ÉCOLE POLYTECHNIQUE

**Département de mathématiques et de génie industriel**

**MTH 2302B – Probabilités et statistique**

**Examen final 16 décembre 2006 - 13h30 à 16h00**

NOM _____ PRÉNOM _____
(caractères d'imprimerie) (caractères d'imprimerie)

SIGNATURE _____ MATRICULE _____

| NUMÉRO | 1a | 1b | 1c | | 2a | 2b | 2c | 2d | | 3a | 3b | 3c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| valeur | 1 | 1 | 2 | | 1 | 1 | 1 | 2 | | 1 | 1 | 1 |
| obtenu | | | | | | | | | | | | |

| NUMÉRO | 3d | 3e | 3f | | 4a | 4b | 4c | | | | | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| valeur | 1 | 1 | 1 | | 1 | 1 | 3 | | | | | 20 |
| obtenu | | | | | | | | | | | | |

## NUMÉRO 1 (4 points)

Des mesures de turbidité sur des échantillons d'eau potable ont donné les résultats suivants :

| | | | | | |
|---|---|---|---|---|---|
| 26,7 | 25,8 | 24,0 | 24,9 | 26,4 | 25,9 |
| 24,4 | 21,7 | 24,1 | 25,9 | 26,3 | 27,1 |

**1a)** Calculez un intervalle de confiance avec coefficient de confiance à 95% pour la moyenne de la population d'eau de laquelle provient cet échantillon.
**L'intervalle de confiance calculé est : [24,30 26,24]**

Précisez votre notation et les hypothèses de la procédure statistique employée.
**X = turbidité distribuée $N(\mu,\sigma^2)$ $\sigma$ inconnu**

**1b)** Après le calcul de l'intervalle de confiance en **1a),** on prend une observation additionnelle de turbidité d'eau et on obtient 28,5. Cette valeur est située à l'extérieur de l'intervalle de confiance calculé en **1a)**. Cette constatation semble une contradiction. Est-ce le cas? Expliquez votre raisonnement. **Il n'y a pas de contradiction.**

**Il n'y a aucun rapport entre l'intervalle de confiance et l'observation d'une nouvelle valeur. Celle valeur peut se situer à l'extérieur de l'intervalle de confiance.**

**1c)** Calculez la taille échantillonnale nécessaire pour que la longueur de l'intervalle de confiance à 95% pour la moyenne ne dépasse pas $\sigma / 2$ où $\sigma$ est l'écart type de la population. **n = 62**

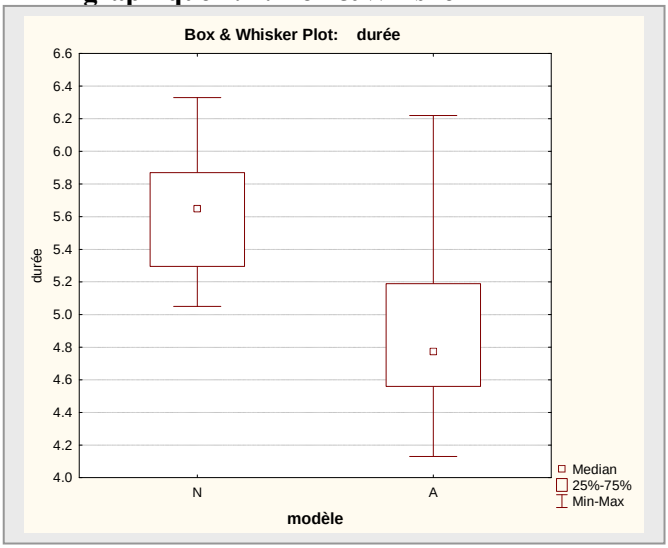
## NUMÉRO 2 (5 points)

Les données du tableau 2.1 ci-dessous sont des mesures de durée de fonctionnement (en heure), obtenues pour deux modèles N et A de pile employés dans les téléphones portables.

