ÉCOLE
POLYTECHNIQUE
MONTRÉAL

Le génie sans frontières

Questionnaire examen final

MTH2302D

Sigle d

Nom : La	
Signature :	**Groupe :**

Sigle et titre du cours	*Groupe*	*Trimestre*
MTH2302D Probabilités et statistique	**TOUS**	**HIVER 2009**
Professeur	***Local***	***Téléphone***
Marc BOURDEAU	**A-520.12**	**4941**

Jour	*Date*	*Durée*	*Heures*
Mercredi	**29 Avril 2009**	**2h30**	**9h30 à 12h00**

Documentation	*Calculatrice*	
☐ Aucune	☐ Aucune	**Les cellulaires, agendas électroniques ou téléavertisseurs sont interdits.**
☐ Toute	☐ Toutes	
☒ Voir directives particulières	☒ Non programmable	

Directives particulières

Voir directives à la prochaine page.

Important

Cet examen contient **3** questions sur un total de **10** pages **(incluant cette page)**

La pondération de cet examen est de **50** **%**

Vous devez répondre sur : ☐ le questionnaire ☒ le cahier ☐ les deux

Vous devez remettre le questionnaire : ☐ oui ☒ non

L'étudiant doit honorer l'engagement pris lors de la signature du code de conduite.

Directives

- Vous pouvez conserver le cahier des questions de l'examen ;
- Le questionnaire à proprement parler, les pages 3 à 10 de ce cahier, comporte 20 courts items répartis sur 3 questions, pour 50 points au total (15 + 10 + 25) ;
- Vous pouvez avoir à disposition pour l'examen les cahiers (non modifiés), disponibles sur le site du cours : (1) un mémento de 18 pages, ainsi que (2) 7 tables et abaques utiles pour l'examen. Les surveillants vérifieront le mémento ainsi que les tables pendant l'examen pour prévenir les fraudes possibles ;
- Même si ce n'est pas toujours indiqué, on doit *expliquer & justifier les réponses* pour avoir les points.

Temps d'exécution

1. **(Total 15 points : a, b, 3 points chaque item ; c, 4 points ; d, 5 points)** Supposons qu'il y ait constamment des requêtes en attente à un serveur informatique, que le temps de service d'une requête suive une loi exponentielle de moyenne un centième de seconde, et que ces temps de traitement soient indépendants les uns des autres.

   On sait qu'une somme de k exponentielles indépendantes de même paramètre λ suit une loi d'Erlang dont la cumulative est :

$$F(t; k, \lambda) = 1 - \sum_{n=0}^{n=k-1} \frac{e^{-\lambda t}(\lambda t)^n}{n!}.$$

   (a) Quelles sont la moyenne et l'écart type de la loi T d'exécution de 5 requêtes ?

   (b) Quelle est la probabilité que ce temps de service, T, dépasse 0,03 secondes ?

   (c) Supposez que le théorème central de la limite (TCL) s'applique pour le temps de traitement de 100 requêtes indépendantes. Calculez la valeur pour le temps de traitement de 100 requêtes qui ne sera dépassée que 1% des fois.

   (d) On aimerait tester si un certain changement dans l'architecture du système a permis d'améliorer le temps d'exécution d'une requête. Une expérience avec $n = 25$ réalisations indépendantes du temps de traitement cumulatif, noté aussi T, de 100 requêtes indépendantes au serveur a donné $\overline{T} = 0{,}985\ 74$ sec, avec un écart type $s_T = 0{,}099\ 8$ sec. On aimerait savoir si l'évidence de cet échantillon est suffisante pour penser que la vraie moyenne est inférieure à celle prévue dans l'ancienne architecture.

   Quelle est votre décision : avez-vous une évidence suffisante pour croire que la nouvelle architecture est meilleure que l'ancienne ?

   Note : Vous préciserez vos choix.

Conformités à des normes

2. **(Total 10 points : a, c, 2 points chacun ; b, d, 3 points chacun)** Une société produit environ 5 millions de résistances de nominal 300Ω par jour ouvrable. Les normes de production spécifient que la résistance doit suivre la loi suivante : $R \sim \mathcal{N}(\mu = 300\Omega, \sigma_R = 5\Omega)$. Les normes spécifient aussi les valeurs acceptables de conformité pour la résistance de 285Ω à 315Ω. La société désire améliorer sa production pour obtenir moins de 2200 résistances non conformes par million de production, au lieu du 2700 actuel.

