1. Rappels statistiques et introduction

Variable aléatoire (v.a.) : fonction dont les résultats possibles sont connus mais dont le résultat final ne peut être déterminé, à priori, avant d'effectuer la mesure.

ex. : - teneur de cuivre d'une carotte de 1 m
- épaisseur d'une veine minéralisée
- concentration d'un polluant dans l'eau souterraine
- pH de l'eau de pluie

Description d'une v.a. : sans connaître la valeur que prendra le résultat final, on peut parfois connaître la probabilité qu'une v.a. prenne chacun des résultats possibles. C'est la description la plus complète que l'on puisse faire de la v.a.

La fonction qui décrit ces probabilités est la fonction de densité (pour les v.a. continues; pour les v.a. discrètes, c'est la fonction de masse).

Propriétés : $f_X(x) \geq 0$, toute probabilité est positive

$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$, l'intégrale de la fonction de densité donne 1

$\int_a^b f_X(x)\,dx = P(a \leq X \leq b)$, probabilité que x prenne une valeur comprise entre [a et b]

Certaines quantités résument les caractéristiques principales de la variable aléatoire.

***Mesures de tendance centrale:**

- mode : x tel que $f_x(x)$ est maximum
- médiane : x tel que $P(X < x) = 0.5$
- moyenne (ou espérance mathématique) :

$$\mu_X \text{ ou } E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\,dx$$

***Mesures de dispersion :**

-Variance :

$$\sigma_X^2 = E[(X - E[X])^2]$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)\,dx$$

-Écart-type :

$$\sigma_X = \sqrt{\sigma_X^2}$$

-Asymétrie :

$$E\left[\left(\frac{X - E[X]}{\sigma_X}\right)^3\right]$$

-Aplatissement : $$E\left[\left(\frac{X - E[X]}{\sigma_X} \right)^4 \right]$$

Toutes ces quantités sont généralement, à priori, inconnues. On doit donc les estimer à partir d'un ensemble d'observations appelé l'échantillon (par abus de langage, on parlera souvent des échantillons pour désigner ces observations).

À partir de l'échantillon, on peut construire des estimateurs:

de la moyenne:

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

de la variance:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \bar{\sigma}^2 \qquad \text{ou} \qquad \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$$

de la fonction de densité : histogramme,

de la fonction de densité cumulative : courbe des fréquences cumulées $F_X(x) = P(X \leq x)$ estimée par: rang $(x_i)/n$

Une des caractéristiques importantes d'un estimateur est d'être sans biais i.e. d'avoir la même espérance mathématique que la quantité qu'il cherche à évaluer.

Ex. :

$$E[\bar{X}] \quad \frac{1}{n} E\left[\sum_{i=1}^{n} X_i \right] = \mu_X \quad \therefore \quad \bar{X} \text{ } est \text{ } sans \text{ } biais \text{ } pour \quad \mu_X$$

de même, s^2 est sans biais pour σ_x^2 alors que $\hat{\sigma}^2$ est biaisé

Passage à plus d'une variable :

On peut aussi étudier et décrire le comportement simultané de plus d'une variable aléatoire.
La fonction de densité conjointe : f_{xy} (x,y) donne la probabilité que, simultanément X = x et Y = y.

On a :

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx\, dy = 1 , \quad f_{XY}(x, y)$$

$$P[\, x_1 < X < x_2 , \; y_1 < Y < y_2 \,] = \int_{x1}^{x2} \int_{y1}^{y2} f_{XY}(x, y)\, dx\, dy$$

Deux mesures additionnelles permettent de décrire des caractéristiques importantes de fonction de densité conjointe.

La covariance:

$Cov(X,Y) = E[\,(X - \mu_X)(Y - \mu_Y)\,]$ mesure la force du lien linéaire entre les variables X et Y.

