs aux matrices, nécessaires pour la compréhension de l'ACP.

III.2. Rappels sur les Matrices

III.2.1. Terminologie et opérations simples

Une matrice est un tableau rectangulaire de nombres. Une Matrice A d'ordre n*m est une matrice à n lignes et m colonnes. On la note $A = [a_{ij}]$ avec :

$i = 1$ ..... à n

$j = 1$ ..... à m

Les nombres a_{ij} appelés éléments de la matrice sont marqués par deux indices dont le premier (i) indique la ligne d'appartenance et le second (j) indique la colonne d'appartenance.

Exemple 1 : $n = 2$; $m = 3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

Exemple 2 : $n = 2$, $m = 4$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix}$$

Une matrice d'ordre « n*1 » est un vecteur colonne ; une matrice d'ordre « 1*m » est un vecteur ligne et une matrice d'ordre « 1*1 » est un scalaire.

Opérations simples

Addition de deux matrices : s'applique uniquement pour deux matrices de même ordre. L'addition se fait en additionnant les éléments de mêmes indices.

Soit $A = [a_{ij}]$ et $B = [b_{ij}]$ alors $A + B = [a_{ij} + b_{ij}]$

Soustraction de deux matrices : mêmes principes que l'addition.

Multiplication par un scalaire : $\lambda * A[a_{ij}] = A[a_{ij}] * \lambda = [\lambda a_{ij}]$

Multiplication de deux matrices.

Elle n'est possible que si le nombre de colonnes de la 1ère matrice est égale au nombre de lignes de la 2ème matrice.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} x & y \\ z & t \end{pmatrix} = \begin{pmatrix} ax + bz & ay + bt \\ cx + dz & cy + dt \end{pmatrix}$$

III.2.2. Déterminant et trace de la matrice :

Matrice carrée d'ordre 2 $\quad \boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$

Trace $= a_{11} + a_{22}$

Déterminant

Signes $\begin{vmatrix} + & - \\ - & + \end{vmatrix}$

Det A $= a_{11} * a_{22} - a_{21} * a_{12}$

Matrice carrée d'ordre 3 $\quad A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$

Trace $= a_{11} + a_{22} + a_{33}$

Déterminant :

Signes $\begin{vmatrix} + & - & + \\ - & + & - \\ + & - & + \end{vmatrix}$

Det A $= \sum(a_{(ij)} * Min_{(ij)})$ tenant compte des signes.

Le Mineur $(Min_{(ij)})$ étant le déterminant de la matrice d'ordre n-1 obtenue en éliminant la ligne « i » et la colonne « j ».

Il faut donc procéder en réduisant la matrice en une série de déterminant (2*2).
Prenons comme vecteur la ligne 1. On commence par la détermination des mineurs ($Min_{(ij)}$).

$$\mathrm{Min}_{11} = \mathrm{Det}\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} ; \mathrm{Min}_{12} = \mathrm{Det}\begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} ; \mathrm{Min}_{13} = \mathrm{Det}\begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{323} \end{pmatrix}$$

$$\mathbf{Det\ (A) = a_{11}*Min_{11} - a_{12}*Min_{12} + a_{13}Min_{13}}$$

Matrice carrée d'ordre supérieur à 3
Le nombre d'opérations est de plus en plus élevé. Il devient essentiel pour réduire ce nombre à choisir la ligne ou la colonne qui contient le plus grand nombre de 0 et/ou de 1pour faciliter les calculs.

Exemple Matrice d'ordre 4 $$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

Trace : $a_{11} + a_{22} + a_{33} + a_{44}$
Déterminant
Signes $\begin{vmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{vmatrix}$
Prenons là aussi comme vecteur la ligne 1.

$$\mathbf{Det(A) = a_{11}*Min_{11} - a_{12}*Min_{12} + a_{13}Min_{13} - a_{14}*Min_{14}}$$

Le calcul des déterminants des matrices d'ordre supérieur à 3 nécessitent plusieurs opérations. A titre d'exemple, le calcul du déterminant d'une matrice d'ordre 10 nécessite 32.10^6 opérations. Heureusement, les ordinateurs sont là.

III.2.3. Matrices particulières

- **Matrice carrée** : c'est une matrice dont le nombre de colonnes est égale au nombre de lignes (exp : matrice des variances-covariances et matrice des corrélations).
- **Matrice transposée** : la transposition d'une matrice est obtenue en inversant la séquence des indices de tel sorte que les lignes deviennent les colonnes et vis-versa.

Exemple

$$\mathrm{A} = \begin{pmatrix} 1 & 2 & 0 \\ 4 & 3 & -1 \end{pmatrix} \qquad {}^{t}A = \begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 0 & -1 \end{pmatrix}$$

La matrice ^{t}A est dite la transposée de A.

- **Matrice symétrique** : c'est une matrice égale à sa transposée (A = ^{t}A).
- **Matrice diagonale** : c'est une matrice dont les seuls éléments non nuls se trouvent sur la diagonale.
- **Matrice Identité** (généralement notée *In*) : c'est une matrice diagonale dont tous les éléments diagonaux valent 1. Cette matrice multiplie une matrice de même ordres sans la modifier.

III.2.4. Valeurs propres et Vecteurs propres

Valeurs propres

On appelle valeur propre de la matrice A une valeur non triviale (différente de 0) de λ pour laquelle le système $A\vec{x} = \lambda I\vec{x}$ donc $(A - \lambda I)\vec{x} = \vec{0}$ où *I* est la matrice identité de même ordre que A.

Pour calculer les valeurs propres d'une matrice (A) il faut donc résoudre l'équation :

$$\det(A - \lambda I) = 0$$

Une matrice d'ordre « n » possède au maximum « n » valeurs propres. Si la matrice est symétrique, toutes les valeurs propres sont réelles.

La somme des valeurs propres d'une matrice est égale à sa trace et leur produit est égal au déterminant.

$$\sum \lambda_i = \mathrm{tr}(A) \qquad \text{et} \qquad \prod \lambda_i = \det(A)$$

Exemple 1 : matrice carré d'ordre 2 $\quad A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$

$\det(A - \lambda I) = 0 \quad \Rightarrow \quad det\begin{pmatrix} 1-\lambda & 2 \\ 3 & 2-\lambda \end{pmatrix} = 0$

$(1-\lambda)*(2-\lambda) - 6 = 0$

$\lambda^2 - 3\lambda - 4 = 0$ (équation de deuxième degré à $\Delta = 25$)

$$\lambda_1 = \frac{(-b+\sqrt{\Delta})}{2a} = 4 \qquad \text{et} \qquad \lambda_2 = \frac{(-b-\sqrt{\Delta})}{2a} = -1$$

Les valeurs propres de la matrice $A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ sont $\lambda_1 = 4$ et $\lambda_2 = -1$

Exemple 2 : matrice carré d'ordre 3 $\quad A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}$

$$\det(A - \lambda I) = det\begin{pmatrix} 2-\lambda & 1 & 0 \\ 1 & 1-\lambda & 1 \\ 0 & 1 & 2-\lambda \end{pmatrix}$$

Prenons la ligne 1 comme vecteur.

$\mathrm{Det}\,(A - \lambda I) = a_{11}\mathrm{Min}_{11} - a_{12}\mathrm{Min}_{12} + a_{13}\mathrm{Min}_{13}$

$\mathrm{Det}\,(A - \lambda I) = (2-\lambda)[(1-\lambda)(2-\lambda) - 1)] - 1[1(2-\lambda) - 0] + 0$

$\det(A - \lambda I) = -\lambda^3 + 5\lambda^2 - 6\lambda$ (équation de troisième degré)

Dans ce cas, la réduction de l'équation est évidente

$-\lambda^3 + 5\lambda^2 - 6\lambda = -\lambda(\lambda^2 - 5\lambda + 6)$

$\det(A - \lambda I) = 0 \qquad \Rightarrow \qquad -\lambda(\lambda^2 - 5\lambda + 6) = 0$

Les valeurs propres de cette matrice sont donc : $\lambda_1 = 3$; $\lambda_2 = 2$ et $\lambda_3 = 0$

Vecteurs propres

Soit la matrice $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ de valeurs propres λ_1 et λ_2

Le vecteur propre $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ associé à chaque valeur propre λ est calculé en supposant que :

$$(A - \lambda I)\vec{y} = \vec{0}$$

donc $\begin{pmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{pmatrix} * \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\begin{pmatrix} (a_{11} - \lambda)y_1 & (a_{12})y_2 \\ (a_{21})y_1 & (a_{22} - \lambda)y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Les équations sont :

$(a_{11} - \lambda)y_1 + a_{12}y_2 = 0$ pour la première ligne

$a_{21}(y_1) + (a_{22} - \lambda)y_2 = 0$ pour la deuxième ligne

Exemple : matrice $A = \begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$ dont les valeurs propres sont $\lambda_1 = -1$ et $\lambda_2 = 2$

- Calcul du vecteur propre pour $\lambda_1 = -1$

$$\det(A - \lambda I)y = 0$$

$$\begin{pmatrix} (a_{11} - \lambda)y_1 & (a_{12})y_2 \\ (a_{21})y_1 & (a_{22} - \lambda)y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} (5 - (-1))y_1 & -3y_2 \\ 6y_1 & (-4 - (-1))y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

L'équation est, ici, la même pour les deux lignes : $6y_1 - 3y_2 = 0$ donc $y_2 = 2y_1$

Donc la base du vecteur propre associé à λ_1 est $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. En ACP, les logiciels utilisent les vecteurs normés.

- Calcul du vecteur propre pour $\lambda_2 = 2$

$$\det(A - \lambda I)y = 0$$

$$\begin{pmatrix} (a_{11} - \lambda)y_1 & (a_{12})y_2 \\ (a_{21})y_1 & (a_{22} - \lambda)y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} (5 - (2))y_1 & -3y_2 \\ 6y_1 & (-4 - 2)y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Les équations sont

$3y_1 - 3y_2 = 0$ pour la première ligne

$6y_1 - 6y_2 = 0$ pour la deuxième ligne

D'où $y_2 = y_1$

Donc la base du vecteur propre associé à λ_2 est $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Les bases des vecteurs propres de la matrice $\begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$ sont $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

III.3. Analyse en Composantes Principales (ACP)

III.3.1. Principe

L'ACP se base sur les acquis des analyses unidimensionnelle et bidimensionnelle. De l'analyse unidimensionnelle, elle retient, pour chaque variable, la moyenne et l'écart-type. De l'analyse bidimensionnelle elle retient la matrice des variances-covariances (si toutes les variables sont exprimées dans la même unité) et la matrice des corrélations (indépendamment des unités).