   (a) Quelle est la valeur de σ_R, tout en supposant que la production est encore centrée sur le nominal de 300Ω, qui assurerait moins de 2200 résistances non conformes par million de production ?

   (b) Après une étude prospective des sources de variation de la production, diverses expériences furent menées pour en diminuer la variabilité. On a fini par établir une nouvelle séquence des étapes de la production qui, on l'espère, satisfera aux nouvelles exigences.

   Après une période de rodage de la nouvelle procédure pour la stabiliser, un échantillon de taille $n = 101$, pris tout au long d'une journée de production, a donné les valeurs suivantes : $\overline{R} = 300{,}547$ et $s_R = 4{,}122$.

      i. Sauriez-vous croire que le nominal de la production est encore valide dans la 'nouvelle production', ou s'il a dérivé, au seuil $\alpha_T = 0{,}05$?

      ii. Sauriez-vous croire que l'écart type $\sigma_R = 5$ est conservé dans la 'nouvelle production', ou s'il ne serait pas devenu plus petit, au seuil $\alpha_T = 0{,}05$?

   (c) Calculez un intervalle de confiance *bilatéral* de niveau 95% pour la vraie valeur de σ_R de la 'nouvelle production'.

   (d) En vous basant sur l'intervalle de confiance calculé à l'item 2c, et de la conclusion en 2(b)i, déterminez le nombre maximum de produits non conformes par million produits qu'on peut attendre dans la nouvelle production. On admettra aussi que la loi de $\overline{R}$ dans la nouvelles production est encore gaussienne.

Modéliser

Les réactions en génie génétique doivent être contrôlées de façon la plus économique et sûre possible. Dans l'équation suivante, Q représente la proportion d'une réaction qui reste à s'accomplir, A et B sont des constantes qui dépendent des conditions initiales et de la vitesse d'évolution de la réaction, variable pour chaque bac de réaction. On dit que Q suit un modèle *logistique* :

$$Q(t) = \frac{1}{1 + Ae^{Bt}}. \qquad (1)$$

On trouvera à la Fig. 1 une courbe typique d'évolution qu'on retrouve approximativement dans notre exemple (l'abscisse est en heures).

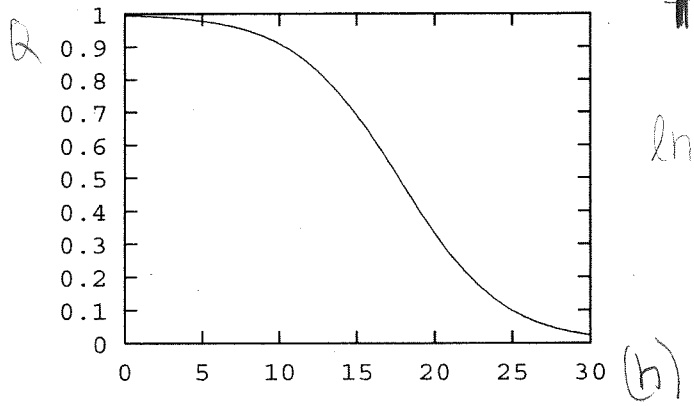

FIG. 1 – Un exemple de logistique sur presque tout son développement, avec des constantes près de celles du problème. L'abscisse est en heures.

On comprendra aux Fig. 2 que les mesures de la variable Q dans les début et fin de la réaction sont très difficiles à effectuer, étant donné la précision des appareils de mesures.

On obtient expérimentalement des valeurs (t_i, Q_i), où les t_i sont en heures. Pour déterminer les coefficients A et B, on utilise la régression linéaire simple effectuée sur la linéarisation suivante :

$$Q_1(t) \equiv \ln(\frac{1}{Q(t)} - 1) = \ln(A) + Bt \equiv \beta_0 + \beta_1 t. \qquad (2)$$

On a dû changer de fournisseur de l'enzyme qui sert de catalyseur. On doit tester alors les paramètres de la nouvelle réaction. On sait que pour l'ancienne souche on avait, suite à une longue utilisation, $\beta_0 = -5{,}24$, $\beta_1 = 0{,}31$.

