Principes et Méthodes Statistiques

Durée : 3 heures. Tous documents autorisés.

Les résultats vus en cours ou en TD peuvent être utilisés sans être redémontrés.

Il sera grandement tenu compte de la qualité de la rédaction (présentation et justification des réponses) dans la notation.

Le sujet comporte trois parties indépendantes et quatre pages.

Barème indicatif - Partie 1 : 4 pts, Partie 2 : 4 pts, Partie 3 : 13,5 pts.

Première partie

Deux sondages réalisés à 15 jours d'intervalle donnent des résultats d'une variabilité étonnante sur les intentions de vote à une élection entre deux candidats. Le premier sondage, réalisé sur 1 050 personnes, indique un score pour le candidat A de 49% alors que le second sondage, réalisé sur 980 personnes, indique un score de 53% pour ce même candidat A. À l'aide des instructions exécutées sous `R` et des résultats présentés ci-dessous, vous répondrez aux deux questions suivantes en choisissant, pour chaque question, un seul test parmi les deux proposés. Pour chacune des deux questions :

- vous justifierez votre choix de test parmi les deux proposés ;
- vous décrirez brièvement le test (H_0 et H_1, expression de la zone de rejet) ;
- vous conclurez très clairement sur le test.

1. Peut-on conclure à une hausse du score du candidat A entre les deux sondages ?
2. Peut-on conclure à partir du second sondage que le candidat A sera élu (c'est-à-dire qu'il emportera 50% ou plus des voix) ?

Les intentions de vote sont codées 1 pour le candidat A et 0 pour le candidat B. Les vecteurs S1 et S2 correspondent aux 1 050 et 980 observations, respectivement pour le premier sondage et le second sondage.

```
> summary(S1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4914  1.0000  1.0000
> table(S1)
```

```
S1
  0   1
534 516
> summary(S2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5327  1.0000  1.0000
> table(S2)
S2
  0   1
458 522
> SS1<-sum(S1)
> SS2<-sum(S2)
> binom.test(SS2,n=980,p=0.4914,alternative="greater")

        Exact binomial test

data:  SS2 and 980

number of successes = 522, number of trials = 980, p-value = 0.005363
alternative hypothesis: true probability of success is greater than 0.4914
95 percent confidence interval:
 0.5058804 1.0000000
sample estimates:
probability of success
             0.5326531

> prop.test(c(SS1,SS2),c(1050,980),alternative="less", correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(SS1, SS2) out of c(1050, 980)
X-squared = 3.4476, df = 1, p-value = 0.03167
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.004738413
sample estimates:
   prop 1    prop 2
0.4914286 0.5326531
```

Deuxième partie

Les occurrences de chaque chiffre dans la suite des 1 000 021 premières décimales du nombre d'Euler e = exp(1) sont données dans le tableau suivant :

0	1	2	3	4
99 425	100 136	99 848	100 231	100 390

5	6	7	8	9
100 089	100 481	99 913	99 816	99 692

On souhaite vérifier, à l'aide d'un test, si cette suite peut être utilisée comme un générateur de chiffres aléatoires de loi uniforme sur $\{0, 1, \ldots, 9\}$. Définir les variables aléatoires X_j considérées et préciser le test utilisé ; en exprimer formellement les hypothèses nulle et alternative. Donner ensuite la région critique et un encadrement de la p-valeur. Que répondriez-vous à la question initiale ?

Troisième partie

On considère un échantillon $(X_1, \ldots, X_n)$, dont la loi a pour densité

$$f(x;\theta) = (1-\theta)^{\mathbf{1}_{[-1/2,0]}(x)}\,(1+\theta)^{\mathbf{1}_{]0,1/2]}(x)}\,\mathbb{1}_{[-1/2,1/2]}(x) = \begin{cases} 1-\theta & \text{si } x \in [-1/2, 0] \\ 1+\theta & \text{si } x \in]0, 1/2] \\ 0 & \text{sinon,} \end{cases}$$

où θ est un paramètre inconnu dans $]0, 1[$.

1. Calculer la fonction de répartition, l'espérance et la variance de cette loi.
2. Dans le script `R` suivant, expliquer ce que fait la fonction `rdistrib`. Justifier votre réponse par un calcul.

```
> rdistrib = function(n, theta) {
    u = runif(n) # simule n variables i.i.d. de loi uniforme sur [0,1]
    x = rep(0, n) # vecteur nul en dimension n
    i1 = u <= 0.5 * (1-theta)
    i2 = u > 0.5 * (1-theta)
    x[i1] = -0.5 + u[i1] /(1-theta)
    x[i2] = (u[i2] - (1-theta)/2) / (1+theta)
    x # rdistrib retourne la valeur x
    }
> theta = 0.7
> n = 1000
> x = rdistrib(n, theta)
> plot(sort(x), seq(1,n)/n) # sort trie les valeurs de x
```

3. Tracer un graphe qui vous paraisse représentatif du résultat de la commande `plot(sort(x), seq(1,n)/n)`
(expliquer).

4. Calculer l'estimateur de θ par la méthode des moments, $\tilde{\theta}_n$. Montrer qu'il est sans biais et convergent en moyenne quadratique.

5. Donnez une expression de l'estimateur de maximum de vraisemblance $\hat{\theta}_n$, en fonction de n et de $K = \sum_{j=1}^{n} \mathbb{1}_{[-\frac{1}{2};0]}(X_j)$.
Indication. On rappelle que pour tout ensemble A et toute variable aléatoire X, $\mathbb{1}_A(X)$ est une variable aléatoire qui vaut 1 si $X \in A$ et 0 sinon.

6. Montrer que l'estimateur de maximum de vraisemblance $\hat{\theta}_n$ est sans biais. Montrer ensuite que $\mathcal{I}_1(\theta) = \dfrac{1}{(1-\theta)(1+\theta)}$ et en déduire que $\hat{\theta}_n$ est efficace.
Indications. On rappelle que pour tout ensemble A et toute variable aléatoire X, $\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$. Par ailleurs, on a aussi $\mathbb{1}_A(X) \times \mathbb{1}_A(X) = \mathbb{1}_A(X)$.

7. En utilisant la propriété

$$\sqrt{n\mathcal{I}_1(\theta)}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

et en vous inspirant de la méthodologie mise en œuvre dans le cours dans le cas d'une proportion, proposez un intervalle de confiance asymptotique de seuil α pour θ.
Indication : Vous montrerez qu'avec une probabilité $1-\alpha$, θ satisfait une certaine inéquation du second degré.