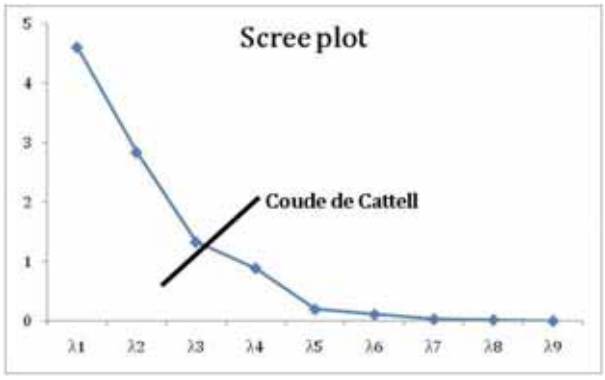
Soit « n » individus décrits par « p » variables quantitatives. L'ACP permet de faire le bilan des liaisons (corrélations) entre les « p » variables et des ressemblances (proximités) entre les « n » individus. Le but de l'ACP est donc d'expliquer le plus de variance possible avec un nombre de composantes le plus restreint possible. C'est-à-dire qu'on cherche des axes portant le maximum d'inertie. Autrement dit, on cherche à définir « k », nouvelles variables combinaisons linéaires de « p » variables initiales qui feront perdre le moins d'information possible. Ces variables seront appelées « composantes principales = facteurs principaux » et les axes qu'elles déterminent sont les « axes principaux = axes factoriels ».

L'élément clé de l'ACP est la détermination des valeurs propres et des vecteurs propres de la matrice d'inertie. Chaque valeur propre détermine un facteur (axe factoriel). Un axe factoriel F_k est donc une variable artificielle, combinaison linéaire de « p » variables initiales. Plus la valeur propre est élevée, plus le facteur explique une proportion significative de la variance totale.

Après vérification ($\sum\lambda$ = trace de la matrice), les valeurs propres doivent être classées dans le sens décroissant. Le % de chaque valeur propre par rapport à la trace de la matrice $\left(\frac{\lambda_i}{\sum\lambda}\right)$ traduit le taux de l'information restituée par elle, ou, la part d'inertie expliquée par l'axe qui lui est associé. Le taux $\frac{\lambda_1+\lambda_2}{\sum\lambda}$ traduit la part d'inertie expliquée par le premier plan principal, ou, le taux de l'information restituée par les deux plus grandes valeurs propres. De manière générale, les % cumulés rendent compte du taux de l'information restituée par les axes retenus. L'une des questions qui se posent donc est de savoir combien d'axes sont intéressants pour l'analyse. Deux types de critères peuvent être utilisés.

Un critère « absolu » (critère de Kaiser) : ne retenir que les axes dont les valeurs propres sont supérieures à l'inertie moyenne **(IM = $\sum\lambda_i$/nombre de variables).** Dans le cas de la matrice des corrélations (IM = 1) on retient donc les axes dont les valeurs propres sont supérieures à 1.

Un critère « relatif » (coude de Cattell) : retenir les axes dont les valeurs propres dominent les autres c'est-à-dire retenir les axes associés aux valeurs propres situées avant la cassure et ignorer les axes associés aux valeurs propres non significatives, situées après la cassure (éboulis factoriels).



Dans l'exemple ci-contre (exercice II.1. de la série TD2) on retiendra les trois premiers axes, les autres étant situés après la cassure et leurs valeurs propres sont inférieures à l'Inertie moyenne (1).

Le meilleur plan de projection est dirigé par les deux premiers axes (axes associés aux deux plus grandes valeurs propres de la matrice d'inertie). Les cordonnés sur l'axe 1 sont calculés sur la base du Vecteur propre associé à λ_1 et les cordonnés sur l'axe 2 sont calculés sur la base du Vecteur propre associé à λ_2. Toutefois, les représentations graphiques des variables et des individus ne se font pas dans les mêmes repères.

Projection des variables

La carte des variables est le cercle des corrélations (ou cercle corrélatoire). C'est un cercle de centre 0 et de rayon 1. Le diamètre (= 2) varie de -1 à 1. La carte des variables s'interprète en termes de liaisons (corrélations entre les variables). La matrice d'inertie est la matrice des corrélations. Les Cordonnés factoriels des variables sont calculés par la formule :

$$F_{ki} = \sqrt{\lambda_k} * V_{ki}$$

Deux variables dont les projections sont très proches sur le cercle corrélatoire sont fortement corrélées positivement. Deux variables dont les projections sont diamétralement opposées sont fortement corrélées négativement. Tant que les points représentant deux variables sont proches, tant que ces deux variables suivent la même évolution.

Projection des individus.

La carte des individus (plan factoriel) s'interprète en termes de proximités (ressemblances entre les individus). Si les variables sont de même nature (exprimées dans la même unité), la projection des individus peut se baser sur la matrice des variances-covariances en utilisant les **données seulement centrées** $(\boldsymbol{Z_i = x_i - \mu})$. On parle alors d'ACP centrée.

Si la matrice d'inertie utilisée est la matrice des corrélations, il faut se baser sur les données **normées**, c'est-à-dire **centrées et réduites** $(\boldsymbol{Z_i = \frac{x_i - \mu}{\sigma}})$. On parle alors d'ACP normée.

Soit Z_i les données centrées (ou centrées et réduites) et V_{ki} les valeurs du vecteur propre associé à la valeur propre λ_k de la matrice utilisée. La position d'un individu (i) sur l'axe K est : $\sum \mathbf{Z_i V_{ki}}$

Des individus dont les points sont très proches sur le plan factoriel sont assez similaires, en tenant compte de toutes les variables. Des individus dont les points sur le plan factoriel sont diamétralement opposés sont les plus différents en tenant compte de toutes les variables.

En Pratique, les étapes de l'ACP peuvent se résumer comme suit :

- Résultats sur les variables
  - Matrice des corrélations
  - Valeurs propres (*Eigenvalues*)
  - Vecteurs propres (*Eigenvectors*)
  - Calcul des coordonnées factorielles ($F_{ki} = \sqrt{\lambda_k} * V_{ki}$)
  - Projection des variables (cercle des corrélations)
- Résultats sur les individus
  - Matrice d'inertie (variances-covariances ou corrélations)
  - Valeurs propres (*Eigenvalues*)
  - Vecteurs propres (*Eigenvectors*)
  - Données centrées ($\boldsymbol{Z_i = x_i -}$) ou centrées et réduites ($\boldsymbol{Z_i = \frac{x_i - \mu}{\sigma}}$)
  - Calcul des coordonnées factorielles ($\sum \mathbf{Z_i V_{ki}}$)
  - Projection des individus (plan factoriel)

Rappel. La matrice d'inertie pour la projection des variables est la matrice des corrélations. Pour la projection des individus :

- si toutes les variables sont de même nature, on peut utiliser les données seulement centrées ($\boldsymbol{Z_i = x_i} - \mu$) et la matrice des variances-covariances.
- si non, on utilise les données centrées et réduites ($Z_i = \frac{x_i - \mu}{\sigma}$) et la matrice des corrélations.

III.3.2. Exemple simple pour assimiler la méthode.

Des cas géologiques concrets seront étudiés en TD et TP. En cours, la démarche à suivre sera expliquée par un exemple simple. Les notes de 9 étudiants en 4 modules de STU-S5 (Hydrogéologie, Géophysique, Métallogénie et Géochimie).

	Données brutes				Données centrées				Données centrées et réduites			
	Hydro	Gphy	Métallo	Gchi	Hydro	Gphy	Métallo	Gchi	Hydro	Gphy	Métallo	Gchi
Amina	6	6	5	5,5								
Brahim	8	8	8	8								
Chafik	6	7	11	9,5								
Driss	14,5	14,5	15,5	15								
Mohamed	14	14	12	12,5								
Naima	11	10	5,5	7								
Rachida	5,5	7	14	11,5								
Saleh	13	12,5	8,5	9,5								
Zineb	9	9,5	12,5	12								
Moyenne												
Ecart-type												

Matrice des variances-covariances					Matrice des corrélations				
	Hydro	Gphy	Métallo	Gchi		Hydro	Gphy	Métallo	Gchi
Hydro					Hydro				
Gphy					Gphy				
Métallo					Métallo				
Gchi					Gchi				

Valeurs propres et taux d'information restituée

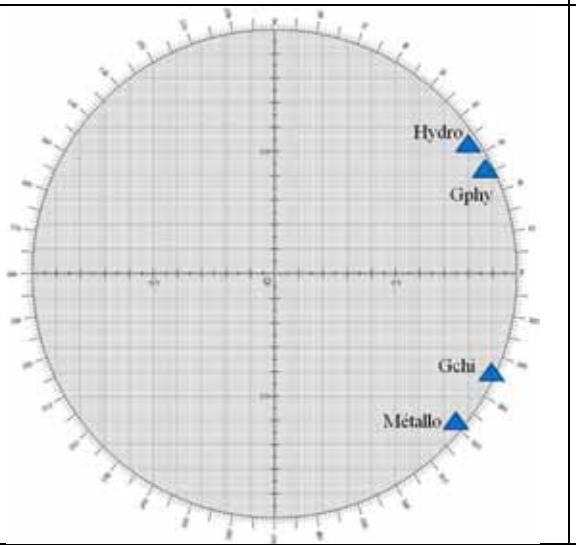
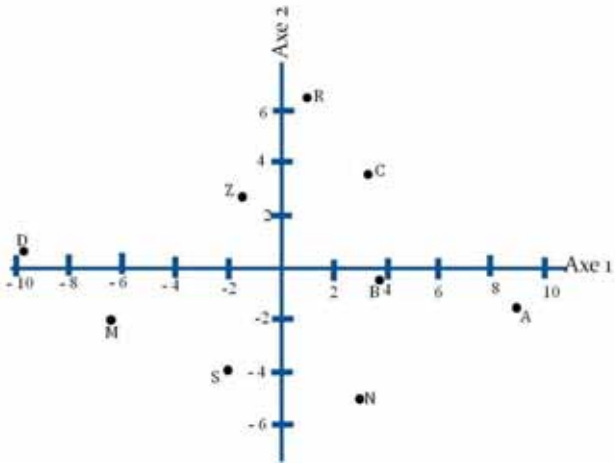
	Matrice des variances-covariances				Matrice des corrélations		
	λ	%λ	% cumulés		λ	%λ	% cumulés
λ_1							
λ_2							
λ_3							
λ_4							
Somme							