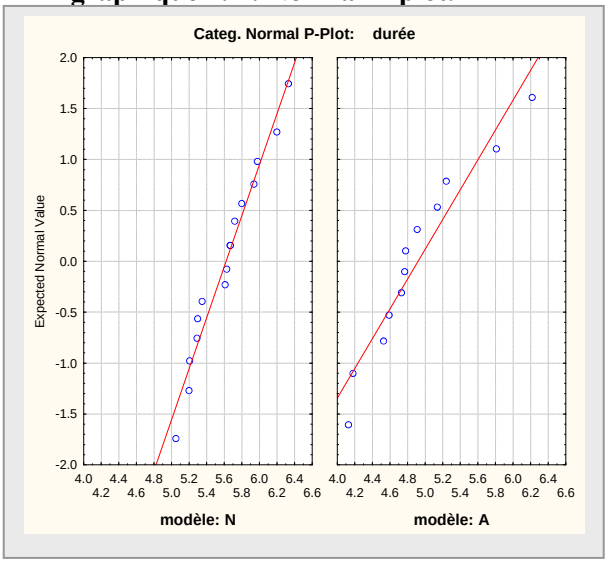
**tableau 2.1 : mesures de durée (en heures) - statistiques descriptives**

| modèle | mesures de durées (heure) | moyenne échantillonnale | écart type échantillonnal |
|---|---|---|---|
| N | 5,21 6,33 5,29 5,94 5,67 5,61 5,72 6,20<br>5,98 5,05 5,67 5,63 5,35 5,20 5,30 5,80 | 5,622 | 0,372 |
| A | 5,14 4,91 4,77 5,81 4,53 4,59 4,73 4,13<br>6,22 4,18 4,78 5,24 | 4,919 | 0,614 |

**graphique 2.1 : Box &Whisker**




**graphique 2.2 : Normal P-plot.**




**vos réponses aux questions 2a / 2b / 2c / 2d doivent contenir les informations suivantes**

- *la définition des variables, la notation employée, les hypothèses de la procédure statistique;*
- *la formulation de l'hypothèse nulle et de la contre hypothèse (alternative)*
- *l'emploi d'un seuil critique (niveau) de 5 % pour un test d'hypothèse;*
- *l'utilisation d'un coefficient de confiance de 95% pour un intervalle de confiance.*

**2a)** Pour chacun des deux modèles de pile, l'hypothèse d'une distribution normale des mesures de durée vous parait-elle plausible? Justifiez votre réponse.

**L'examen du graphique normal P-plot supporte l'hypothèse de la normalité.**

**2b)** Pour chacun des deux modèles de pile, donner une estimation de l'écart type σ des mesures de durée si on suppose que les 2 populations (N et A) ont le même écart type σ. **$s_p$=0,699**

**2c)** Peut-on affirmer que le modèle N a en moyenne une durée de fonctionnement supérieure à 5,5 heures?

**$H_0$ : $\mu = \mu_0 = 5,5$ $H_1$ : $\mu >= 5,5$ ; avec $\alpha = 0,05$**

**$t_0 = 1,29 < t_{15}(0,05) = 1,753$ $H_0$ n'est pas rejetée. Non, on ne peut pas affirmer …**

**2d)** Le modèle N est nouveau et le fabricant aimerait (avant sa commercialisation) comparer sa durée de fonctionnement à celle du modèle concurrent (modèle A) actuellement sur le marché. Peut-on conclure que le modèle N a une durée de fonctionnement supérieure à celle du modèle A?

**$H_0$ : $\mu_1 = \mu_2$** **X1 : durée pile N distribuée $N(\mu_1, \sigma^2)$**
**$H_1$ : $\mu_1 > \mu_2$** **X2 : durée pile A distribuée $N(\mu_2, \sigma^2)$**

**$t_0 = 3{,}763 > t_{26}(0{,}05) = 1{,}706$** **$H_0$ est rejetée**
**le modèle N a une durée moyenne plus grande que la durée moyenne de A**

## NUMÉRO 3 (6 points)

Les données suivantes ont été obtenues au cours d'une étude visant à évaluer le niveau de pollution résultant du transfert de carburant des réservoirs souterrains dans des camions citernes. Les variables mesurées :

- X1 : température du camion citerne
- X2: température du carburant du réservoir souterrain
- X3: pression dans du camion citerne
- X4: pression dans le réservoir souterrain
- Y : quantité d'hydrocarbures libérée dans l'atmosphère

Les données sont présentées dans le tableau 3.1 et les statistiques descriptives dans le tableau 3.2

