POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

MTH2302D - PROBABILITÉS ET STATISTIQUE

Devoir - Hiver 2021

Date de remise : 20 avril avant 18h00 (dans Moodle)

DIRECTIVES :

- ✓ Vous devez remettre un rapport individuel au plus tard lundi le **20 avril avant 18h00**, dans Moodle, sous la forme d'un fichier électronique de format PDF, nommé **`matricule.pdf`**. Le rapport doit contenir votre nom, prénom, matricule et toutes les informations requises sur la page de présentation dont un modèle est disponible sur le site du cours. Aucune remise papier ou par courriel ne sera acceptée. La note zéro sera attribuée à toute remise qui ne respecte pas les directives.
- ✓ Pour le **problème I**, vous devez d'abord obtenir avec votre matricule les valeurs numériques des lettres A, B, C et D avant de donner une réponse numérique à la question. **Veuillez consulter les instructions** (page de présentation).
- ✓ Pour le **problème II**, vous devez d'abord obtenir avec votre matricule un ensemble personnalisé de données. Toutes vos réponses doivent correspondre à votre ensemble de données. **Veuillez consulter les instructions** de la procédure sur la page de présentation.
- ✓ Chacune de vos réponses doit être **complète**, **expliquée** et **justifiée**. Lors de la correction, il sera tenu compte de la qualité de la présentation, la pertinence des analyses et l'initiative dont vous ferez preuve dans votre rapport. Sur les **40** points, **38** sont alloués aux analyses, commentaires pertinents, etc. et **2** à la présentation.
- ✓ Les analyses statistiques, les tableaux et les graphiques du rapport doivent être produits avec le logiciel R.
- ✓ Lorsque nécessaire et selon le contexte, utiliser un seuil critique de **5 %**, ou un niveau de confiance de **95 %**.
- ✓ Tout cas de plagiat sera sévèrement sanctionné (Comité d'Examen des Fraudes), ainsi qu'une note **F** à ce cours.

PROBLÈME I (10 points). Un magasin propose un nouveau produit saisonnier en vedette. Soit N la variable aléatoire qui désigne le nombre de clients qui se présentent au magasin durant la saison, où $N \sim \text{Poi}(A)$. On estime que la probabilité qu'un client achète ce nouveau produit est de B et ce indépendamment d'un client à l'autre.

a) (3 points) On suppose ici que le magasin dispose d'un stock illimité de ce produit. Soient les variables aléatoires X et Y telles que $X =$ le nombre de clients qui achètent le produit ; $Y =$ le nombre de clients qui n'achètent pas le produit. Les variables X et Y sont-elles indépendantes ? Justifier.

b) (7 points) Le magasin a un profit de C \$ pour chaque unité vendue. Chaque unité non vendue devrait être stockée pour l'année prochaine au coût de D \$.

Déterminer la valeur du nombre d'unités stockées n que le magasin devrait avoir pour maximiser son profit moyen.

PROBLÈME II (28 points). Ce problème est une étude de cas qui consiste en une analyse de données tirées du magazine *Motor Trend* de 1975. Elles avaient été recueillies au terme d'une étude portant sur l'efficacité énergétique de divers modèles et marques de véhicules, particulièrement leur consommation d'essence.

Les données. Les données à analyser sont constituées d'un échantillon de **160** observations sur cinq variables mesurant un certains nombre de caractéristiques des véhicules de l'étude. Le Tableau 1 ci-dessous présente les différentes variables de l'étude (numéro de colonne dans le fichier, symbole, nom, et description). Les symboles et les numéros de colonnes sont tels qu'ils apparaissent dans votre ensemble de données personnalisées.

Col. nº	Symbole (Nom dans le fichier)	Description
1	Y (mpg)	La consommation du véhicule (en milles par gallon)
2	X_1 (horsepower)	La puissance du moteur du véhicule (en livres par pied)
3	X_2 (weight)	Le poids du véhicule (en livres)
4	X_3 (origin)	Le code du pays d'origine du véhicule (**1** : États-Unis et **0** : Autres pays)

Tableau 1 : Les variables de l'analyse.