Valeurs propres et vecteurs propres

Valeurs propres (λ)	Matrice des variances-covariances					Matrice des corrélations			
	λ_1	λ_2	λ_3	λ_4		λ_1	λ_2	λ_3	λ_4
$\sqrt{\lambda}$									
Vecteurs propres									

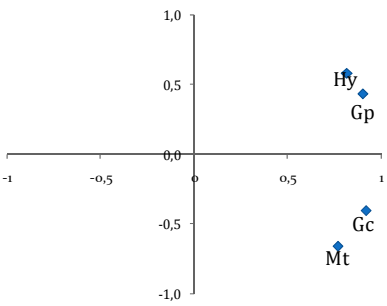
Cordonnés factoriels des variables

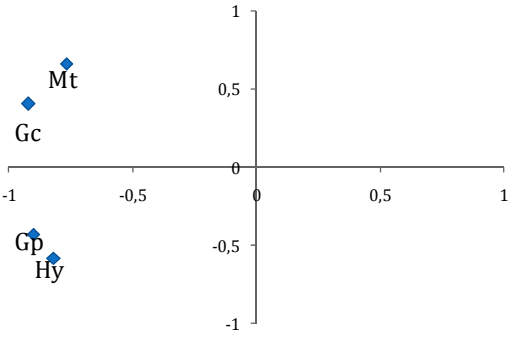
Variables	Vecteur associé à λ_1 (v_{1i})	Cordonnés sur l'axe 1 ($\sqrt{\lambda_1} * v_{1i}$)	Vecteur associé à λ_2 (v_{2i})	Cordonnés sur l'axe 2 ($\sqrt{\lambda_2} * v_{2i}$)
Hydro				
Gphy				
Métallo				
GChi				

Projection des variables sur le **cercle des corrélations**	Projection des individus sur le **plan factoriel**
	

On remarque que l'Hydrogéologie et la Géophysique d'une part ; la Géochimie et la Métallogénie d'autre part sont fortement corrélées (ce qui peut-être déduit également de la lecture de la matrice des corrélations). C.-à-d. que les étudiants qui sont bons en Hydro sont bons également en Gphy ; ceux qui sont mauvais en Hydro sont mauvais également en Gphy (même chose pour la Métallo et la Gchi).
Sur l'axe 1 du plan factoriel, les bons étudiants à gauche, les mauvais à droite (la moyenne la plus élevée (14,88) a été obtenue par Driss (D), la moyenne la plus basse (5,6) a été obtenue par Amina (A). Sur l'axe 2, en bas les étudiants mauvais en Métallo et Gchi ; en haut les étudiants mauvais en Hydro et Gphy.

Remarque : les signes des vecteurs propres sont fixés arbitrairement. Ils peuvent êtres inversés selon les calculateurs utilisés. Cependant, les positions relatives entre les variables (corrélations) et les individus (proximités) sont préservées.





Géostatistique

I. Généralités et Historique

Dans plusieurs phénomènes naturels, les corrélations diminuent et la variabilité augmente avec la distance. Autrement dit, deux observations situées l'une près de l'autre devraient, en moyenne, se ressembler davantage que deux observations éloignées. On parle alors de données spatialisées (régionalisées). La Géostatistique, appelée également Statistique Spatiale, est la branche de la Statistique qui cherche à déterminer la structure spatiale d'une variable. La structure spatiale, au sens large, d'une variable régionalisée est tout simplement la manière dont se comporte la variable dans son champ.

On classe habituellement les approches de la Statistique spatiale en trois grands domaines, qui correspondent à des types de données distincts.

- Lorsque les données sont échantillonnées irrégulièrement mais peuvent en principe être mesurées en tout point d'un domaine continu (exp. les teneurs en un métal dans un secteur minier, les teneurs en matière organique dans un champ agricole, la profondeur du toit d'une couche, le pH du sol etc.) on utilise le formalisme des champs aléatoires continus ; **c'est le domaine habituel de la Géostatistique** (seul ce cas sera traité).
- Lorsque, par leur nature même, les données sont liées à un réseau (pour des images ou des données récoltées sur des entités administratives, etc.), il est certes possible d'utiliser le formalisme de la Géostatistique, mais celui des champs de Markov est souvent plus adapté.
- Enfin, lorsque ce sont les coordonnées qui portent l'information principale (positions des arbres dans une forêt, points d'impact de la foudre dans une région, etc.), on parle de processus de points.

Durant les années 1950, l'ingénieur minier sud-africain D.G. Krige (1919-2013) avait développé une série de méthodes statistiques empiriques pour estimer la teneur d'un bloc de minerai à partir d'échantillons pris autour du bloc à exploiter. Le français Georges Matheron (1930-2000) a formalisé cette méthode qu'il a appelée, en hommage à son initiateur, le krigeage (et non pas krigéage). Signalons que Matheron fut également le premier à utiliser le terme Géostatistique (1962), correspondant à un rapprochement entre deux domaines :

- problèmes techniques du monde minier ;
- méthodes mathématiques de régression.

Durant la première période (1950-1960), l'utilisation des méthodes géostatistiques portaient essentiellement sur les estimations minières. Jusqu'aux années 1970-1980 les méthodes géostatistiques n'étaient connues que de quelques chercheurs. Depuis longtemps, le champ de la Géostatistique ne se limite plus à la Géologie et la Géographie. Actuellement, les méthodes de la Géostatistique sont appliquées dans des domaines très divers (démographie, Economie, Météorologie, pollutions, épidémiologie, trafique routier, Pédologie, etc.) à tel point que certains chercheurs contestent la conservation du terme « Géostatistique ». A titre d'exemple, Matheron lui-même avait proposé, en 1978, l'utilisation du terme « Modèles topo-probabilistes ».

Lorsqu'on mesure une caractéristique en un point, on peut considérer la valeur obtenue comme la réalisation d'une variable aléatoire en ce point. Il en est de même pour tous les points d'un site donné. On a donc un grand nombre (ou une infinité) de v.a. représentant conjointement un site. Les valeurs d'une variable dans un champ ne sont connues que dans quelques points (ou lignes). Le géostatisticien cherche à estimer les valeurs de cette variable en dehors des points (ou lignes) données. Si les estimations sont calculées pour des zones situées à l'intérieur du champ d'étude, il s'agit d'une interpolation. Si elles (estimations) sont calculées pour des zones en dehors du champ d'étude, il s'agit d'une extrapolation.

Il y a deux étapes principales dans une étude géostatistique :

- identification des caractéristiques des v.a., l'outil principal utilisé est le variogramme ;
- utilisation de ces caractéristiques et des valeurs connues pour l'estimation optimale aux points (ou lignes ou volumes) non mesurés ; la méthode utilisée est le krigeage.

Ces deux étapes n'ont rien de figé et il est naturel en cours de travail de revenir en arrière et d'affiner ou corriger un modèle à la lumière des premiers résultats obtenus.

Dans ce cours, qui n'est qu'un aperçu introductif à la Géostatistique, l'accent sera mis sur les principes et les problèmes méthodologiques plus que sur les techniques mathématiques. Il sera question de la notion de variographie puis de quelques notions sur le krigeage linéaire univarié.

II. Variographie

II.1. Eléments de vocabulaire

Le variogramme est la mesure de la demi-variance d'une variable régionalisée. Il est indépendant de la position géographique et dépend uniquement de la distance entre les points. Soit z_i et z_j les valeurs d'une variable régionalisée en deux points.

La dissemblance entre les deux points est : $\frac{(z_i - z_j)^2}{2}$.

La moyenne des dissemblances est donc $\frac{\sum(z_i - z_j)^2}{2n} = \frac{1}{2n}\sum(z_i - z_j)^2$

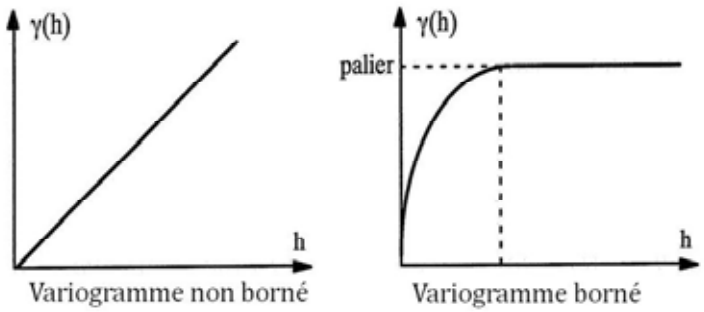
Un point du variogramme expérimental est la moyenne des dissemblances entre les valeurs pour toutes les paires reliées par un vecteur distance « h ». Sa formule est :

$$\gamma_{(h)} = \frac{1}{2nh}\sum(z_i - z_{i+h})^2$$

avec « nh » le nombre de paires pour le vecteur distance « h ».

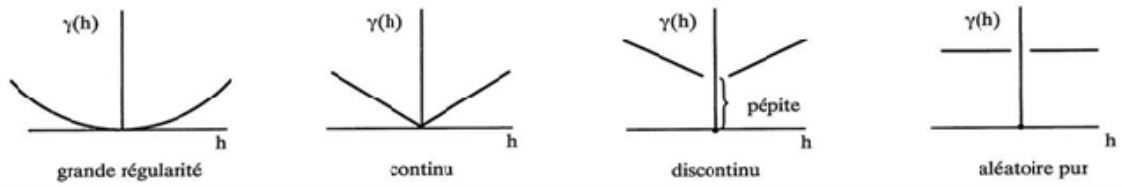
Lorsque la dissemblance moyenne des valeurs est constante pour toutes les distances h, il y a une absence complète de structuration spatiale des valeurs.

Le variogramme peut être borné (avec palier) ou non borné (sans palier), avec ou sans effet de pépite.



Variogramme non borné Variogramme borné

Effet de pépite (*nugget* en anglais) : variation à très courte échelle (erreurs de localisation, erreurs d'analyse et précision analytique) due à un saut de discontinuité à l'origine du variogramme souvent notée **(C_0)**.



