École Polytechnique de Montréal
Département de mathématiques et de génie industriel et département C.G.M.

MTH2302C **–** *Probabilités et statistique*

Examen final – Hiver 2011, Jeudi 28 avril de 13h30 à 16h00

QUESTION 1 (10 points)

Soit un échantillon aléatoire de taille « n » (i.e. les X_i sont indépendants, i=1...n).

6 pts a) Quel est l'estimateur de σ^2 obtenu par la méthode des moments ?

L'estimateur $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ est sans biais pour σ^2.

4 pts b) Calculer le biais de l'estimateur obtenu par la méthode des moments en a) ?

QUESTION 2 (10 points)

Souvent, à propos des résultats d'un sondage donnant les intentions de vote, on lit ou entend : « Les résultats sont précis à plus ou moins 3 %, 19 fois sur 20. »

3 pts a) Que représente la partie d'énoncé « 19 fois sur 20 » ?

4 pts b) D'où vient le chiffre 3 % (donner l'expression le définissant) et à quelle taille d'échantillon correspond-t-il approximativement ?

3 pts c) Selon la théorie, a-t-on besoin d'un échantillon de taille très différente selon que la population sondée est constituée de 35 millions de personnes ou de 100 000 personnes ? Justifier.

QUESTION 3 (20 points)

On mesure la conductivité hydraulique (k en m/s) et la porosité (X sans unité) sur un échantillon de taille 30 prélevé dans un dépôt de sable. Les observations sont considérées indépendantes. On désire établir un modèle de régression avec constante liant $Y=\log_{10}(k)$ à la porosité :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

On vous fournit les quantités suivantes :

$\overline{y} = -4.93$	$\overline{x} = 0.26$	$\sum_{i=1}^{30}(y_i - \overline{y})^2 = 38.12$	$\sum_{i=1}^{30}(x_i - \overline{x})^2 = 0.35$	$\sum_{i=1}^{30}(y_i - \overline{y})(x_i - \overline{x}) = 3.49$

4 pts a) Calculer les estimateurs b_0 et b_1 des coefficients du modèle.

4 pts b) Calculer R^2.

3 pts c) Calculer SCE et CME.

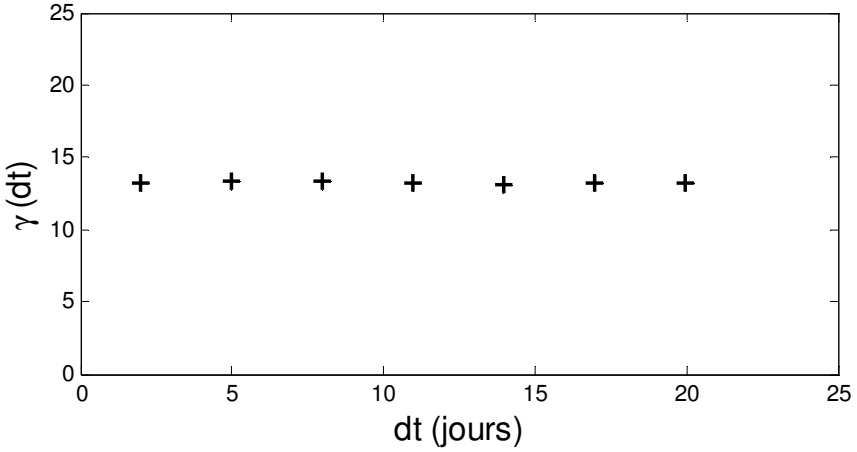
3 pts d) Est-ce que la régression est significative au seuil $\alpha = 5\ \%$ ($H_0\ \beta_1 = 0$ vs $H_1\ \beta_1 \neq 0$) ?

On vous fournit la matrice $(X'X)^{-1}$:

$$(X'X)^{-1} = \begin{bmatrix} 0.23 & -0.75 \\ -0.75 & 2.83 \end{bmatrix}$$

4 pts *e) Une nouvelle observation montre une porosité de 0.34. Donner la valeur prédite et l'intervalle de confiance de niveau 90 % sur la droite de régression à cette valeur 0.34 (i.e. intervalle de confiance sur $E[Y|x_i=0.34]$).*

On ordonne les données (et donc les résidus) selon le jour où la mesure a été effectuée. On calcule ensuite le variogramme expérimental des résidus du modèle en fonction de l'écart de temps entre les mesures (en jours) et l'on obtient :



2 pts *f) A-t-on l'évidence d'une dérive instrumentale dans le temps? Justifier.*

QUESTION 4 (15 points)

On veut estimer la tendance régionale dans un relevé gravimétrique (variable G; n=100 données) à l'aide d'une régression polynomiale en fonction des localisations (x,y). L'ordre du polynôme étant inconnu, on considère successivement des ordres croissants. Les résultats sont présentés dans le tableau suivant :