La corrélation

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \, \sigma_Y}$$

comme la Cov mais avec des unités "normalisées"

Propriétés de ρ_{XY} :

$$-1 \leq \rho_{XY} \leq 1$$

$$\rho_{XY} = \rho_{aX,bY}$$

(avec a et b des constantes quelconques)

Note : $\rho_{XY} = 0$ ---> absence de lien linéaire
≠ indépendance de x et y (en effet, on a indépendance ssi $f_{XY}(x,y) = f_X(x).f_Y(y)$).
Par contre, l'indépendance de X et Y ---> $\rho_{XY} = 0$.

L'interprétation propre à la géostatistique

Les v.a. sont régionalisées i.e. elles dépendent de leur localisation dans le gisement.

Z(x) Ex. Z : teneur de cuivre mesurée au point x.
(ou dans un volume centré en x)

Différentes visions du même gisement : G

- collection infinie de valeurs ponctuelles

$$Z_G = \frac{1}{G} \int_G Z(x)\, dx$$

Z_G est la teneur moyenne du gisement obtenue en faisant la moyenne de toutes les valeurs ponctuelles.

- collection finie de petits blocs v

$$Z_G = \frac{1}{N} \sum_{i=1}^{N} Z_v(x)$$

- collection finie de gros blocs V

$$Z_G = \frac{1}{M} \sum_{i=1}^{M} Z_V(x)$$

et ainsi de suite... Le gisement est donc assimilé à un ensemble fini ou infini (cas ponctuel) de variables aléatoires. Si on connaît le comportement de la variable aléatoire au niveau ponctuel (ou quasi-ponctuel) alors on peut aussi décrire le comportement de Z_v, Z_V et Z_G.

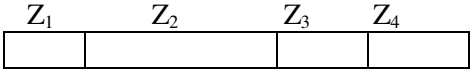
Cette collection de variables aléatoires s'appelle fonction aléatoire. Le gisement en est une réalisation limitée dans le temps et dans l'espace. On cherchera à caractériser Z(x) pour pouvoir dire quelque chose sur Z_v, Z_V et Z_G.

Support des observations :

Dans la pratique, Z(x) ne sera jamais mesuré sur un support ponctuel mais sur un support physique relativement très petit par rapport à la taille du gisement (disons v avec v << G). Il est de toute première importance de s'assurer que toutes les observations proviennent de supports identiques.

En effet, les statistiques habituelles calculées sur des supports différents n'ont aucun sens physique précis.

Ex.



La teneur de la carotte entière n'est pas donnée par la simple moyenne arithmétique des teneurs des bouts de carotte; i.e.:

$$Z_c \neq \frac{1}{4} \sum_{i=1}^{4} Z_i$$

De plus, on pourrait démontrer que Var (Z_1) > Var (Z_3) > Var (Z_4) > Var (Z_2). Les variances sont inversement proportionnelles aux tailles des supports.

Ex.: Sans perte de généralité, supposons que les valeurs des teneurs de cuivre mesurées dans des carottes de 1 m ne montrent aucune corrélation d'une carotte à l'autre (i.e. Cov(Z_1,Z_1')=0).

Supposons que l'on regroupe les carottes de 1 m en carottes de 2 m. i.e. la teneur moyenne d'une carotte de 2 mètres (Z_2) formée de deux carottes de 1m. (Z_1 et Z_1') est:

$Z_2=(Z_1+Z_1')/2$

Si on avait

$$Var(Z_1) = \sigma_1^2 = Var(Z_1')$$

on aura maintenant

$$Var(Z_2) = \frac{\sigma_1^2}{2}$$

en effet

$$Var(Z_2) = Var\left(\frac{1}{2}(Z_1 + Z_1')\right) = \frac{1}{4}\left[Var(Z_1) + Var(Z_1') + 2\,Cov(Z_1, Z_1')\right]$$

$$= \frac{1}{2}\sigma_1^2$$

S'il y a des corrélations entre les carottes, on aura quand même Var (Z_2) < Var (Z_1).

On voit donc que la distribution statistique d'une v.a. est toujours définie en relation avec un support physique.

Quelques propriétés des distributions normales et lognormales :

Normale :

$$Z \longrightarrow N(\mu, \sigma^2) \rightarrow \frac{Z - \mu}{\sigma} \quad N(0,1)$$

Une table unique d'une N(0,1) suffit pour calculer les probabilités de toute loi normale.