**tableau 3.1 : données**

| | X1 | X2 | X3 | X4 | Y |
|---|---|---|---|---|---|
| 1 | 33 | 53 | 3,32 | 3,42 | 29 |
| 2 | 36 | 54 | 3,20 | 3,41 | 27 |
| 3 | 59 | 60 | 4,60 | 4,41 | 32 |
| 4 | 60 | 62 | 4,31 | 4,42 | 34 |
| 5 | 90 | 64 | 7,32 | 6,70 | 40 |
| 6 | 61 | 62 | 3,91 | 4,08 | 29 |
| 7 | 63 | 62 | 4,30 | 4,30 | 37 |
| 8 | 59 | 62 | 4,39 | 4,53 | 34 |
| 9 | 31 | 36 | 3,10 | 3,26 | 24 |
| 10 | 35 | 35 | 3,03 | 3,03 | 21 |
| 11 | 60 | 60 | 4,53 | 4,53 | 34 |
| 12 | 60 | 36 | 4,27 | 3,94 | 23 |
| 13 | 90 | 60 | 7,32 | 7,20 | 46 |
| 14 | 59 | 42 | 3,75 | 3,45 | 22 |
| 15 | 60 | 61 | 4,02 | 4,10 | 37 |
| 16 | 37 | 35 | 2,75 | 2,64 | 19 |

**tableau 3.2 : description des variables**

| | moyenne | minimum | maximum | écart type |
|---|---|---|---|---|
| X1 | 55,81 | 31,00 | 90,00 | 17,95 |
| X2 | 52,75 | 35,00 | 64,00 | 11,56 |
| X3 | 4,26 | 2,75 | 7,32 | 1,33 |
| X4 | 4,21 | 2,64 | 7,20 | 1,21 |
| Y | 30,50 | 19,00 | 46,00 | 7,59 |

On propose d'établir un lien entre les émissions d'hydrocarbures Y et les variables X1, X2, X3, X4 en utilisant des modèles de régression. Le premier modèle proposé est:

**modèle 1 :** $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \varepsilon$ où $\varepsilon \sim N(0, \sigma_1^2)$

Les tableaux 3.3 et 3.4 suivants présentent des résultats de l'ajustement du modèle 1

**tableau 3.3 : ajustement du modèle 1**
**régression variable dépendante Y: $R^2 = 0{,}931$ $R^2$ ajusté = 0,905**

| | estimation des coefficients β (beta) | écart type de l'estimation | statistique de Student t(11) | p-level |
|---|---|---|---|---|
| Intercepte | - 1,205 | 2,878 | - 0,419 | 0,6835 |
| X1 | - 0,016 | 0,103 | - 0,159 | 0,8764 |
| X2 | 0,251 | 0,089 | 2,818 | 0,0167 |
| X3 | - 5,827 | 4,390 | - 1,327 | 0,2113 |
| X4 | 10,492 | 4,450 | 2,358 | 0,0380 |

**tableau 3.4 : analyse de la variance du modèle 1**

| Source variation | Somme carrés (SS) | degrés de liberté (df) | Carrés moyens (MS) | F | p-level |
|---|---|---|---|---|---|
| Régression | 804,04 | ? | ? | 36,88 | 0,000003 |
| Résiduelle | ? | ? | ? | | |
| Totale | 864,00 | ? | | | |

**3a)** Déterminer les six (**6**) valeurs manquantes (représentées par **?**) dans le tableau 3.4. Placer vos réponses directement dans le tableau.

**3b)** Pourquoi peut-on conclure que les variables X1et X3 ne contribuent pas significativement au modèle? Expliquez votre réponse.

**3c)** Sur la base d'un intervalle de confiance à 95 % pour le coefficient $\beta_2$ peut-on dire que ce coefficient est significativement différent de 0? Expliquez votre réponse.