Ordre	Modèle	Nombre total de paramètres dans le modèle	R^2
0	$G=\beta_0+\varepsilon$	1	-
1	$G=\beta_0+\beta_1 x+\beta_2 y+\varepsilon$	3	0.2
2	$G=\beta_0+\beta_1 x+\beta_2 y+\beta_3 x^2+\beta_4 xy+\beta_5 y^2+\varepsilon$	6	0.35
3	$G=\beta_0+\beta_1 x+....+\beta_6 x^3+\beta_7 x^2 y+\beta_8 xy^2+\beta_9 y^3+\varepsilon$	10	0.39
4	$G=\beta_0+\beta_1 x+....+\beta_{12} x^2 y^2+\beta_{13} xy^3+\beta_{14} y^4+\varepsilon$	15	0.40
5	$G=\beta_0+\beta_1 x+....+\beta_{17} x^3 y^2+\beta_{18} x^2 y^3+\beta_{19} xy^4+\beta_{20} y^5+\varepsilon$	21	0.42

De plus, on sait que $SCT_m = 300$.

3 pts *a) À quelle statistique descriptive courante est égale le b_0 obtenu pour le modèle d'ordre 0 ?*

.

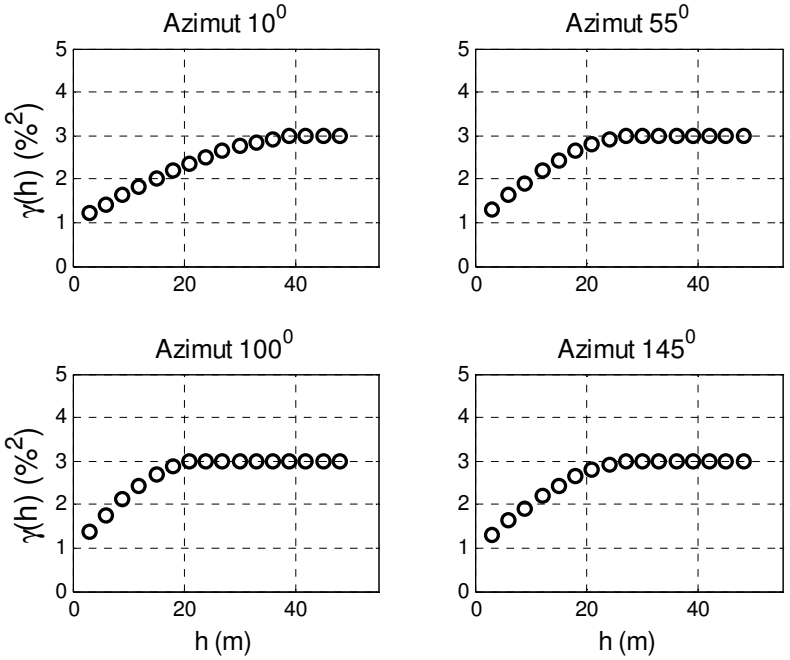
3 pts *b) Quel est le R^2 du modèle d'ordre 0 ?*

6 pts *c) Selon les résultats obtenus, au seuil de signification α = 5%, quel ordre de polynôme devrait-on retenir pour ce problème ?* (Aide : Partir du modèle d'ordre 0 comme référence et accroître l'ordre tant que l'ajout des variables est significatif).

3 pts *d) Supposons qu'en c) vous reteniez un polynôme d'ordre « k ». Si l'on calculait le variogramme des résidus obtenus avec le modèle d'ordre k-1 en fonction des distances entre localisations (x,y), verrait-on un effet de pépite pur sur les variogrammes expérimentaux des résidus ? Justifier.*

QUESTION 5 (10 points)

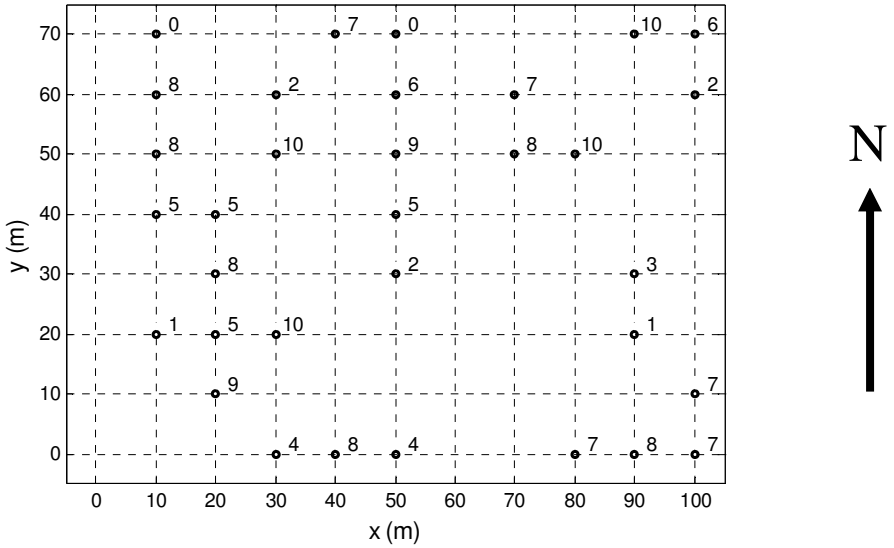
La figure suivante montre les variogrammes expérimentaux calculés selon quatre directions différentes dans un niveau d'un gisement de cuivre (Cu mesuré en %). Les directions pour le calcul sont choisies en fonction des directions géologiques préférentielles.