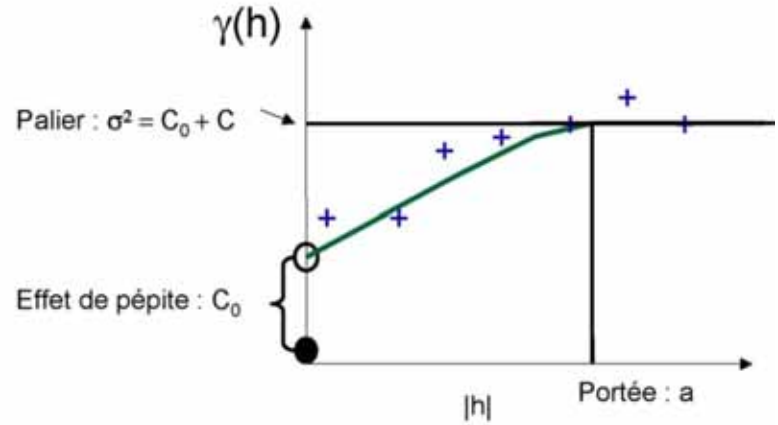
Exemples de comportement de variogramme à l'origine

Le variogramme est une fonction paire ($\gamma_{(h)} = \gamma_{(-h)}$) nulle en h = 0 et positive partout ailleurs.

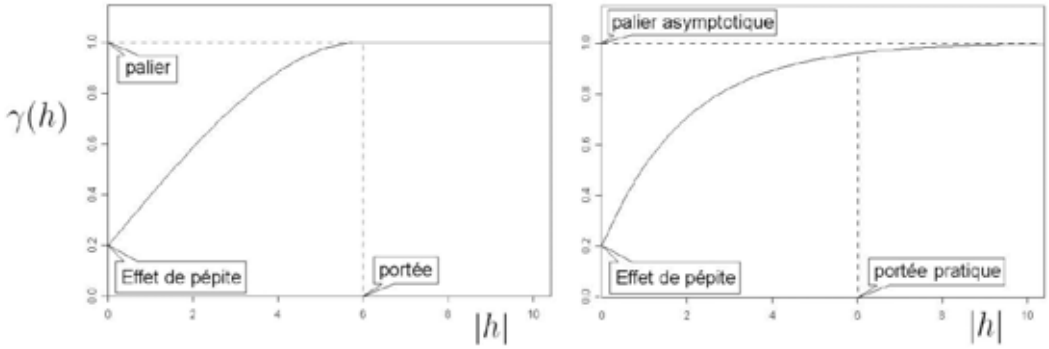
Palier (*sill* en anglais) : il correspond à la variance du champ d'étude et est définie par l'équation

$$(\sigma^2 = C_0 + C)$$

- C_0 est l'effet de pépite ;
- C est la variance expérimentale des données qui peut s'exprimer comme la demi-moyenne des carrés des écarts de tous les couples qui peuvent être formés à partir de ces données.



Portée (*range* en anglais) : distance (a) à laquelle le variogramme atteint son pallier, soit de façon exacte (portée vraie), soit asymptotiquement (portée pratique). La portée pratique est définie par la distance à laquelle le semi-variogramme atteint 95% de la valeur de son palier. Pour les distances h < a, les échantillons sont corrélés ; pour les distances h ≥ a ils ne le sont pas.



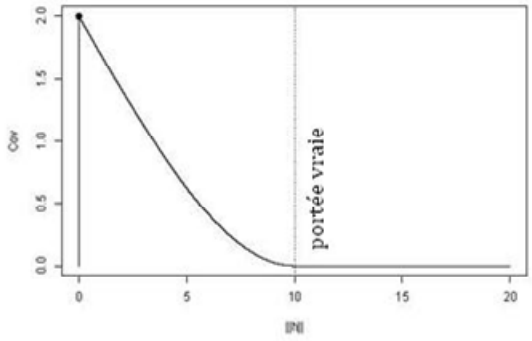
Rappel : une **droite asymptote** à une courbe est une droite telle que, lorsque l'abscisse ou l'ordonnée tend vers l'infini, la distance de la courbe à la droite tend vers 0.
Remarque : la portée est propriété du champ, non de la variable régionalisée elle-même.

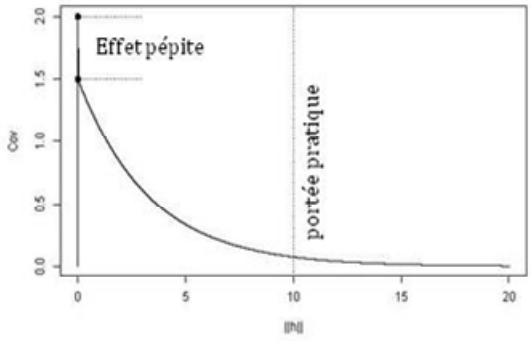
II.2. Covariogramme et variogramme

Le covariogramme $C_{(h)}$ est la fonction covariance de variable régionalisée.

$$(C_{(h)} = Cov(z_i, z_{i+h})$$

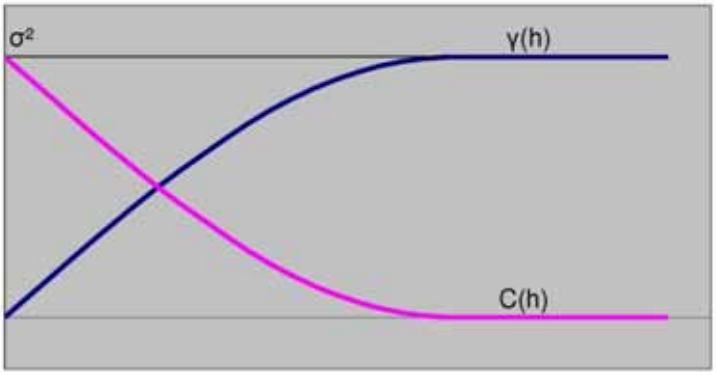
C'est également une fonction paire ($C_{(h)} = C_{(-h)}$). Là aussi, existent des modèles avec, et des modèles sans, effet de pépite. La fonction de la covariance est bornée et n'excède pas la variance. Elle s'éteint lorsque la portée vraie est atteinte. La portée pratique est définie selon le même principe que celle du variogramme (95%).





L'existence du covariogramme implique l'existence du variogramme. L'implication inverse n'est vraie que si le variogramme est borné. Il se peut donc que le covariogramme d'une fonction ne soit pas défini. Par conséquent, le variogramme, d'ailleurs plus facile à estimer, est le plus utilisé.

Le variogramme est alors simplement la variance totale (σ^2) moins la covariance ($C_{(h)}$) en fonction de la distance (h) entre les points.



$$C_{(h)} = \sigma^2 - \gamma_{(h)}$$

$$\gamma_{(h)} = \sigma^2 - C_{(h)}$$

II.3. Variogramme expérimental

Il mesure la variabilité à différentes échelles d'une variable régionalisée. Normalement, il faut au moins une trentaine de paires pour le calcul des points du variogramme. Évidemment, une certaine régularité (homogénéité) du phénomène est requise dans le champ. Des zones très différentes géologiquement doivent être traitées séparément.

Variogramme directionnel. On parle de variogramme directionnel lorsqu'on calcule un variogramme pour une distance h ayant une direction donnée. Dans ce cas, le minimum de points pour que l'analyse variographique soit statistiquement admissible est 60 points. Toutefois, pour comprendre le principe, prenons un exemple avec un nombre réduit de paires.

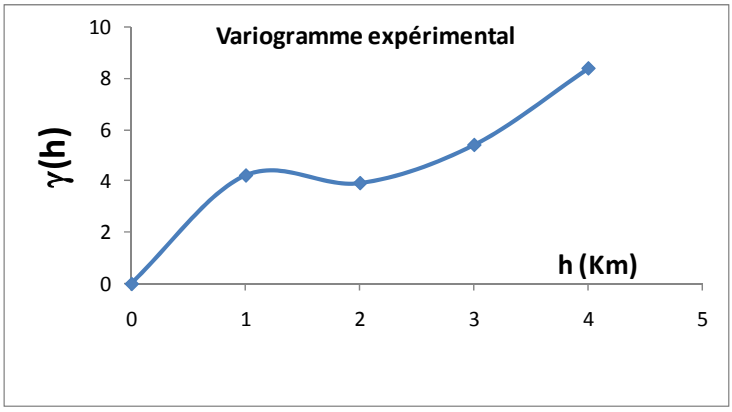
Exemple simple : soit, pour une variable régionalisée, les valeurs (z_1 à z_8) mesurées en huit points alignés (x_1 à x_8) distants, l'un de l'autre d'une distance de 1 km.

Le tableau ci-dessous résume les différentes étapes pour calculer les différents points (γ_h) du variogramme expérimental.

	x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	Σ	nh	$2\gamma_{(h)}$	$\gamma_{(h)}$
h	z_i	**3**	**6**	**5**	**7**	**2**	**2**	**4**	**0**				
1 km	$(z_i - z_{i+1})^2$	9	1	4	25	0	4	16		59	7	8,43	4,21
2 km	$(z_i - z_{i+2})^2$	4	1	9	25	4	4			47	6	7,83	3,92
3 km	$(z_i - z_{i+3})^2$	16	16	9	9	4				54	5	10,80	5,40
4 km	$(z_i - z_{i+4})^2$	1	16	1	49					67	4	16,75	8,38
5 km	$(z_i - z_{i+5})^2$	1	4	25						30	3	10,00	5,00
6 km	$(z_i - z_{i+6})^2$	1	36							37	2	18,50	9,25
7 km	$(z_i - z_{i+7})^2$	9								9	1	9,00	4,50

Des distances entre les échantillons plus grandes que la moitié de la région étudiée ne sont pas utilisées, car le nombre de paires serait faible. Pour la même raison (nombre de paires faible) le traçage ne va pas, en général, au-delà de la moitié de la distance maximale entre deux points. Les points du semi-variogramme expérimental de l'exemple sont donc :

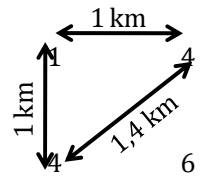
h (km)	$\gamma_{(h)}$
0	0
1	4,21
2	3,92
3	5,40
4	8,38



Variogramme omnidirectionnel. Il est toujours possible, sur un jeu de données distribuées dans l'espace, de calculer un variogramme expérimentale. Il suffit, pour un vecteur distance « h » donné, de considérer tous les couples de valeurs distants de « h ».

On parle de variogramme omnidirectionnel si les points ne sont pas tous alignés. Le minimum de points pour que l'analyse variographique soit statistiquement admissible est 30 points, mais il est recommandé d'avoir une centaine de points.