Le deuxième modèle proposé est :

**modèle 2 :** $Y = \gamma_0 + \gamma_2 X2 + \gamma_4 X4 + \varepsilon$ où $\varepsilon \sim N(0, \sigma_2^2)$

Les tableaux 3.5 et 3.6 présentent les résultats de l'ajustement du modèle 2

**tableau 3.5 : ajustement du modèle 2**
**régression variable dépendante Y: $R^2$=0,911 $R^2$ ajusté = 0,897**

| | estimation des coefficients γ (gamma) | écart type de l'estimation | statistique de Student t(13) | p-level |
|---|---|---|---|---|
| Intercepte | - 1.334 | 2,973 | - 0,449 | 0,6610 |
| X2 | 0,320 | 0,070 | 4,549 | 0,0005 |
| X4 | 3,551 | 0,669 | 5,308 | 0,0001 |

**tableau 3.6 : analyse de la variance du modèle 2**

| Source variation | Somme Carrés (SS) | degrés de liberté (df) | Carrés moyens (MS) | F | p-level |
|---|---|---|---|---|---|
| Régression | 768,83 | 2 | 393,42 | 66,28 | 0,000000 |
| Résiduelle | 77,17 | 13 | 5,94 | | |
| Total | 864,00 | | | | |

**3d)** Faites un choix entre le modèle 1 et le modèle 2. Justifier votre choix et préciser vos critères.

**3e)** En utilisant le modèle 2, pour quelle valeur de X2 et quelle valeur de X4 a-t-on une émission d'hydrocarbures minimale si on contraint les variables à l'intérieur de l'espace de variation observé dans le tableau 3.2.

**3f)** On a calculé un **intervalle de confiance** à 95% pour la quantité moyenne d'hydrocarbures émise lorsque X2 = 35 et X4 = 2,64. On a aussi calculé un **intervalle de prédiction** à 95% pour la quantité d'hydrocarbure émise pour les mêmes valeurs de X2 et X4.
Les 2 intervalles calculés sont présentés ici dans un ordre quelconque :
intervalle A **:** 13,39 à 25,07 intervalle B : 16,70 à 21,77
Lequel est l'intervalle de confiance? Lequel est l'intervalle de prédiction?
Expliquez pourquoi les 2 intervalles ont des longueurs différentes.

## RÉPONSES – NUMÉRO 3

**3a)**

**tableau 3.4 : analyse de la variance du modèle 1**

| Source variation | Somme carrés (SS) | degrés de liberté (df) | Carrés moyens (MS) | F | p-level |
|---|---|---|---|---|---|
| Régression | 804,04 | ? 4 | ? 201,01 | 36,88 | 0,000003 |
| Résiduelle | ? 59,96 | ? 11 | ? 5,45 | | |
| Totale | 864,00 | ? 15 | | | |

**3b)** **Les variables X1 et X3 ne contribuent pas significativement au modèle car le test de $H_0 : \beta_1 = 0$ et celui de $H_0 : \beta_3 = 0$ ne sont pas rejetés**

**les p-level sont respectivement 0,876 et 0,211. Le résultat est identique avec des tests T.**

**3c)** **l'intervalle de confiance à 95% pour $\beta_2$ est : $0{,}055 < \beta_2 < 0{,}446$.**

**L'intervalle ne contient pas 0 : donc la valeur de $\beta_2$ est significativement différent de 0**

**3d)**

| modèle | $R^2$ | $R^2$ ajust |
|---|---|---|
| 1 | 0,931 | 0,905 |
| 2 | 0,911 | 0,897 |

**Les deux modèles ont des coefficients de détermination ajustés quasi identiques.**

**Le modèle 2 est préférable au modèle 1 car ses 2 termes sont significatifs.**

**Ce n'est pas le cas du modèle 1 qui contient 2 termes non significatifs.**

**De plus, on cherche toujours à proposer le modèle le plus simple qui explique les données.**

**3e)** **Puisque les coefficients β du modèle 2 sont positifs, Y sera minimale pour les valeurs minimales de X2 et de X4. Selon les données, $X_{02} = 35$ et $X_{04} = 2{,}64$ valeur minimale de Y est Y = -1,334 + 0,320*35 + 3,551*2,64 = 19,23**

**3f)** **Selon les formules développées en régression – voir HMGB p. 403 et 404 équations 14.23 / 14.25, l'intervalle de confiance pour la quantité moyenne**

**d'hydrocarbures est le plus petit des 2 intervalles. C'est l'intervalle B variant de 16,70 à 21,77**

**L'intervalle de prédiction pour la quantité d'hydrocarbure est le plus grand des 2 intervalles. C'est l'intervalle A variant de 13,39 à 25,07.**

**L'intervalle de prédiction est plus grand car on doit ajouter à la valeur prédite $Y_0$ l'incertitude du terme d'erreur dans le modèle.**

## NUMÉRO 4 (5 points)

**4a)** Dans toute expérience planifiée, il y a une variable de réponse (ou dépendante) Y et des facteurs (variables) contrôlés $X_1$, $X_2$, …, Xk. On fait varier les facteurs X à des valeurs choisies selon un plan expérimental (série d'essais ou de tests) défini par l'expérimentateur. On mesure la réponse Y à chacun de ces essais. De plus, il y a toujours la présence de l'erreur aléatoire (ou expérimentale) mais celle-ci n'est pas mesurée et elle n'est pas contrôlée.
Précisez une méthode qui permet de détecter la présence de l'erreur expérimentale dans la conduite d'une expérience.