Décrivez le modèle 2D permettant un ajustement adéquat, simultanément, pour les quatre variogrammes expérimentaux.

QUESTION 6 (15 points)

On vous fournit le plan suivant où l'épaisseur de sédiments meubles est connue en quelques points.

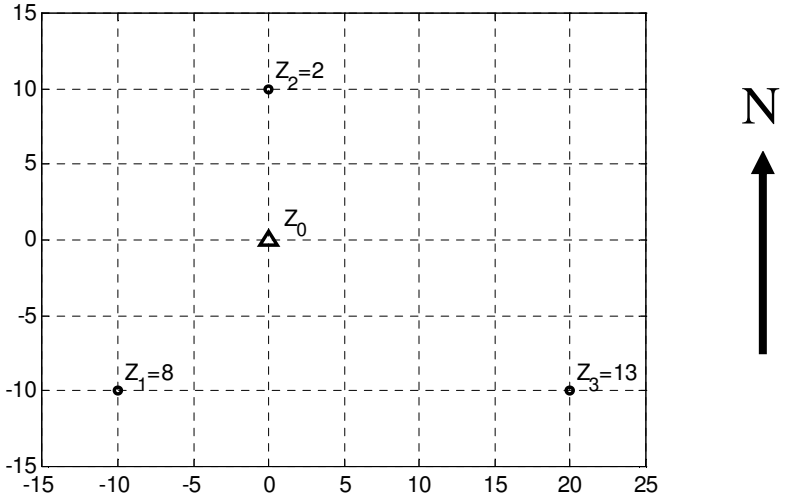


8 pts *a) Calculez le variogramme expérimental selon exactement la direction (azimut) 135^o pour la distance* $h = 10\sqrt{2}$ *m.*(Rappel : l'azimut est l'angle dans le sens horaire par rapport au nord). Indiquer clairement sur la figure toutes les paires utilisées pour le calcul.

7 pts *b) Indiquer deux raisons pour utiliser le variogramme, plutôt que d'estimer directement la fonction de covariance, pour caractériser la continuité spatiale.*

QUESTION 7 (20 points)

Un sable contaminé aux hydrocarbures montre les concentrations de carbone organique total (COT) indiquées sur la figure suivante (en %). On a identifié un variogramme sphérique avec anisotropie géométrique et portées $a_g = 50$ m et $a_p = 20$ m. La direction de plus grande portée est 20^0 (azimut). On note un effet de pépite de 4 $\%^2$ et un C = 10 $\%^2$. On veut kriger la concentration inconnue Z_0.



10 pts *a) Compléter le système de krigeage suivant en calculant la valeur A; fournir tous les détails du calcul.*(Note : les entrées dans la matrice et le vecteur suivent l'ordre des indices)

$$\begin{bmatrix} 14 & A & 0 & 1 \\ A & 14 & 0 & 1 \\ 0 & 0 & 14 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \mu \end{bmatrix} = \begin{bmatrix} 4.40 \\ 6.27 \\ 0 \\ 1 \end{bmatrix}$$

Supposons que l'on utilise les poids $\alpha_1 = 0.7$, $\alpha_2 = 0.0$ et $\alpha_3 = 0.3$ pour former l'estimateur $Z_0^* = \sum_{i=1}^{3} \alpha_i Z_i$ (note, ces poids **ne sont pas** les poids de krigeage).

3 pts *b) Cet estimateur est-il sans biais ? Justifier.*

4 pts *c) Quelle est la variance de l'erreur d'estimation (ou variance d'estimation) associée à cet estimateur ?*

3 pts *d) Que devient la variance d'estimation si les valeurs Z_1 à Z_3 doublent mais que l'on garde le même variogramme pour calculer les covariances ?*

Corrigé

Question 1

a) E[X] estimé par $\bar{x}$, E[X^2] estimé par $\frac{1}{n}\sum_{i=1}^{n} x_i^2$

or σ^2 = E[X^2]-E[X]^2 estimé par: $\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 1/n\,(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2) = (n-1)/n\ S^2$

b) E[estimateur]= (n-1)/n σ^2 , donc le biais = -1/n σ^2

En b), enlever 1 pt si signe inversé pour le biais

Question 2

a) C'est le niveau de signification de l'intervalle de confiance (19/20 = 95 %)

b) 3 % = .03 = $Z\alpha/2\sqrt{\frac{p(1-p)}{n}}$ Cette valeur est maximale pour p=0.5 d'où on tire que n = $\frac{1.96^2}{4\,(.03)^2}$ = 1067

c) Non la taille de la population n'intervient pas dans la formule tant que la population est assez grande pour que l'on puisse considérer le tirage avec remise.