La fonction de densité est:

$$\frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

Note: La moyenne, la médiane et le mode d'une loi normale sont égaux à μ.

Lognormale :

Z est lognormale avec moyenne "m" et variance s^2 si ln Z ~ N (u , $ß^2$).

Lien entre m, s^2 et u , $ß^2$

$$m = e^{\mu+\frac{\beta^2}{2}} \qquad \sigma^2 = m^2\,(e^{\beta^2} - 1)$$

Inversant les relations, on obtient:

$$\beta^2 = \ln\left(\frac{\sigma^2}{m^2} + 1\right) \quad \text{et} \quad \mu = \ln(m) - \frac{\beta^2}{2}$$

Note: Pour la loi lognormale, la médiane vaut e^{μ} et le mode vaut $e^{\mu-\beta^2}$.

Application des lois normale et lognormale: réserves en fonction d'une teneur de coupure c.

Notons : T(c) = tonnage au-dessus de la teneur de la coupure.
Q(c) = quantité de métal au-dessus de la coupure.
m(c) = teneur moyenne de ce qui est au-dessus de c.

T(c) en termes statistiques peut s'écrire P (Z > c) · T_o où T_o désigne le tonnage total du gisement.

$$T(c) = T_o \int_c^{\infty} f_Z(z)\, dz$$

$$Q(c) = T_o \int_c^{\infty} z\, f_Z(z)\, dz$$

$$m(c) = Q(c) / T(c)$$

Si on connaît la loi de distribution (et ses paramètres), on peut calculer ces quantités.

Loi normale	**Loi lognormale**
$T(c) = T_o \left(1 - F\left(\frac{c-m}{\sigma}\right)\right)$	$T(c) = T_o\, F\left(\frac{1}{\beta} \ln\left(\frac{m}{c}\right) - \frac{\beta}{2}\right)$
$Q(c) = m\,T(c) + T_o\, \sigma f\left(\frac{c-m}{\sigma}\right)$	$Q(c) = m\, T_o\, F\left(\frac{1}{\beta} \ln\left(\frac{m}{c}\right) + \frac{\beta}{2}\right)$

où : m: moyenne des Z

σ écart-type des Z
ß: écart-type des ln Z

$F(t) = \int_{-\infty}^{t} f_{N(0,1)}(x)\, dx$ (cumulative d'une N(0,1))

$f(t) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}$, (fonction de densité d'une N(0,1))

Conséquences: Les réserves in-situ vont dépendre de la loi de distribution et de ses paramètres.

Ex. : Supposons que l'on connaisse m et σ^2 pour les unités de sélection déterminées, voyons quelle est l'influence du type de la loi.

Soit m = 3%
c = 4.5%
σ = 1.5%
T_o = 1 Mt

Si la distribution est normale, on trouve :

$$T(4.5) = 1Mt\left(1 - F\left(\frac{4.5 - 3}{1.5}\right)\right) = 0.16\,Mt \qquad \text{note: F(1) = 0.84}$$

$$Q(4.5) = 3\% * 0.16Mt + 1Mt * 1.5\% f\left(\frac{4.5 - 3}{1.5}\right) = 0.84 \times 10^{-2}\ Mt \quad \text{note: f(1) = 0.24}$$

$$m(4.5) = \frac{0.84x10^{-2}\,Mt}{0.16\,Mt} = 5.25\%$$

Si la distribution est lognormale, on a alors:

$$\beta = \left\{\ln\left(\frac{2.25\%^2}{9\%^2} + 1\right)\right\}^{1/2} = 0.472$$

et

$$T(4.5) = 1Mt\ F\left(\frac{1}{0.472}\ln\left(\frac{3}{4.5}\right) - \frac{0.472}{2}\right) = 1Mt\ F\ (-1.095) = 0.14\,Mt$$

$$Q(4.5) = 1Mt * 3\% * F\left(\frac{1}{0.472}\ln\left(\frac{3}{4.5}\right) + \frac{0.472}{2}\right) = 1Mt * 3\% * F(-0.623) = 0.80 \times 10^{-2}\ Mt$$

$$m(4.5) = \frac{0.80x10^{-2}\,Mt}{0.14\,Mt} = 5.71\%$$

Si l'exploitation impose la sélection sur de plus gros blocs ayant un écart-type de 1% (au lieu de 1.5% dans l'exemple précédent), alors, toujours en supposant une loi lognormale,

ß = 0.325

et

T(4.5) = .08 Mt
$Q(4.5) = .42\times10^{-2}$ Mt
M(4.5) = 5.25%

la quantité de métal est réduite de moitié!