**4b)** Lequel de ces énoncés représente le mieux le concept de l'erreur expérimentale :

**a)** il y a une possibilité de faire une erreur de manipulation dans l'exécution des essais (tests);
**b)** les appareils de mesure ne donnent pas toujours des valeurs justes et précises;
**c)** il faut tenir compte de l'erreur humaine qui est toujours possible dans un test;
**d)** il est difficile de tout contrôler dans la conduite d'une expérience;
**e)** autre chose (à définir) que les énoncés a - b - c –d.

Si vous choisissez **a)** ou **b)** ou **c)** ou **d)** : expliquer votre choix.
Si votre choix est **e)**, proposer votre définition de l'erreur expérimentale. (5 lignes maximum)

**4c)** On étudie l'influence de 3 facteurs A, B, C sur une variable de réponse Y.
On fait varier les facteurs A, B et C à 2 modalités (valeurs) seulement.
Le plan d'expérience est un plan factoriel complet.
Le plan est exécuté une première fois.
Par la suite, le plan est exécuté deux autres fois.
L'analyse des données est basée sur le modèle suivant:

$$\mathbf{Y = \beta_0 + \beta_1 * XA + \beta_2 * XB + \beta_3 * XC + \beta_4 * XA * XB + \beta_5 * XA * XC + \beta_6 * XB * XC + \varepsilon \quad (1)}$$

où XA, XB, XC représentent les variables de codage associés aux facteur A, B, C;
les $\beta$ sont les coefficients du modèle de régression;
$\varepsilon$ est l'erreur aléatoire; on suppose que $\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$ (distribution normale)

La moyenne de tous les essais réalisés est de 100.
**Compléter les valeurs manquantes** (**notées ?**) dans le tableau d'analyse de la variance.

**Analyse de la variance du modèle (1)**

| Source variation | Somme de carrés SS | Degrés de Liberté df | Carrés Moyens MS | F | Effet significatif? au seuil de 0,05 réponse : oui ou non |
|---|---|---|---|---|---|
| A | 300 | 1 | ? 300 | ? | ? |
| B | 200 | 1 | ? 200 | ? | ? |
| C | 50 | 1 | ? 50 | ? | ? |
| AB | 150 | ? | ? | ? | ? |
| AC | 130 | ? | ? | ? | ? |
| BC | 20 | ? | ? | ? | ? |
| Résiduelle (erreur) | ? | ? | ? | | |
| totale | 1000 | ? | | | |

## RÉPONSES – NUMÉRO 4

**4a) En faisant des répétitions avec des valeurs fixées des variables $X_1, X_2, \ldots, X_k$**

**4b) Réponse = e**

**L'erreur expérimentale est générée par la somme de toutes les sources de variation inconnues et non contrôlées qui affectent la réponse autres que les facteurs connus et contrôlés $X_1, X_2, \ldots, X$**

**4c)**

| Source variation | Somme de Carrés SS | Degrés de Liberté df | Carrés Moyens MS | F | Effet significatif? au seuil de 0,05 réponse : oui ou non |
|---|---|---|---|---|---|
| A | 300 | 1 | ? 300 | ? 34,01 | ? oui |
| B | 200 | 1 | ? 200 | ? 22,67 | ? oui |
| C | 50 | 1 | ? 50 | ? 5,67 | ? oui |
| AB | 150 | ? 1 | ? 150 | ? 17,01 | ? oui |
| AC | 130 | ? 1 | ? 130 | ? 14,74 | ? oui |
| BC | 20 | ? 1 | ? 20 | ? 2,67 | ? non |
| Résiduelle (erreur) | ? 150 | ? 17 | ? 8,82 | | |
| totale | 1000 | ? 23 | | | |