	ING2-GI **EXAMEN DE STATISTIQUES INFERENTIELLES 2017-2018**
Durée : 2h	Calculatrice autorisée 4 feuilles manuscrites R/V autorisées

<u>Questions à choix multiples</u> :

Question 1 : Dans un problème de test d'hypothèses, on fixe à un niveau donné :

A) l'erreur de première espèce et l'erreur de deuxième espèce

B) la puissance

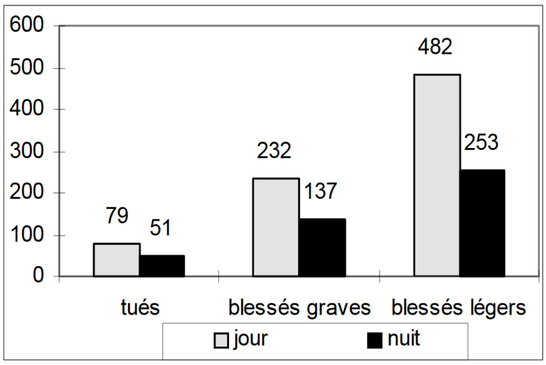
C) l'erreur de première espèce

Question 2 : Pour déterminer la région critique d'un test, il faut pouvoir :

A) déterminer une loi de probabilité sous l'hypothèse H_0

B) déterminer une loi de probabilité sous l'hypothèse H_1

Question 3 : Dans un article paru dans le journal "La dépêche" du 10 novembre 1999 et intitulé "Des chiffres qui font mal" on trouve le graphique ci-dessous qui concerne l'ensemble des personnes accidentées de la route en 1998 pour la Haute-Garonne. Quel test doit-on mettre en place pour déterminer si la gravité des accidents est liée avec le moment de l'accident ?



A) Test de Shapiro

B) Test de Student

C) Test de Kruskall

D) Test du chi-deux

<u>Exercice 1 : IDC</u>

Chaque semaine, le magasin Monoprix de A. sélectionne un échantillon de 100 clients, pour estimer le montant moyen des dépenses de chaque client. En se fondant sur les nombreuses enquêtes précédentes, Monoprix suppose que la dépense de chaque client suit approximativement une loi normale, d'écart-type 10 euros.

Cette semaine, la moyenne de l'échantillon observé par le Monoprix de A. est 45 euros. On note µ le montant moyen dépensé par les clients cette semaine.

1) Déterminez un intervalle de confiance pour µ avec alpha=5%
2) Comment évolue l'amplitude de l'intervalle de confiance si 10 euros est l'écart-type estimé sur l'échantillon ?
3) Comment évolue l'amplitude de l'intervalle de confiance si alpha=1% ?
4) Comment évolue l'amplitude de l'intervalle de confiance si on réduit l'échantillon à 50 personnes ?

Exercice 2 : comparaison d'estimateurs

On considère la répartition des ménages français en deux classes C_1 et C_2 suivant qu'ils habitent dans une commune urbaine ou dans une commune rurale.

On désigne respectivement par p_1 et p_2 les proportions dans chaque classe des ménages effectuant un tri sélectif des déchets.

Pour estimer p_1 et p_2, on a tiré au hasard et avec remise n_1 ménages dans C_1 et n_2 ménages dans C_2 et pris pour estimateurs les fréquences F_1 et F_2 de ménages effectuant le tri sélectif dans chaque échantillon. On suppose ici que $p_1=p_2=p$ où p est la proportion sur toute la population.

1) Les estimateurs F_1, F_2 et $F_3=(F_1+F_2)/2$ sont-ils sans biais ?
2) Quel est le meilleur des trois suivant les valeurs de n_1 et n_2?

Exercice 3 : test d'hypothèses

On s'intéresse au problème des algues toxiques qui atteignent certaines plages de France ; après étude on constate que p=10% des plages sont atteintes par ce type d'algues et on veut tester l'influence de rejets chimiques nouveaux sur l'apparition de ces algues. Pour cela 50 plages proches de zones de rejet chimiques, sont observées ; on compte alors le nombre de plages atteintes par l'algue nocive : on constate que 10 plages sont atteintes par l'algue. Nous allons construire un test pour répondre à la question « Les rejets chimiques ont-t-il modifié, de façon significative, avec le risque $\alpha=5\%$, le nombre de plages atteintes ?»

1) Etablissez les hypothèses dans le cas d'un test bilatéral sur la proportion p.
2) Déterminez l'estimateur qui va être calculé sur l'échantillon. Quelle est sa loi ? Justifiez.
3) Faire un graphique représentant la région critique, l'erreur de première espèce et l'erreur de deuxième espèce.
4) Déterminez les seuils de décision de la région critique.
5) Quelle est votre conclusion ?

Exercice 4 : Régression linéaire et ANOVA

On dispose d'un échantillon de 20 champions du monde de saut en hauteur pour lesquels on dispose de leur performance, leur taille et leur provenance.

	Provenance	Taille	Performance
Annys	Autre	1,87	2,36
Austin	US	1,84	2,40
Brumel	URSS	1,85	2,28
...			

1) On souhaite savoir si la taille permet d'expliquer la performance. On réalise une régression linéaire.

```
Call:
lm(formula = Performance ~ Taille, data = tab)

Residuals:
      Min        1Q    Median        3Q       Max
-0.084835 -0.012198 -0.000999  0.012596  0.047492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.73125    0.17958   9.641 1.56e-08 ***
Taille       0.34248    0.09606   3.565  0.00221 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03176 on 18 degrees of freedom
Multiple R-squared:  0.4139,    Adjusted R-squared:  0.3813
F-statistic: 12.71 on 1 and 18 DF,  p-value: 0.002212
```

a) Quel est le modèle construit ?
b) Pourquoi les p-valeurs en gras sont-elles identiques ? Que pouvez-vous en conclure ?
c) Que pouvez-vous conclure à l'aide de la p-valeur de la constante (intercept)? Est-ce que cela est cohérent ?
d) Que pensez-vous de ce modèle ?

2) On souhaite maintenant savoir si la provenance a un impact sur la performance. On effectue un test de l'ANOVA.

```
Analysis of Variance Table

Response: Performance
           Df    Sum Sq   Mean Sq F value Pr(>F)
Provenance  2 0.0026133 0.0013067  0.7831 0.4728
Residuals  17 0.0283667 0.0016686
```

a) A quoi correspond la somme de la colonne "Sum Sq"?
b) Quelle est l'hypothèse H_0 de ce test ?
c) Quelles sont les conditions pour que ce test soit valable ?
d) Que pouvez-vous conclure ?