Si on définit le profit par T(c) . (m(c) - c), on passerait de 0.169×10^{-2} Mt à $.06\times10^{-2}$ Mt donc une réduction de près des 2/3 du profit escompté.

Deux problèmes complexes:

L'exemple précédent illustre l'importance de tenir compte du **support** sur lequel s'opère la sélection lors de l'exploitation. On reconnaît donc un premier problème: l'information dont on dispose est définie sur de petites unités échantillonnales (carottes, échantillons en vrac, canelures). Comment, à partir de cette information, prédire ce que sera la distribution d'unités de sélection d'un volume très supérieur?

Supposons que l'on connaisse la loi de distribution des teneurs des bloc. On peut calculer, comme on l'a fait à l'exemple précédent, les réserves récupérables en fonction des différentes teneurs de coupure. Ces réserves calculées correspondent à ce que nous obtiendrions si l'on connaissait effectivement la vraie teneur de tous les blocs du gisement. En pratique, ces vraies valeurs ne sont jamais connues et doivent être estimées à partir de **l'information disponible**. Quel estimateur choisir? Quelle quantité d'échantillonnage effectuer? Peut-on prédire maintenant ce qui sera effectivement récupéré plus tard à partir d'une quantité d'information supérieure?

Ces deux problèmes sont fondamentaux en géostatistique, on leur a donné un nom: **l'effet support et l'effet information**. L'effet support indique que la distribution des teneurs dépend de la taille des blocs que l'on considère. Ainsi pour un même tonnage extrait et supposant que l'on connaisse les vraies valeurs des blocs, *on retire toujours plus de métal si la sélection s'effectue sur de petits blocs plutôt que sur des gros blocs* (l'opération sur de petits blocs est plus sélective). L'effet information indique que l'on ne dispose pas des vraies teneurs des blocs qui nous intéressent mais seulement d'une estimation de celles-ci. Pour un même tonnage extrait, la sélection s'effectuant sur des blocs d'une taille donnée, *on récupérera toujours moins de métal avec un estimateur qu'avec les vraies valeurs.* Normalement plus on améliore l'estimateur, soit en recourant à de meilleures méthodes d'estimation, soit en augmentant le nombre de données, plus on retire de métal pour un même tonnage.

Un problème très important relié à l'effet information et à l'effet support est le problème de **biais conditionnel.** Très souvent, pour un tonnage extrait donné, on aura retiré beaucoup moins de métal que ne le prévoyait l'estimation, ce qui risque d'être ruineux pour la compagnie minière. Pour minimiser ce biais conditionnel, il faut utiliser des estimateurs qui tiennent compte à la fois de l'effet support et de l'effet information. C'est ce que fait le krigeage.

Exemple numérique (tiré de Armstrong, 1998[1])

Supposons une portion de gisement formée de 16 blocs eux-mêmes divisés en 4 parcelles. On connaît la vraie teneur de la parcelle du coin supérieur gauche. La teneur de coupure est 300 (représente les coûts pour extraire et traiter le bloc). On définit le profit comme $\sum_{i=1,t_i>300}^{16}(t_i - 300)$.

735		45	125	167
450		337	95	245
124		430	230	460
75		20	32	20

Les vraies teneurs des blocs sont également connues:

505	143	88	207
270	328	171	411
102	220	154	263
101	54	44	155

Profit maximum que l'on puisse atteindre?
Profit: (505-300)+(328-300)+(411-300)=205+28+111=344

Quels seraient les profits prévu et réalisé si l'on estime la teneur de chaque bloc par la parcelle connue?