Exemple simple : les valeurs d'une variable régionalisée dans un secteur de 16 km^2 échantillonné selon une grille régulière à nœuds distants de 1 km sont :



1	4	5	4	0
4	6	6	6	3
5	6	8	7	5
2	5	5	4	1
0	1	2	1	0

Pour cet exemple, plusieurs variogrammes expérimentaux peuvent être tracés. En se limitant aux quatre exemples les plus simples (Est-Ouest ; Nord-Sud ; SW-NE et NW-SE), calculez les points des variogrammes expérimentaux (TD).

II.4. Variogramme théorique (modèle)

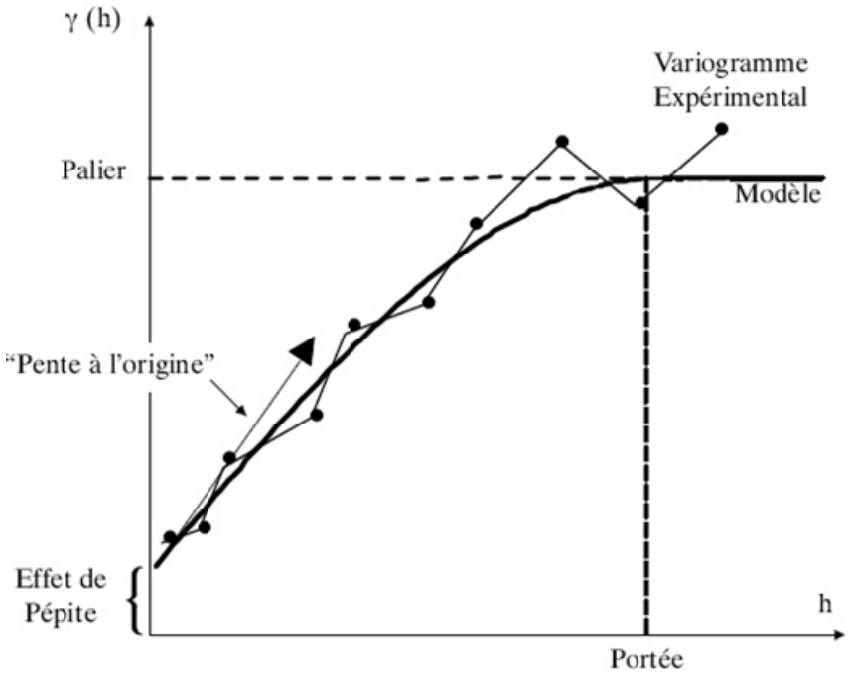
Le variogramme expérimental est basé sur un nombre fini de distances seulement. Le vrai variogramme est celui qui serait calculé à partir d'une connaissance exhaustive de la variable sur son champ. En pratique, un tel variogramme est inconnu.

Le géostatisticien dégage les caractéristiques importantes du variogramme expérimental et peut alors lui ajuster une fonction mathématique (le variogramme théorique ou modèle de variogramme), qui peut être considéré comme une estimation du vrai variogramme. On dit que la substitution d'une fonction mathématique à la structure empirique permet d'estimer les fonctions théoriques C et γ sous-jacentes aux fonctions empiriques C* et γ*. C'est ce qu'on appelle l'ajustement ou la modélisation.

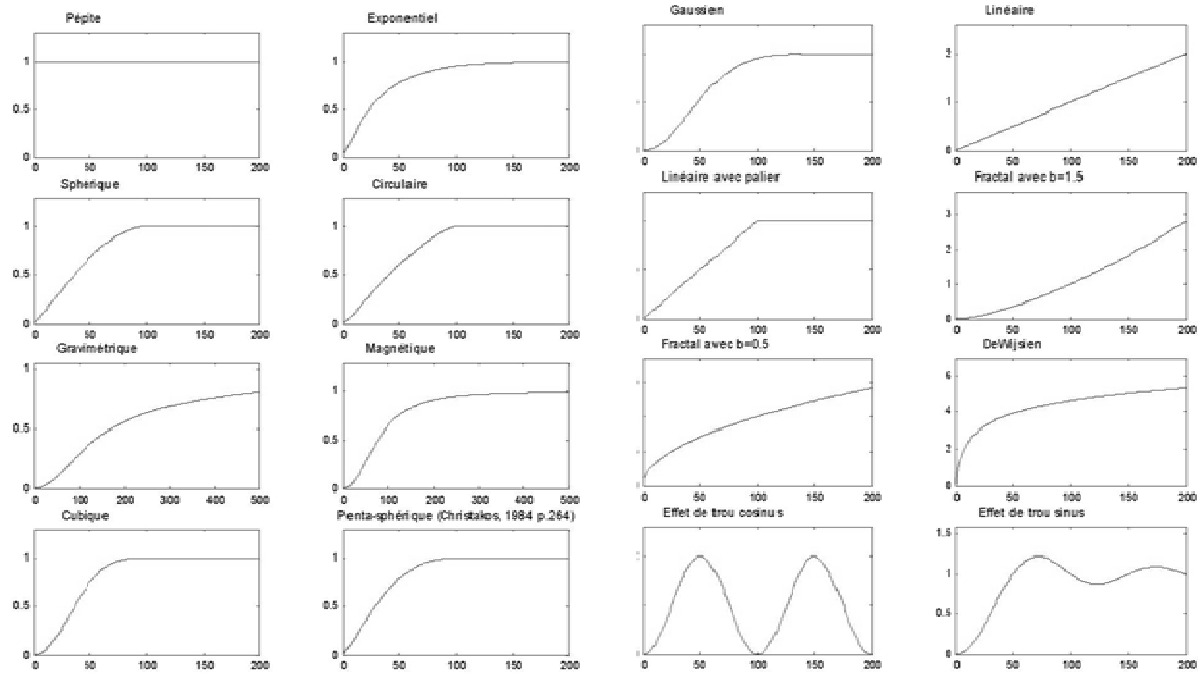
L'intérêt de la modélisation d'un variogramme expérimental est essentiellement pratique : il s'agit de passer d'une fonction définie par des points à une fonction continue dans l'espace et possédant une expression mathématique.

L'étape de l'analyse variographique qui influence le plus le résultat final de l'interpolation est la sélection du modèle variographique. Ce choix doit idéalement tenir compte des analyses variographiques antérieures effectuées sur des variables régionalisées modélisant le même phénomène naturel, de l'analyse exploratoire et de l'allure du semi-variogramme expérimental. Par exemple, une apparente discontinuité à l'origine du semi-variogramme expérimental suggère l'ajout d'un effet de pépite dans le modèle.

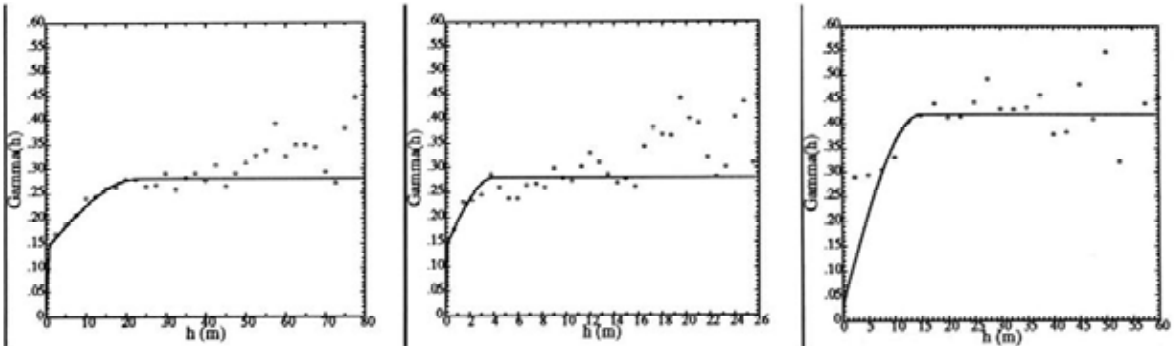
L'ajustement peut se faire à l'œil, mais il s'effectue habituellement à l'aide de méthodes d'estimation. Si l'utilisateur désire automatiser l'ajustement, il doit sélectionner une procédure d'estimation. Les différentes procédures devraient mener à des estimations assez semblables. Il est bon de toujours vérifier après coups le bon sens des paramètres estimées obtenus. Notamment, les estimations d'un effet de pépite, d'une portée ou d'un palier doivent toujours être positives.



Il existe un catalogue de modèles. La figure ci-dessous montre quelques exemples de ce catalogue (d'autres modèles peuvent êtres facilement consultés sur internet). Toute somme de modèles élémentaires est aussi un modèle admissible

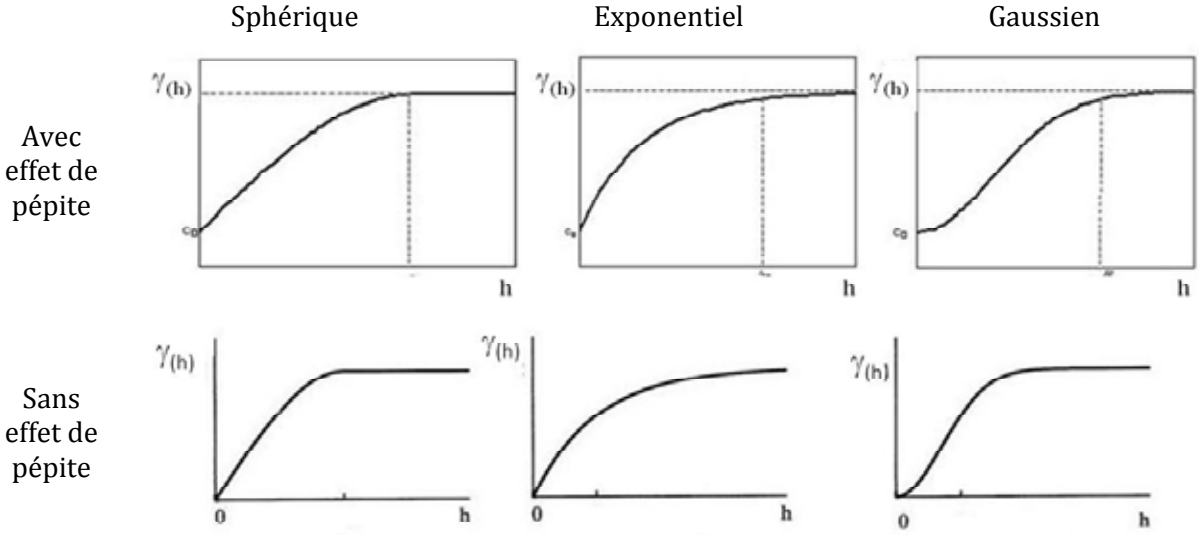


Le choix et l'ajustement d'une fonction au semi-variogramme est la partie la plus délicate du Krigeage. On peut dire que les modèles (variogrammes théoriques) sont des expressions analytiques que l'on tente d'ajuster le mieux possible aux points des variogrammes expérimentaux. Ces modèles tentent d'ajuster au mieux la croissance du variogramme pour les courtes distances (excepté le premier point), mais ne tiennent pas compte des oscillations du variogramme expérimental pour les grandes distances.