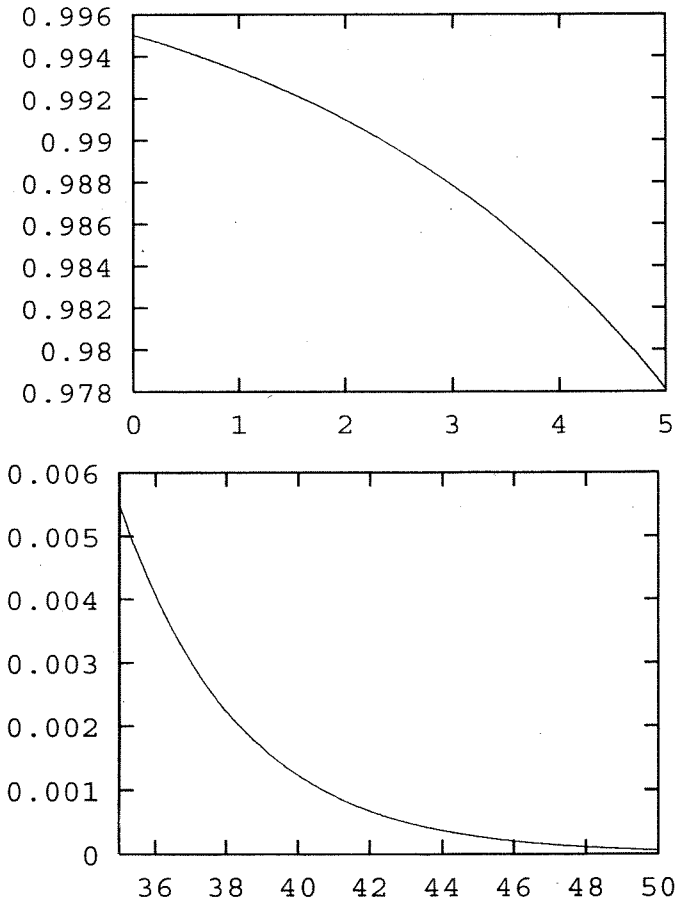

FIG. 2 – La lenteur de la réaction au début et à la fin de son développement complique la mesure.

On a effectué 5 essais avec la nouvelle souche. On a pris des mesures des valeurs de Q aux heures avec cette nouvelle enzyme. Les résultats des régressions simples, avec les coefficients b_0 et b_1 (issus du modèle (2)) sont présentés au tableau de la Fig. 3. On admettra que le modèle est valide.

Résultats des régressions pour 5 essais

		St. Err.		St. Err.		
N=38	BETA	of BETA	B	of B	t(36)	p-level
Intercpt			-5.47977	.073996	-74.0554	.000000
VAR1	.997960	.010641	.30838	.003288	93.7846	.000000
		St. Err.		St. Err.		
N=37	BETA	of BETA	B	of B	t(35)	p-level
Intercpt			-5.26596	.026879	-195.910	0.00
VAR1	.999705	.004109	.29757	.001223	243.325	0.00
		St. Err.		St. Err.		
N=38	BETA	of BETA	B	of B	t(36)	p-level
Intercpt			-5.31358	.031511	-168.627	0.00
VAR1	.999609	.004660	.29973	.001397	214.497	0.00
		St. Err.		St. Err.		
N=37	BETA	of BETA	B	of B	t(35)	p-level
Intercpt			-5.55870	.096102	-57.8414	.000000
VAR1	.996579	.013970	.31455	.004409	71.3350	.000000
		St. Err.		St. Err.		
N=39	BETA	of BETA	B	of B	t(37)	p-level
Intercpt			-5.21980	.065705	-79.4431	.000000
VAR1	.998288	.009615	.29509	.002842	103.8294	0.000000

FIG. 3 – Résultats principaux pour les paramètres du modèle linéarisé pour 5 essais avec prises de mesures aux heures avec la nouvelles enzyme.