Question 3

a) b_1 = 3.49/0.35 = 9.97 b_0 = -4.93 -9.97 * 0.26 = -7.52

b) r= 3.49 / (38.12 * 0.35)$^{0.5}$ = 0.96 => R^2 = 0.91

c) SCT_m = 38.12 SCE = (1-0.91) * 38.12 = 3.43 ; CME = 3.43/28 = 0.12

d) SCR_m = 0.91* 38.12 = 34.69 => F_0 = (34.69/1) / 0.12 = 289 >> $F_{1,28,\,5\%}$ = 4.2 => on rejette très fortement H_0 => régression très significative.

e) x_i=[1 0.34] on calcule CME*$x_i(X'X)^{-1}x_i'$ = 0.12*0.05 , $t_{28,\,5\%}$ = 1.7

valeur prédite : $x_i'b$ = -4.13 + ou – 1.7 (0.12*0.05)$^{0.5}$ = -4.13 + ou – 0.13 => [-4.26, -4.00]

f) Non, il n'y a pas de corrélation dans le temps (effet de pépite pur) donc pas de dérive instrumentale.

Question 4

a) à la moyenne arithmétique

b) SCE = SCT_m (par définition) or SCE+SCR_m = SCT_m => SCR_m = 0 => R^2 =0

c) ordre 1 vs 0 : F0 $= \frac{(0.2 - 0) * SCT_m / 2}{(1 - 0.2) * SCT_m /(100 - 3)} = 12.13$ > $F_{2,97,\,5\%}$ = 3.09, donc ajout significatif

ordre 2 vs 1 : F0 $= \frac{(0.35 - 0.2) * SCT_m / 3}{(1 - 0.35) * SCT_m /(100 - 6)} = 7.23$ > $F_{3,94,\,5\%}$ = 2.70 donc ajout significatif

ordre 3 vs ordre 2 : $F0 = \frac{(0.39 - 0.35) * SCT_m / 4}{(1 - 0.39) * SCT_m /(100 - 10)} = 1.48 < F_{4,90,5\%} = 2.47$ Donc l'ajout n,est pas significatif, on conserve l'ordre 2.

d) On verrait une structure car n'ayant pas le bon degré de polynôme, les résidus seront fortement corrélés à petite distance. Donc on n'aura pas un variogramme effet de pépite pur.

Question 5
Variogramme sphérique avec anisotropie géométrique et effet de pépite. Les paramètres sont a_g = 40 m, a_p =20 m, C_0 =1%^{2} et C=2 %^{2}. La direction de meilleure continuité (a_g est selon l'azimut 10^o).

Barème : Enlever 1 pt si les unités des portées et palier ne sont pas indiquées
Le mot anisotropie géométrique doit apparaître.
Si l'on indique 4 variogrammes différents => 3/10 au plus

Question 6
a) On trouve 9 paires selon l'orientation 135
donc $1/18 * ((4-9)^2+(9-1)^2+(8-10)^2+(5-8)^2+(1-7)^2+(5-8)^2+(7-6)^2+(7-10)^2+(10-2)^2) = 12.28$
b) On n'a pas à estimer la moyenne, ce qui évite un problème de biais pour l'estimation des covariances
On dispose de modèles sans palier pour lesquels on n,a pas l'équivalent en terme de covariance.

Question 7

a) On calcule $h = (100+400)^{0.5} = 22.36$

direction définie par $x_1 - x_2$ = atan(10/20) = 26.57 =`angle theta avec a_g = 6.57

La portée $a_\theta = 50*20/(20^2 \cos^2(6.57)+50^2 \sin^2(6.57))^{0.5} = 48.37$

C(h) =22.36,theta= 6.57) = $10*(1-(1.5*22.36/48.37 - 0.5*(22.36/48.37)^3)) = 3.56$

b) oui sans biais car somme des poids =1

c) a=[0.7 0 0.4]. $Var(e) = Var(Z_0) + Var(Z_0^*) - 2Cov(Z_0^*, Z_0) = 14 + aKa' - 2a*k =$
$14 + 14*(0.7^2+0.3^2) - 2* 0.7*4.40 = 15.96$

d) La variance ne change pas puisque celle-ci ne dépend que des poids et du modèle de variogramme.