Exemples d'ajustements

Les modèles les plus utilisés en Géologie sont les modèles Sphérique ; Exponentiel et Gaussien, avec ou sans effet de pépite.



Sphérique

- $h = 0 \Rightarrow \gamma_{(h)} = 0$
- $0 < h < a \Rightarrow \gamma_{(h)} = C\left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right]$
- $h \geq a \Rightarrow \gamma_{(h)} = C$

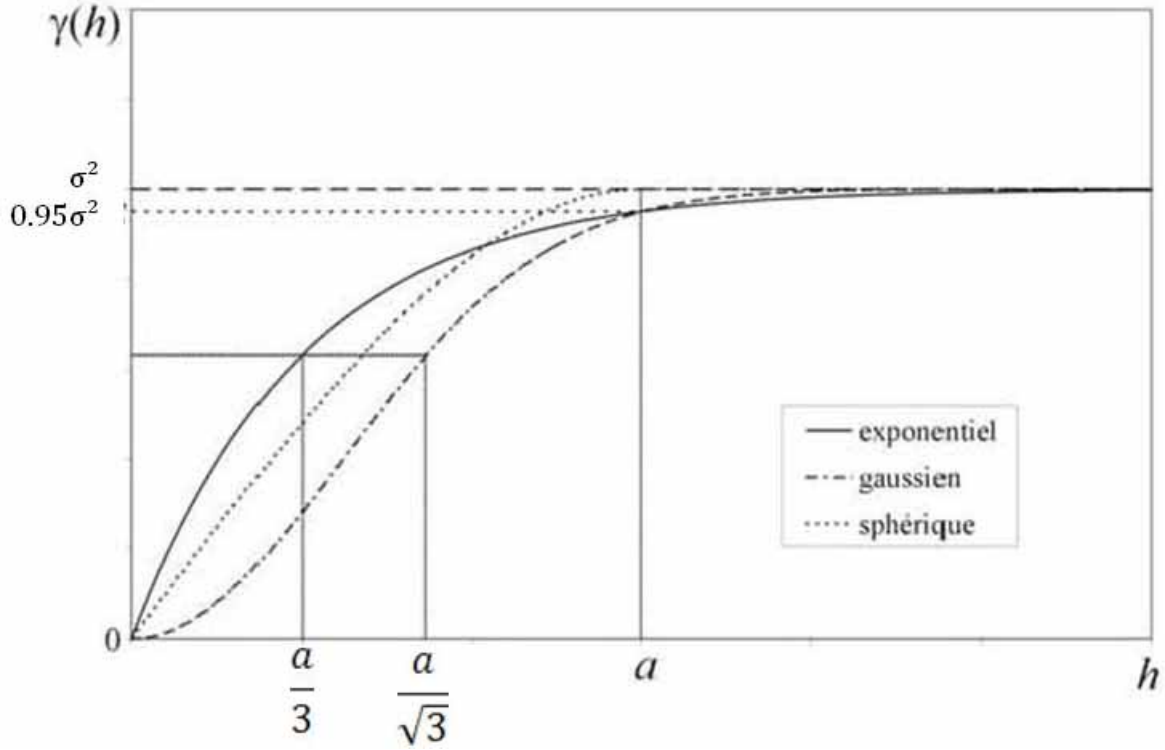
Exponentiel $\gamma_{(h)} = C\left[1 - e^{\frac{-3h}{a}}\right]$

Ces deux modèles (sphérique et exponentiel) sont les plus utilisés pour les teneurs des gisements miniers, les propriétés mécaniques des roches et les analyses géochimiques.

Gaussien $\gamma_{(h)} = C\left[1 - e^{-3\left(\frac{h}{a}\right)^2}\right]$

Ce modèle est utilisé pour les variables continues (topographie, champ gravimétrique, charge hydraulique). Son utilisation est déconseillée dans le cas de variables discrètes.

Le schéma ci-dessous donne un aperçu comparatif des trois modèles les plus utilisés en Géologie.



Cas particuliers

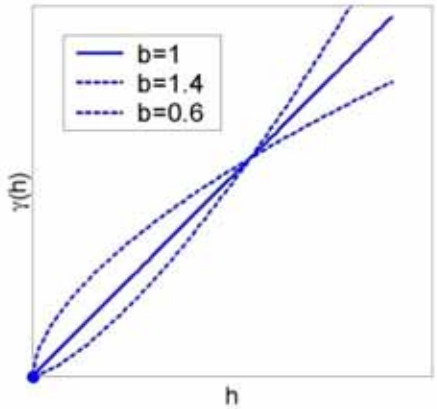
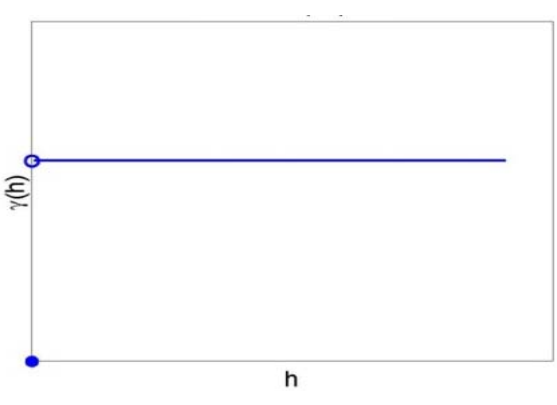
Puissance $\gamma_{(h)} = Ch^b$ avec $0 < b < 2$

Si $b = 1$, on parle de modèle linéaire

Effet pépite pur.

$$h = 0 \Rightarrow \gamma_{(h)} = 0$$
$$h \neq 0 \Rightarrow \gamma_{(h)} = C_0$$

C'est le cas extrême de l'effet de pépite. Il caractérise une absence totale de corrélation entre échantillons.

Puissance	Effet pépite pur
	

III. Krigeage
III.1. Définitions

Du point de vue mathématique, le Krigeage n'est autre qu'une application de la théorie des fonctions aléatoires à l'étude de phénomènes naturels fluctuants dans le temps et/ou l'espace. On dit que le krigeage est une régression multiple de variance minimum, à partir de données corrélées.

Exemples de définition :

- « *En terme miniers, le problème du krigeage consiste à trouver la meilleur estimation lineaire possible de la teneur d'un panneau, compte tenu de l'information disponible, c'est-à-dire des teneurs des différents échantillons qui ont été prélevés, soit à l'intérieur, soit à l'extérieur du panneau que l'on veut estimer* » **(Matheron 1970).**
- « *Le krigeage est une méthode stochastique d'interpolation spatiale qui prévoit la valeur d'un phénomène naturel en des sites non échantillonnés par une combinaison linéaire sans biais et à variance minimale des observations du phénomène en des sites voisins* » **(Baillargeon 2005).**
- « *Le krigeage est, en géostatistique, la méthode d'estimation linéaire garantissant le minimum de variance. Le krigeage réalise l'interpolation spatiale d'une variable régionalisée par calcul de l'espérance mathématique d'une variable aléatoire utilisant l'interprétation et la modélisation du variogramme expérimental. C'est le meilleur estimateur linéaire non-biaisé ; il se fonde sur une méthode objective. Il tient compte non seulement de la distance entre les données et le point d'estimation, mais également des distances entre les données deux-à-deux* » (http://fr.wikipedia.org/wiki/Krigeage consulté le 04 février 2015).

Donc, l'idée de base du krigeage est de prévoir la valeur de la variable régionalisée étudiée en un site non échantillonné par une combinaison linéaire de données ponctuelles adjacentes. On dit que le Krigeage sert à effectuer l'interpolation spatiale, c.-à-d. qu'il permet de prévoir la valeur prise par un phénomène naturel sur un site à partir d'observations ponctuelles de ce phénomène en des sites voisins.

Comparé aux autres méthodes d'interpolation, le Krigeage se distingue par sa prise en compte de la structure de dépendance spatiale des données. Ainsi, le krigeage est souvent considéré comme la méthode optimale (au sens statistique du terme) d'estimation spatiale.

Les fondements du krigeage se basent sur une connaissance de la structure de dépendance spatiale des données. Cependant, en pratique, cette structure n'est pas connue. Elle est estimée lors de l'analyse variographique.
Soit :

- Z_i les valeurs d'une variable régionalisée dans les points (x_i) ;
- (m) l'espérance mathématique de la fonction aléatoire ;
- (x_p) un point du plan non échantillonné.

L'estimation de la valeur de cette variable par krigeage pour le point (x_p) est :

$$\hat{Z}_p = \sum_i WiZi + (1 - \sum Wi)m$$

Les poids Wi (***Weight en anglais***) associés à chacune des valeurs régionalisées observées dépendent de leur localisation par rapport au point x_p. Le krigeage consiste donc à **déterminer la combinaison des poids Wi** qui garantit que les (x_p) se trouvent sur le variogramme.

On se limitera au Krigeage ponctuel univarié. Ils existent plusieurs techniques, trois seulement seront évoquées : Krigeage simple ; Krigeage ordinaire et Krigeage universel. Les deux premières sont stationnaires, la troisième est non stationnaire.

III.2. Modèles stationnaires

III.2.1. Krigeage simple (variable stationnaire de moyenne connue)

Puisque la variable est stationnaire, on peut poser m = 0.

L'équation $\hat{Z}_0 = \sum WiZi \ + \ (1 - \sum Wi)m$ se résume à $\hat{Z}_0 = \ \sum WiZi$.

Le Krigeage simple consiste à calculer les Wi (les poids) de l'équation à l'aide des valeurs de la fonction $C_{(h)}$ ou $\gamma_{(h)}$ correspondant aux n points choisis.

Le problème peut s'écrire sous forme de $[Aij] * [Wi] = \ [Bip]$ avec :

- $[Aij]$ une matrice carrée d'ordre n *n ;
- $[Wn]$ et $[Bip]$ des matrices d'ordre n*1 (vecteurs colonnes).