3. **(Total : 25 points)**

   (a) **(2 points)** Développez la linéarisation qui passe de l'équation (1) à l'équation (2). Donnez aussi les liens entre les valeurs de A et β_0, et de B et β_1, que vous obtiendrez de l'équation (2).

   (b) **(2 points)** Déterminer pour le *premier essai de la nouvelle enzyme*, un intervalle de confiance de niveau 99% pour la valeur de β_0.

   (c) **(1 point)** En déduire un intervalle de confiance de même niveau pour la valeur du paramètre A du modèle original (i.e. le paramètre avant la linéarisation).

   (d) **(4 points)**

      i. Déterminez à partir des 5 réalisations b_1 de β_1 si la valeur moyenne de B pourrait bien celle de l'ancienne enzyme. Vous considérerez la contre hypothèse bilatérale avec $\alpha_T = 0{,}05$.

      ii. Peut-on admettre que les conditions d'application d'un test-T sont satisfaites ?

(c) **(4 points)** Supposons maintenant que vous vouliez tester l'hypothèse provenant de l'ancienne enzyme $H_0 : \beta_1 = 0{,}31$ avec un $\alpha_T = 0{,}01$. On admettra que l'écart type est *inconnu* mais au voisinage de $\sigma = 0{,}01$. Combien d'essais avec la nouvelle enzyme devriez-vous effectuer pour détecter avec une probabilité de 95% un écart de 0,01 de la valeur supposée dans H_0, avec la contre-hypothèse bilatérale ?

Deuxième partie. On sait qu'un modèle linéaire bien validé et de qualité permet de faire des prédictions dites *directes*, i.e. en nos termes, de t vers Q, avec un intervalle de prédiction pour Q.

En réalité, ce qui importe souvent, comme dans notre application, consiste à utiliser la régression dite *inverse*, ici de Q vers t. C'est le problème dit du *calibrage* : quel est le temps à partir duquel je puis être assuré qu'un certain Q est atteint, avec une incertitude contrôlée sur cette valeur ?

En principe, un modèle de régression ne peut être utilisé que pour la régression directe et non pour l'inverse : l'inversion de l'équation linéaire ne permet nullement de trouver des intervalles de confiance pour les prédictions inverses. On n'a qu'à regarder les équations suivantes, d'un modèle linéaire quelconque, pour comprendre :

$$y = \beta_0 + \beta_1 x \Longrightarrow x = \frac{y - \beta_0}{\beta 1} \qquad (\beta_1 \neq 0)\,. \tag{3}$$

La régression inverse n'est pas théoriquement valide. Comment en effet estimer la qualité de l'estimation du x pour un y donné ?

Mais, on a étudié ce problème qui est fréquent dans les applications, et on a montré les résultats suivants qu'on indique ici dans les termes de l'exemple du modèle de l'équation (2), avec les $Q(t)$ transformés par la linéarisation en $Q_1(t)$. L'inversion de l'équation (2) donne :

$$\widehat{t^*} = \frac{Q_1 - b_0}{b_1}\,.$$

Ainsi, cette expression donne la valeur $\widehat{t^*} \approx 40{,}038\ 72\,\text{h}$ pour la valeur de $Q = 0{,}001$ (soit $Q_1 \approx 6{,}91$) dans le cas du dernier essai réalisé cette avec la nouvelle enzyme, cette fois avec très peu d'observations, soit $n = 5$. (*cf.* les deux tableaux du haut de la Fig. 4). Ce $Q = 0{,}001$ indique un produit très bien purifié qui ne demande que peu de traitement ultérieur. Le temps $\widehat{t^*}$ représente une estimation du temps où ce degré de pureté est atteint.