Profit prévu: (735-300)+(450-300)+(337-300)+(430-300)+(460-300)=435+150+37+130+160=912
Profit réalisé: (505-300)+(270-300)+(328-300)+(220-300)+(263-300)=205-30+28-80-37=86

[1] M. Armstrong, 1998, Basic linear geostatistics, Springer.

Quels seraient les profits prévu et réalisé si l'on estime la teneur de chaque bloc par la moyenne des parcelles du bloc et des blocs voisins par le côté?

Rép: Les valeurs estimées seraient alors:

410	310.5	108	179
411.5	271.4	206.4	241.8
269.8	228.2	249.4	238.8
73	139.3	75.5	170.7

Profit prévu: (410-300)+(310.5-300)+(411.5-300)=110+10.5+111.5=232
Profit réalisé: (505-300)+(143-300)+270-300)=205-157-30=18

Une estimation obtenue par krigeage a fourni les valeurs suivantes:

442	190	142	204
354	276	212	279
189	226	216	271
99	81	88	125

Profit prévu: (442-300)+(354-300)=142+54=196
Profit réalisé: (505-300)+(270-300)=205-30=175

Dans cet exemple, on note que:
-Dans tous les cas, le profit réalisé est inférieur au profit optimum (effet information)
-Seul le krigeage a fourni un profit prévu s'approchant du profit réalisé (absence de biais conditionnel).

Quelques exemples de problèmes, dans le domaine minier, auxquels la géostatistique peut apporter une contribution:

- Lorsque les limites du gisement sont floues, sans contrôle structural, définies en fonction d'une teneur de coupure (ex. cuivre porphyrique) et que l'on doit déterminer l'emplacement d'un chantier d'abattage.

- A partir de l'information recueillie lors de l'exploration, déterminer la rentabilité du gisement en regard de différentes méthodes d'exploitation (elles influencent le type de sélection et la taille des blocs) et de différents scénarios économiques. Ceci peut ensuite être utilisé pour comparer divers projets entre eux et leur accorder une cote de priorité.

- Déterminer si des forages (ou un autre type d'échantillonnage) additionnels permettront de dégager suffisamment de profits supplémentaires pour couvrir leurs coûts.

- Aider à déterminer la séquence d'exploitation optimale permettant de maximiser les profits (on a généralement intérêt à exploiter les zones les plus riches d'un gisement le plus tôt possible).

- Aider à déterminer la teneur de coupure optimale.

- Aider à déterminer les contours optimaux d'une fosse d'exploitation à ciel ouvert.

- Prédire la teneur et la variabilité de la teneur du minerai envoyé au concentrateur (moulin) et ainsi aider à prédire le taux de rendement de celui-ci.

- Déterminer si un processus d'homogénéisation ("stockpiling", points de prélèvements multiples, etc.) est justifié afin d'améliorer le rendement du concentrateur. Comparer divers scénarios pour l'homogénéisation.

- Prédire le plus exactement possible la teneur du minerai qui sera exploité à court terme. Déterminer si le minerai ainsi extrait sera suffisant pour alimenter seul une fonderie ou s'il faudra prévoir importer du concentré d'autres mines et/ou d'autres pays.

Exemples d'applications de la géostatistique dans divers domaines.

- Estimation et planification des mines et des gisements pétroliers.
- Prospection géochimique et géophysique.
- Cartographie automatique (par ordinateur).
- Filtrage de signal.
- Simulations d'écoulements, prédiction et simulation de conductivités hydrauliques.
- Caractérisation de sites contaminés.
- Cartographie météorologique.
- Classification de sols.
- Estimation de la biomasse et de sa localisation en pêches.
- Estimation de la compaction du noyau imperméable d'un barrage (géotechnique).
- Répartition spatiale de la déformabilité des roches au pourtour d'une excavation.
- Charges hydrauliques et directions d'écoulement.
- Analyse et caractérisation d'images (biomédical, télédétection).
- Représentation numérique-analytique de surfaces pour la CAO-DAO.