Par simplification, l'équation $[Aij] * [Wi] = \ [Bip]$ est souvent notée $A^*W = B$

$$A = \begin{pmatrix} C_{(h_{11})} & C_{(h_{12})} & \bullet & C_{(h_{1n})} \\ C_{(h_{21})} & C_{(h_{22})} & \bullet & C_{(h_{2n})} \\ \bullet & \bullet & \bullet & \bullet \\ C_{(h_{n1})} & C_{(h2n)} & \bullet & C_{(h_{nn})} \end{pmatrix} ; W = \begin{pmatrix} W_1 \\ W_2 \\ \bullet \\ W_n \end{pmatrix} \text{ et } B = \begin{pmatrix} C_{(h1p)} \\ C_{(h2p)} \\ \bullet \\ C_{(hnp)} \end{pmatrix}$$

ou

$$A = \begin{pmatrix} \gamma_{(h_{11})} & \gamma_{(h_{12})} & \bullet & \gamma_{(h_{1n})} \\ \gamma_{(h_{21})} & \gamma_{(h_{22})} & \bullet & \gamma_{(h_{2n})} \\ \bullet & \bullet & \bullet & \bullet \\ \gamma_{(h_{n1})} & \gamma_{(h_{n2})} & \bullet & \gamma_{(h_{nn})} \end{pmatrix} ; W = \begin{pmatrix} W_1 \\ W_2 \\ \bullet \\ W_n \end{pmatrix} \text{ et } B = \begin{pmatrix} \gamma_{(h1p)} \\ \gamma_{(h2p)} \\ \bullet \\ \gamma_{(hnp)} \end{pmatrix}$$

Les $C_{(hij)}$ sont les valeurs du covariogramme expérimental et les $\gamma_{(hij)}$ celles du variogramme expérimental qui correspondent à la distance entre les points x_i et x_j. Les $C_{(hip)}$ et $\gamma_{(hip)}$ sont calculés à l'aide de la fonction analytique qui a été ajustée aux points du covariogramme et variogramme théoriques.

La matrice des variances-covariances étant symétrique, définie positive, elle est inversible et on résout le système de krigeage en l'inversant : $A^*W = B \ \Rightarrow \ \ W = A^{-1}*B$

Comme nous l'avons évoqué auparavant, l'utilisation du variogramme est plus facile que celle du covariogramme. Dans ce qui suivra, on utilisera donc le variogramme. Il faut donc résoudre les équations :

$$W_1\gamma_{(h11)} + W_2\gamma_{(h12)} + \ldots\ldots + W_n\gamma_{(h1n)} = \gamma_{(h1p)}$$

$$W_1\gamma_{(h21)} + W_2\gamma_{(h22)} + \ldots\ldots + W_n\gamma_{(h2n)} = \gamma_{(h2p)}$$

.

.

$$W_1\gamma_{(hn1)} + W_2\gamma_{(hn2)} + \ldots\ldots + W_n\gamma_{(hnn)} = \gamma_{(hnp)}$$

C'est un système linéaire où le nombre d'équations et le nombre d'inconnues sont égaux aux nombres de points de données utilisées.

Une fois les poids (Wi) des différents points calculés, la valeur estimée pour le point x_p est :

$$\hat{Z}_p = W_1Z_1 + W_2Z_2 + W_3Z_3 \ldots\ldots + W_nZ_n$$

Ou

$$\hat{Z}_p = \sum WiZi$$

Remarque : il n'est pas possible d'effectuer un krigeage simple si le variogramme ne présente pas de palier.

III.2.2. Krigeage ordinaire (variable stationnaire de moyenne inconnue)

Dans la réalité, il est rare que l'on connaisse avec certitude la valeur de la moyenne. Il convient donc d'élargir le krigeage aux cas où celle-ci est inconnue. C'est ce qu'on appelle le Krigeage ordinaire, par ailleurs le plus fréquemment utilisé.
Dans ce cas, m étant inconnue, une condition dite condition d'universalité est imposée.

Condition d'universalité : $\sum Wi = 1$

Par conséquent $(\sum Wi - 1)m = 0$

Il s'en suit que, là aussi, l'équation :

$\hat{Z}_0 = \sum WiZi + (1 - \sum Wi)m$ se résume à $\hat{Z}_0 = \sum WiZi.$

Un degré supplémentaire est utilisé en ajoutant une variable libre (λ) appelée paramètre de Lagrange, dans le but de minimiser l'erreur d'estimation.

Dans ce qui suit, juste pour alléger l'écriture :

- $C_{(hij)}$ sera notée C_{ij} ;
- $\gamma_{(hij)}$ sera noté γ_{ij} ;
- $\gamma_{(hip)}$ sera notée γ_{ip}

Le système du krigeage ordinaire s'écrit alors :

$$\sum W_iC_{ij} + \lambda = C_{ip} \qquad \forall i$$

ou

$$\sum W_i\gamma_{ij} - \lambda = \gamma_{ip} \qquad \forall i$$

Paramètre de Lagrange (λ) : permet de trouver les points stationnaires (maximum, minimum) d'une fonction dérivable d'une ou plusieurs variables, sous contraintes.

Le problème peut s'écrire sous la même forme que le krigeage simple ($[Aij] * [Wi] = [Bip]$) avec dans le cas du krigeage ordinaire :

- $[Aij]$ une matrice carrée d'ordre (n +1)*(n+1) ;
- $[Wn]$ et $[Bip]$ des matrices d'ordre (n+1)*1 (vecteurs colonnes).

La visualisation du système du krigeage ordinaire sous forme matricielle s'écrit donc :

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} & \bullet & C_{1n} & 1 \\ C_{21} & C_{22} & \bullet & \bullet & \bullet & 1 \\ C_{31} & \bullet & \bullet & \bullet & \bullet & 1 \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ C_{n1} & \bullet & \bullet & \bullet & C_{nn} & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} * \begin{pmatrix} W_1 \\ W_2 \\ \bullet \\ \bullet \\ W_n \\ \lambda \end{pmatrix} = \begin{pmatrix} C_{1p} \\ C_{2p} \\ \bullet \\ \bullet \\ C_{np} \\ 1 \end{pmatrix}$$

ou

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \bullet & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \bullet & \bullet & \bullet & 1 \\ \gamma_{31} & \bullet & \bullet & \bullet & \bullet & 1 \\ \bullet & \bullet & \bullet & \bullet & \bullet & 1 \\ \gamma_{n1} & \bullet & \bullet & \bullet & \gamma_{nn} & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} * \begin{pmatrix} W_1 \\ W_2 \\ \bullet \\ \bullet \\ W_n \\ -\lambda \end{pmatrix} = \begin{pmatrix} \gamma_{1p} \\ \gamma_{2p} \\ \bullet \\ \bullet \\ \gamma_{np} \\ 1 \end{pmatrix}$$

Comme évoqué plus haut, le Krigeage consiste à calculer les Wi (les poids) correspondant aux n points choisis de l'équation. Ce système comporte **une inconnue et une équation de plus** que le système de krigeage simple. Il faut donc résoudre les équations :

$$W_1\gamma_{(h11)} + W_2\gamma_{(h12)} + \ldots\ldots + W_n\gamma_{(h1n)} - \lambda = \gamma_{(h1p)}$$
$$W_1\gamma_{(h21)} + W_2\gamma_{(h22)} + \ldots\ldots + W_n\gamma_{(h2n)} - \lambda = \gamma_{(h2p)}$$
$$\vdots$$
$$W_1\gamma_{(hn1)} + W_2\gamma_{(hn2)} + \ldots\ldots + W_n\gamma_{(hnn)} - \lambda = \gamma_{(hnp)}$$
$$W_1 + W_2 + \ldots\ldots + Wn = 1$$

Une fois les poids (Wi) des différents points calculés, la valeur estimée pour le point x_p est :

$$\hat{Z}_p = \sum WiZi$$

III.3. Modèles non stationnaires

III.3.1. Généralités

La stationnarité requise dans les cas précédents (krigeage simple et krigeage oridinaire) suppose que le comportement de la variable régionalisée est invariant dans l'espace. Or dans certains cas, cette supposition ne peut pas être retenue. A titre d'exemples :

- en bathymétrie, la profondeur augmente en s'éloignant de la côte ;
- en géothermie, le gradient augmente avec la profondeur ;
- En hydrologie, la charge hydrique diminue dans le sens de l'écoulement ;
- En environnement, la charge en polluants diminue en s'éloignant de la source polluante.

Dans de tel cas, les modèles stationnaires se révèlent insuffisants pour rendre comptes de la régionalisation, d'où la nécessité de recourir à des approches permettant de traiter les processus non stationnaire (krigeage universel, krigeage intrinsèque d'ordre K, krigeage transitif, Krigeage avec dérive externe, Cokrigeage etc.). Le cas le plus simple, et le seul qui sera évoqué, est celui du krigeage ponctuel univarié à variable non-stationnaire qui contient une tendance connu sous le nom de Krigeage universel.

III.3.2. Krigeage universel

Le krigeage universel est basé sur une dichotomie de la régionalisation, c.-à-d. la décomposition de la fonction aléatoire Z(x) en deux composantes :

$$\boldsymbol{Z(x) = Y(x) + m(x)}$$

- Y(x) résidu aléatoire, qui satisfait de « bonnes » conditions de stationnarité ;
- m(x) dérive de la fonction aléatoire.

La composante Y(x) peut être traitée par les méthodes de la Géostatistique stationnaire.

$$m(x) = \sum a^{\ell} f_i^{\ell}$$

- les a^{ℓ} sont des coefficients inconnus ;
- les f_i^{ℓ} sont un jeu de fonctions de base, avec la première d'indice 0 toujours fixe ($f_0^1 = 1$) et toutes les autres connues.

Le système du krigeage universel s'écrit :

$$\sum W_i \, C_{i,j} + \lambda_o + \sum a^{\ell} f_i^{\ell} = C_{i,p}$$

ou

$$-\sum W_i \, \gamma_{i,j} + \lambda_o + \sum a^{\ell} f_i^{\ell} = -\gamma_{i,p}$$

Avec :
- i = 1, 2 ……n
- j = 1, 2 ……n
- λ_o : Lagrangien du krigeage ordinaire
- $\ell = 0, 1, 2$ ……k
- (p) : le point à estimer

Là aussi, la condition d'universalité ($\sum$Wi = 1) est imposée.

