IFMA/MPE **Statistique de base**

Examen

Jeudi 20 octobre 11h-12h30

Exercice 1. *On considère un échantillon $X = (X_1, \dots, X_n)$ de n variables indépendantes et distribuées selon la loi exponentielle $Exp(\theta)$ de paramètre θ, où $\theta > 0$ est inconnu. La variable X_1 a donc pour densité*

$$f_\theta(x_1) = \theta \exp(-\theta x_1) \mathbf{1}_{\{x_1>0\}}.$$

1. *Donner la vraisemblance $f_\theta(x_1, \dots, x_n)$ de ce modèle.*
2. *Donner l'estimateur $\hat{\theta}_n$ du maximum de vraisemblance de θ.*
3. *On rappelle que la somme de n variables i.i.d. de loi exponentielle $Exp(1)$ de paramètre 1 suit une loi Gamma de paramètre n, de densité $\frac{1}{(n-1)!} x^{n-1} e^{-x} \mathbf{1}_{\{x>0\}}$. Soit Y_n une variable aléatoire suivant une loi Gamma de paramètre n. Montrer que $\sum_{i=1}^n X_i$ a même loi que Y_n/θ.*
4. *Calculer le biais de l'estimateur $\hat{\theta}_n$.*
5. *Est-ce que $\hat{\theta}_n$ est fortement consistant?*
6. *On notera $\gamma_n(t)$ le quantile de la loi Gamma de paramètre n défini par*

$$P(Y_n \leq \gamma_n(t)) = t.$$

   Donner, à n fixé, un intervalle de confiance de niveau $1 - \alpha$ sur θ sous la forme $[a, +\infty[$.

7. *Soit $\theta_0 > 0$ un réel fixé. En déduire un test de $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta > \theta_0$ de risque α.*
8. *Justifier que ce test est uniformément plus puissant parmi tous les tests de risque inférieur à α.*
9. *Déterminer la puissance de ce test en fonction de la loi de Y_n. Montrer que la puissance est croissante en θ et tend vers 1 quand $\theta \to +\infty$.*

Exercice 2. *Pour chaque test, on posera précisément le modèle, les hypothèses faites sur le modèle, les hypothèses à tester, la statistique de test, sa loi, et la règle le décision.*

1. *On veut comparer deux procédés de mesure d'un élément chimique. On dispose de M solutions différentes (de concentration inconnue, et non nécessairement identique, notées $c_1, c_2, \ldots, c_M$). Pour chacune d'elles, on fait une mesure avec la méthode I, et une mesure avec la méthode II. On veut savoir si les deux procédes de mesure donnent le même résultat en moyenne. Quel test effectuer?*

2. *A l'approche d'un réferendum, on effectue un sondage sur les intentions de vote de n individus pris au hasard dans la population. On note $X_i = 1$ si l'individu i pense voter OUI et 0 sinon. On note θ la proportion de OUI dans la population entière et $\hat{\theta}_n$ la proportion de OUI dans notre échantillon. Faire un test asymptotique de risque α visant à savoir si le OUI va l'emporter.*

3. *On recueille les données d'une maternité en France.*

   (a) *On veut savoir si le nombre de naissances varie de façon significative selon le mois de l'année. Construire le test adéquat.*

   (b) *On aimerait savoir s'il y a un lien entre le mois de naissance et le mode d'accouchement (par césarienne ou sans césarienne). Construire le test correspondant.*

Exercice 3. *Nous voulons expliquer la concentration de l'ozone sur Rennes en fonction des variables* `T9, T12, Ne9, Ne12, Vx` *(température à 9h,12h, nébulosité à 9h,12h, vitesse du vent). Les sorties données par* `R` *sont:*

```
Call:
lm(formula = ozone ~ T9 + T12  + Ne9 + Ne12  +   Vx )

Coefficients:
                 Estimate    Std. Error  t value    Pr(>|t|)
(Intercept)     12.24442     13.47190     0.909       0.3656
T9               0.35198     0.06289        ???     1.88e-07 ***
T12             -2.18909     0.93824     -2.333      0.0216 *
Ne9              2.22115     1.43294      1.550      0.1243
Ne12            -0.42102     1.36766     -0.308      0.7588
Vx               0.84455     0.15247      5.3247       2e-08***


Residual standard error: 14.36 on 101 degrees of freedom
Multiple R-squared:  0.7638,        Adjusted R-squared:  0.7405
F-statistic: ??? on ??? and ??? DF,  p-value: < 2.2e-16
```