De plus, la théorie nous indique qu'on peut construire l'intervalle de confiance approximatif de niveau $1-\alpha$ autour d'un $\widehat{t^*}$ calculé en une proportion Q_1 désirée, à l'aide de l'équation suivante :

$$\widehat{t^*} \pm T_{n-2;1-\frac{\alpha}{2}}\ \widehat{\sigma}_{\text{prédiction}}\,, \tag{4}$$

avec, pour $\widehat{\sigma}_{\text{préd-Inv}}$ une expression liée à $\widehat{\sigma}_{\text{préd}}$

$$\widehat{\sigma^2}_{\text{préd-Inv}} = \frac{MS_R}{b_1^2}\left[\,1 + \frac{1}{n} + \frac{(\widehat{t^*} - \overline{t})^2}{SS_{xx}}\,\right], \tag{5}$$

pourvu que l'expression suivante soit vérifée :

$$\frac{\left[T_{n-2;1-\frac{\alpha}{2}}\right]^2 MS_R}{b_1^2\, SS_{xx}} \leq 0{,}1\,. \tag{6}$$

(f) **(2 points)** Expliquer pourquoi la régression inverse de l'équation (3) n'est pas théoriquement valide.

(g) **(2 points)** Donnez une expression générale tirée de (4) d'un intervalle de confiance *unilatéral à gauche* de niveau $(1-\alpha)$ pour une valeur $\widehat{t^*}$.

Attention. On ne demande pas de calcul, mais une expression générale.

Pour les questions qui suivent qui exigent des données, on utilisera les résultats rapportés aux tableaux de la Fig. 4

Note. Dans cet essai on n'a pas pris des mesures toutes les heures, comme pour les résultats rapportés au tableau de la Fig. 4, mais seulement 5 mesures bien réparties dans le temps : le modèle ayant été validé, on peut économiser beaucoup de mesures sur chacun des bacs de réaction pour estimer ce qu'on cherche.

(h) **(3 points)** Utilisez l'*expression* du test-T pour l'hypothèse H_0 : $\beta_1 = 0$, dont la valeur est rapportée au tableau du milieu de la Fig. 4, pour voir sans presque calculer si l'équation (6) est valide, avec $\alpha = 0{,}01$.

(i) **(2 points)** Admettez ici que l'équation (6) est valide, quel que soit le résultat de votre calcul à l'item 3h. Utilisez l'expression et la valeur du $\widehat{\sigma}_{\text{préd-Inv}}$ de la prédiction à la valeur $\widehat{t^*} \approx 40{,}04$, que vous calculez à partir des résultats rapportés à la Fig. 4 (tableau du bas), pour enfin obtenir, sans presque calculer, l'intervalle de confiance unilatéral à gauche de niveau 99,5% pour cette valeur de $\widehat{t^*}$.

(j) **(3 points)** Donnez une interprétation de cet intervalle de confiance à gauche pour la valeur de $\widehat{t^*}$ au point $Q = 0{,}001$, en terme de la décision d'arrêter la réaction.

En d'autres termes, *justifiez* l'arrêt de la production au moment fixé par la limite à droite de l'intervalle de confiance trouvé. On peut répondre à cette question sans avoir répondu correctement aux précédentes.

Résultats de la régression pour 1 essai avec moins d'observations

```
STAT.       Analysis of Variance; DV: LIN25 (enzyme.sta)
MULTIPLE
REGRESS.

                Sums of                      Mean
 Effect         Squares        df          Squares          F          p-level

Regress.       80.14307         1         80.14307     1367.287        .000044
Residual         .17584         3           .05861
 Total         80.31891


STAT.       Regression Summary for Dependent Variable: LIN25
MULTIPLE    R= .99890474 R²= .99781068 Adjusted R²= .99708090
REGRESS.    F(1,3)=1367.3 p<.00004 Std.Error of estimate: .24210

                      St. Err.                     St. Err.
N=5              BETA  of BETA          B           of B          t(3)        p-level

Intercpt                             -5.27903      .179538     -29.4033       .000086
    VAR1         .998905  .027014      .30435      .008231      36.9768       .000044
```

```
STAT.       Predicting Values for (enzyme.sta)
MULTIPLE    variable: LIN25
REGRESS.

                                                    B-Weight
variable          B-Weight          Value           * Value

    VAR1          .304351        40.03872          12.18582
Intercpt                                           -5.27903

Predictd                                            6.90680
-95.0%PL                                            5.87528
+95.0%PL                                            7.93832
```

FIG. 4 – Les résultats de la régression sur modèle linéarisé avec 5 observations seulement (haut) ; les résultats de la prédiction avec $t = 40{,}038\ 72$ (bas).