Probabilités et statistiques pour l'ingénieur
M1 Informatique

Examen, 4 janvier 2016

Ce devoir surveillé consiste en **4 exercices** sur un total de 4 pages.
Les calculatrices ne sont pas autorisées.
Les exercices peuvent être traités dans le désordre.
La notation prendra en compte le **soin et la clarté des réponses**.

Exercice 1.

On considère une pièce de monnaie C et deux dés à six faces A et B. On suppose que

- C est non biaisée : la probabilité que C tombe sur pile et la même que celle qu'elle tombe sur face,
- A est non biaisé : les faces ont toutes la même probabilité,
- B est biaisé : les valeurs 4 et 6 a été remplacée par la valeur 1.

On considère le jeu suivant : on lance C, et après on lance A si on a obtenu pile et on lance B si on a obtenu face. On gagne si on obtient un nombre pair.

1. Calculer la probabilité p de victoire.
2. Calculer la probabilité conditionnelle d'avoir obtenu face sachant qu'on a gagné.
3. On répète le jeu n fois et on considère la variable aléatoire X_n qui compte le nombre de victoires sur les n répétitions. Quelle est la distribution exacte de X_n ? Donner la formule en fonction de n et p.
4. Quel est le nombre attendu de victoires sur $n = 10$ répétitions ?
5. On considère la fréquence empirique de victoire $\frac{X_n}{n}$. Quelle est la limite de $\frac{X_n}{n}$ quand $n \to \infty$? Justifier à l'aide d'un important théorème vu en cours.

Exercice 2.

1. On veut vérifier si un dé à 6 faces est biaisé ou pas. Quelle expérience mèneriez-vous ? Proposez un test statistique pour cela : donner les hypothèses H_0, H_1 et le nom du test.

2. On s'intéresse à savoir si il y a une dépendance entre le revenu X et le niveau d'étude Y d'une personne. Pour cela on considère trois catégories pour X (faible, moyen, élevé) et cinq catégories pour Y (pré-baccalauréat, baccalauréat, licence, master, doctorat). Proposer un test pour étudier l'association entre X et Y : donner les hypothèses et le nom du test.
3. On veut vérifier si il y a une dépendance entre l'âge X et la pression artérielle Y. En particulier on se demande si la pression artérielle augmente avec l'âge. Proposer un test pour cela : donner les hypothèses, le nom du test et sa latéralité.

Exercice 3.

Chaque jour à Paris cinq millions de voyageurs empruntent les transports en commun, dont un certain nombre sans titre de transport. A la suite de contrôles sur un total de n voyageurs en un jour donné, on a sanctionné s fraudeurs. Un estimateur ponctuel de la proportion p de fraudeurs est $\hat{p} = \frac{s}{n}$.

1. Montrer que $\hat{p}$ est un estimateur sans biais de p, c'est à dire que $E[\hat{p}] = p$, où $E[\cdot]$ note l'espérance d'une variable aléatoire.
2. Dans un échantillons de $n = 10^4$ voyageurs, on a compté $s = 1000$ fraudeurs. Donner l'intervalle de confiance symétrique au seuil 95% pour p calculé sur la base de l'échantillon. Indication : on rappelle que

$$\mathbb{P}\left(\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

3. Est-ce que le résultat des contrôles effectués est d'une quelconque utilité pour savoir si le la proportion de fraudeurs est supérieur à 10% ? Justifier la réponse.
4. Donner le nombre minimal de voyageurs qu'il aurait fallu contrôler pour être au 95% certains que l'erreur d'estimation était inférieur à $\pm 0.4\%$, c'est à dire $|\hat{p} - p| \leq 0.004$.

Exercice 4.

On s'intéresse à la relation entre deux variables aléatoires X et Y. Pour cela on a relevé 50 observations (x, y) : voir Fig. 1.

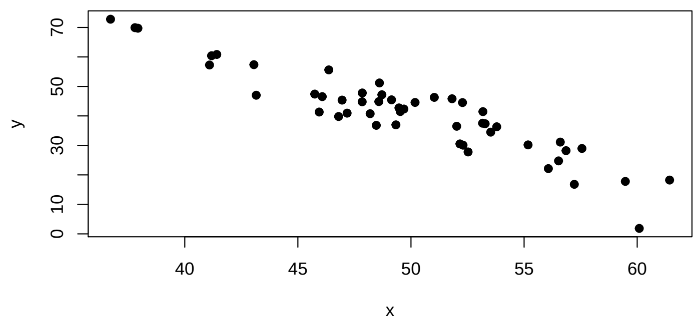

FIGURE 1 – Paires (x, y).

On modélise Y en fonction de X à l'aide d'un modèle de régression linéaire.

1. Ecrire l'équation du modèle en précisant ses hypothèses.

On estime les paramètres du modèle à l'aide du logiciel R et on obtient le listing de la Fig. 2. Au vu de ces résultats :

2. Donner l'estimation des paramètres du modèles. Peut-on conclure que les paramètres sont différents de zéro ? Pourquoi ?
3. Ecrire la droite des moindres carrés.
4. Ecrire la valeur attendu de Y sachant $X = 50$.
5. Donner et commenter une mesure de la *qualité* du modèle.
6. Donner une estimation (approximative) du coefficient de corrélation entre X et Y.

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-16.196  -3.666   1.099   3.608   9.310

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 152.0841     6.7749   22.45   <2e-16 ***
x            -2.2304     0.1351  -16.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.445 on 48 degrees of freedom
Multiple R-squared:  0.8503,       Adjusted R-squared:  0.8471
F-statistic: 272.6 on 1 and 48 DF,  p-value: < 2.2e-16
```

Figure 2 – Régression linéaire modélisant Y en fonction de X, sortie R