En plus, une autre condition s'ajoute : $\sum W_i f_i^{\ell} = f_p^{\ell}$

Le problème semble consister à estimer les Wi et les a^{ℓ}. En fait, la variance de l'erreur d'estimation ne dépend pas des a_{ℓ} (la démonstration mathématique dépasse les ambitions de ce cours). Par conséquent, la visualisation du système du krigeage universel sous forme matriciel s'écrit :

$$\begin{pmatrix} C_{ij} & f_i^{\ell} \\ {}^{t}f_i^{\ell} & 0 \end{pmatrix} * \begin{pmatrix} W_i \\ \lambda^{\ell} \end{pmatrix} = \begin{pmatrix} C_{ip} \\ f_p^{\ell} \end{pmatrix} \quad \text{ou} \quad \begin{pmatrix} \gamma_{ij} & f_i^{\ell} \\ {}^{t}f_i^{\ell} & 0 \end{pmatrix} * \begin{pmatrix} W_i \\ -\lambda^{\ell} \end{pmatrix} = \begin{pmatrix} \gamma_{ip} \\ f_p^{\ell} \end{pmatrix}$$

Les composantes C_{ij} ; γ_{ij} ; Wi et γ_{ip} étant celles des matrices du système de krigeage ordinaire et λ^{ℓ} est le vecteur des paramètres de Lagrange.

La présentation matricielle (cas de $\gamma_{(ij)}$) est la suivante :

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \bullet & \gamma_{1n} & 1 & f_{01} & f_{11} & \bullet & f_{k1} \\ \gamma_{21} & \bullet & \bullet & \bullet & 1 & f_{02} & \bullet & \bullet & f_{k2} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \gamma_{n1} & \bullet & \bullet & \gamma_{nn} & 1 & f_{0n} & \bullet & \bullet & f_{kn} \\ 1 & 1 & \bullet & 1 & 0 & 0 & \bullet & \bullet & 0 \\ f_{01} & f_{02} & \bullet & f_{0n} & 0 & \bullet & \bullet & \bullet & \bullet \\ f_{11} & f_{12} & \bullet & f_{1n} & 0 & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ f_{k1} & \bullet & \bullet & f_{kn} & 0 & 0 & \bullet & \bullet & 0 \end{pmatrix} * \begin{pmatrix} W_1 \\ W_2 \\ \bullet \\ \bullet \\ W_n \\ -\lambda_0 \\ -\lambda_1^{\ell} \\ \bullet \\ \bullet \\ \bullet \\ \lambda_k^{\ell} \end{pmatrix} = \begin{pmatrix} \gamma_{1p} \\ \gamma_{2p} \\ \bullet \\ \bullet \\ \gamma_{np} \\ 1 \\ f_p^1 \\ \bullet \\ \bullet \\ \bullet \\ f_p^k \end{pmatrix}$$

En supposant qu'il y ait « n » points de donnée et (k + 1) fonctions de base, il s'agirait alors d'un système à n + 1 + k équations et inconnues (les Wi et les λ^{ℓ}).

$$Wi = \sum_{\ell} W_{\ell}^{i} f_0^{\ell} \qquad \lambda^{\ell} = \sum_{n} \lambda_i^{\ell} f_i^{\ell}$$

III.4.Comparaison

Comme noté auparavant, il ne s'agit pas de krigeage"s" différents mais d'un seul et unique krigeage dont le système de base est A*W= B. Soit n le nombre de données et k le nombre de fonctions de base, le tableau ci-dessous résume, pour les trois techniques discutées, les dimensions des matrices, les conditions et les nombres d'équations et d'inconnues.

	Krigeage simple	Krigeage ordinaire	Krigeage universel
Matrice A	(n*n)	(n + 1) *(n + 1)	(n+ 1+ k)*(n + 1 + k)
	$\left[\gamma_{ij}\right]$	$\begin{bmatrix} \gamma_{ij} & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} \gamma_{ij} & 1 & f_i^{\ell} \\ 1 & 0 & 0 \\ {}^{t}f_1^{\ell} & 0 & 0 \end{bmatrix}$
Matrice W	(n*1)	(n + 1)*1	(n + 1 + k)*1
	$[W_i]$	$\begin{bmatrix} W_i \\ -\lambda_0 \end{bmatrix}$	$\begin{bmatrix} W_i \\ -\lambda_0 \\ -\lambda^{\ell} \end{bmatrix}$
Matrice B	(n*1)	(n + 1)*1	(n + 1 + k)*1
	$\left[\gamma_{ip}\right]$	$\begin{bmatrix} \gamma_{ip} \\ 1 \end{bmatrix}$	$\begin{bmatrix} \gamma_{ip} \\ 1 \\ f_p^{\ell} \end{bmatrix}$
Conditions	Variogramme borné	∑Wi = 1 Avec i = 1…. à n	∑Wi = 1 $\sum W_i f_i^{\ell} = f_p^{\ell}$ Avec i = 1, 2, …..à n ℓ = 0,1,2 …..à k
Equations & inconnues	n	n + 1	n + k + 1

III.5. Validation croisée

Comme déjà souligné, les deux étapes (variographie et krigeage) ne sont pas figées et il est naturel en cours de travail de revenir en arrière et d'affiner ou corriger un modèle à la lumière des premiers résultats obtenus. La méthode la plus simple pour cette démarche est connue sous le nom de validation croisée. Elle consiste à retirer une à une les observations (z_i) pour ensuite les estimer par krigeage ($\hat{z}_i$) à partir des autres données. L'écart est appelé erreur d'estimation (ε). On calcul alors la moyenne de cette erreur (μ_ε) appelée **EQM** pour « Erreur Quadratique Moyenne » ; l'**EQNM** (Erreur Quadratique standardisée (Normalisée) Moyenne) et leurs variances.

- EQM : $\mu_\varepsilon = \frac{1}{n}\sum(\hat{z}_i - z_i)$
- Variance de l'EQM : $\sigma_\varepsilon^2 = \frac{1}{n}\sum(\hat{z}_i - z_i)^2$
- EQNM : $\widehat{\mu_\varepsilon} = \frac{1}{n}\sum\left(\frac{(\hat{z}_i - z_i)}{\sigma}\right)$
- Variance de l'EQNM : $\widehat{\sigma_\varepsilon^2} = \frac{1}{n}\sum\left(\frac{(\hat{z}_i - z_i)}{\sigma}\right)^2$

L'EQM qui **doit tendre vers zéro**, permet de vérifier si le krigeage est effectivement non-biaisé. Les points pour lesquels EQNM est comprise dans l'intervalle [-2,5 ; 2,5] (par analogie au cas de la loi normale centrée réduite, où cet intervalle contient 95% des valeurs) sont dits robustes et retenus pour la suite des calculs.
La variance de l'EQNM qui **doit se rapprocher de 1**, correspond au rapport entre les variances expérimentale et théorique de krigeage. Elle permet de vérifier que les erreurs de krigeage sont cohérentes avec la variance calculée.
On cherche à minimiser l'EQM. On test plusieurs modèles et ceux obtenant les plus petits EQM sont jugés meilleurs. Une façon du choix du modèle (variogramme théorique) est donc de retenir celui qui minimise l'EQM. Cependant, il peut être préférable de sélectionner un modèle présentant une EQM légèrement moins bonne, si la variance de l'EQNM est proche de 1.

IV. Références (il ne s'agit pas d'une liste bibliographique exhaustive mais des travaux d'où émane l'essentiel de ce document).

Allard D. 2012. *Statistiques spatiales : introduction à la Géostatistique.* Universités Montpelier (I et II).

Baccini A., 2010. Statistique descriptive élémentaire. *Publications de l'Institut de Mathématique de Toulouse.*

Baillargeon S. 2005. *Le krigeage : revue de la théorie et application à l'interpolation spatiale des données de précipitations*. Université de Laval.

Benzécri J.P., 1976. Histoire et préhistoire de l'analyse des données. Partie I La préhistoire. *Les cahiers de l'analyse des données,* tome 1, n° 1, p. 9-32.

Chevet P. 2008. *Aide-Mémoire de Géostatistique linéaire*. Les Presses de l'Ecole des mines de Paris.

Cressie N.A.C., 1993. *Statics for spatial data*. Iowa University, Wiley-Interscience Publication.

Cronbach, L. J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, *16, p. 297–334.*

David, H. A., Hartley, H. O., & Pearson, E. S., 1954. The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika, 41(3/4), p. 482-493.*

Dixon W. J., 1953. Processing data for outliers. *Biometrics, 9(1), p. 74-89*.

Emery X. 2001. *Géostatistique linéaire*. Centre de Géostatistique de l'Ecole des Mines de Paris.

Gratton Y. 2002. *Le krigeage : la méthode optimale d'interpolation spatiale*. Institut National de la Recherche Scientifique (Eau-Terre-Environnement) du Quebec.

Grubbs F. E., 1969. Procedures for detecting outlying observations in samples. *Technometrics, 11(1), p. 1-21.*

Guerreau A., 2004. Statistique pour historiens. *Cours pour l'école des Chartes, Sorbonne, 164 p.*

Hengl T. 2007. *A Pratical Guide to Geostatistical mapping of Environnemental variables*. Office for Official Publications of the European Communities (Luxembourg).

Matheron G., 1969. *Le krigeage universel (Recherche d'estimateurs optimaux en présence d'une dérive)*. Edité par l'Ecole Nationale Supérieure des Mines de Paris.

Matheron G., 1970. *La Théorie des variables régionalisées, et ses applications.* Les cahiers du Centre de Morphologie Mathématique de Fontainbleau.

Oriol J.-C., 2007. Formation à la Statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un Département d'IUT. *Thèse de Doctorat, Université Lyon 2, 277 p.*

Saulnier S. 2012. « Arithmétique politique » et bataille de(s) chiffres. *Revue : Mots. Les langages du politique, n° 100, p. 15-29.*

Rivoirard J., 1995. *Concepts et méthodes de la Géostatistique*. Centre de Géostatistique, Ecole de Mines de Paris.

Vessereau A., 1972 (12ème édition). La Statistique. *Revue Que sais-je, n° 281* (1ère édition 1947 rééditée plusieurs fois par «*Imprimerie des Presses universitaires de France* »).

Wackernagel H. 1993. *Cours de Géostatistique Multivariable*. Centre de Géostatistique, Ecole de Mines de Paris.

Webographie

- http://wikistat.fr/
- http://www.jehps.net/