Le but visé est d'analyser ces données afin de déterminer les liens possibles entre différentes variables, et de déterminer un modèle statistique permettant de décrire et de prédire la consommation d'un véhicule à partir de la caractéristique la plus pertinente.

***Phase 1* : Analyse statistique descriptive et inférence**.

On demande de répondre aux questions suivantes en utilisant des techniques appropriées de statistique (statistique descriptive et inférence), illustrées par des diagrammes pertinents.

a) (4 points) Pour la variable *consommation* (mpg), produisez les graphiques et les tableaux demandés et interprétez brièvement le résultat dans chaque cas :

- un histogramme et un diagramme de Tukey (ou «Box Plot») ;
- une droite de Henry (ou «Normal Probability Plot») et un test de normalité (Shapiro-Wilk) ;
- un tableau de statistiques descriptives comprenant : *moyenne, quartiles, écart type, intervalle de confiance pour la moyenne*.

b) (7 points) On veut vérifier si la consommation d'un véhicule est significativement différente selon le pays d'origine. Pour cela on peut considérer la variable *consommation* (mpg) divisée en deux groupes selon le code du pays d'origine (origin) et effectuer une comparaison des deux groupes en termes de moyenne, symétrie et variabilité. Pour ce faire, effectuez les analyses suivantes et donnez une brève conclusion :

- deux histogrammes juxtaposés, et deux diagrammes de Tukey (ou «Box Plot») juxtaposés ;
- un tableau des statistiques descriptives par groupe : *moyenne, quartiles, écart type, intervalle de confiance pour la moyenne* ;
- un test d'hypothèse sur l'égalité des variances pour les deux groupes ;
- un test d'hypothèse sur l'égalité des moyennes pour les deux groupes.

***Phase 2* : Recherche du meilleur modèle**.

On s'intéresse dans cette phase à la détermination d'un modèle permettant d'expliquer la consommation en fonction des différents facteurs considérés. Pour ce faire, on envisage des modèles de régression simple en considérant la consommation (mpg) comme variable dépendante Y.

c) (12 points) On considère les six modèles suivants :

Modèle 1 : $Y = \beta_0 + \beta_1 X_1 + \varepsilon$; **Modèle 2 :** $Y = \beta_0 X_1^{\beta_1} e^{\varepsilon}$; **Modèle 3 :** $Y = \beta_0 e^{\beta_1 X_1 + \varepsilon}$;

Modèle 4 : $Y = \beta_0 + \beta_1 X_2 + \varepsilon$; **Modèle 5 :** $Y = \beta_0 X_2^{\beta_1} e^{\varepsilon}$; **Modèle 6 :** $Y = \beta_0 e^{\beta_1 X_2 + \varepsilon}$,

où β_0 et β_1 sont des paramètres et ε une erreur aléatoire.

Remarque : Les coefficients β_0 et β_1 ainsi que l'erreur ε ne sont pas les mêmes d'un modèle à l'autre.

Pour chacun des six modèles ci-dessus :

- *(5 points)* Effectuez l'ajustement (i.e. obtenir le tableau des coefficients de régression, le tableau d'analyse de la variance).
- *(5 points)* Tester la signification du modèle et effectuez une analyse des résidus (normalité, homoscédasticité, points atypiques, etc.)
- *(2 points)* En conclusion : effectuez une comparaison et dire lequel des six modèles est préférable aux autres. Justifiez votre choix en précisant les critères utilisés.

d) (5 points) Sur la base du meilleur modèle que vous avez obtenu en c), calculez un intervalle de prévision pour la consommation (mpg) d'un véhicule ayant les caractéristiques suivantes : puissance ($X_1 = 120$), poids ($X_3 = 2200$). Commentez brièvement le résultat.