1. *Ecrire le modèle de la régression en précisant les hypothèses du modèle.*

2. *Dire comment trouver les élements cachés par* `???` *à partir des autres données de la sortie (ne pas faire les calculs numériques). On notera $R^2 (= 0.7638)$ le R^2 de la régression.*

3. *Rappeler le test pour savoir si le coefficient de la variable T9 est significativement non nul au seuil de 5%. Lire la sortie et conclure.*

4. *Donner l'expression, en fonction des données de la sortie R et de quantiles bien choisis, d'un intervalle de confiance au niveau 95% sur le coefficient de la variable T12.*

5. *La régression est-elle significative?*

6. *Nous souhaitons tester la nullité simultanée des paramètres correspondant aux variables* `Ne9, Ne12`. *Proposer un test et exprimer la statistique de test en fonction de R^2 et $\tilde{R}^2$, représentant le R^2 de la régression de* `ozone` *respectivement contre toutes les variables, et contre les variables* `T9, T12, Vx`.

Exercice 4. *Soit $X = (X_1, \dots, X_{N_1})$ un échantillon de N_1 variables i.i.d. de loi $\mathcal{N}(\mu_X, \sigma^2)$ et $Y = (Y_1, \dots, Y_{N_2})$ un échantillon de N_2 variables i.i.d. de loi $\mathcal{N}(\mu_Y, \sigma^2)$. Les paramètres μ_X, μ_Y, σ^2 sont inconnus. On fera bien attention au fait que les variances de X_1 et Y_1 sont supposées égales.*

On note $Z = (Z_1, \dots, Z_{N_1+N_2})'$ le vecteur de $\mathbb{R}^{N_1+N_2}$ de coordonnées $Z_k = X_k$ si $k \leq N_1$ et $Z_k = Y_{k-N_1}$ si $N_1 + 1 \leq k \leq N_1 + N_2$. Soit le vecteur $V_1 = (V_{1,1}, \dots, V_{1,N_1+N_2})'$ de coordonnées $V_{1,k} = \mathbf{1}_{\{k \leq N_1\}}$ et le vecteur $V_2 = (V_{2,1}, \dots, V_{2,N_1+N_2})'$ de coordonnées $V_{2,k} = \mathbf{1}_{\{k > N_1\}}$. Ainsi,

$$Z = \begin{pmatrix} X_1 \\ \dots \\ X_{N_1} \\ Y_1 \\ \dots \\ Y_{N_2} \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 1 \\ \dots \\ 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0 \\ \dots \\ 0 \\ 1 \\ \dots \\ 1 \end{pmatrix}.$$

1. *Soit $\mathcal{V}$ l'espace vectoriel engendré par les vecteurs V_1 et V_2. Donner la projection orthogonale notée $Z_{\mathcal{V}}$ de Z sur $\mathcal{V}$ et celle notée $Z_{\mathcal{V}^\perp}$ de Z sur $\mathcal{V}^\perp$.*

2. *Donner les vecteurs moyenne $E[Z_{\mathcal{V}}]$ et $E[Z_{\mathcal{V}^\perp}]$.*

3. *Donner la loi des variables aléatoires*
$$\frac{1}{\sigma^2}||Z_{\mathcal{V}} - (\mu_X V_1 + \mu_Y V_2)||^2 \qquad , \qquad \frac{1}{\sigma^2}||Z_{\mathcal{V}^\perp}||^2.$$

4. *On note $\bar{X} = \frac{1}{N_1}\sum_{i=1}^{N_1} X_i$. Montrer que $\bar{X}$ est indépendant de $Z_{\mathcal{V}^\perp}$.*

5. *En déduire la loi de*
$$\frac{\bar{X} - \mu_X}{\hat{\sigma}/\sqrt{N_1}}$$
où $\hat{\sigma}^2 = \frac{1}{N_1+N_2-2}||Z_{\mathcal{V}^\perp}||^2$.

6. *Soit μ_0 un réel. En déduire un test possible de $(H_0) : \mu_X = \mu_0$ contre $(H_1) : \mu_X \neq \mu_0$.*