**Université Joseph Fourier**
**L2/STA230**

---

TP : Tests statistiques

---

# 1 Introduction :

L'objectif de ce TP est de savoir réaliser différents **tests statistiques**. Nous rappelons qu'un test statistique est une décision prise au vu des observations. Réaliser un test statistique se fait en 2 étapes :

1. La **construction du test** est la construction de la règle de décision. **Cette etape ne nécessite pas l'utilisation des observations.**
2. **Tester $H_0$ à partir des données.** Cette étape consiste simplemeent à appliquer la règle de décision construite précédement aux données.

On construit un **test paramétrique** pour répondre à la question : "Peut on considérer que le poids moyen des bébés est 2,5 Kg ?". On construit un **test de comparaison** d'échantillons **non appariés** pour répondre à la question "Le poids moyen des bébés issus de mère fumeuse est il identique au poids moyen des bébés issus de mère non fumeuse ?". On construit un **test de comparaison** d'échantillons **appariés** pour répondre à la question "le bras droit est il plus long que le bras gauche ?" ; Le test est apparié parce que les variables à comparer (longueur de bras gauche et longueur de bras droit) concernent un même individu. On construit un **test d'ajustement** pour répondre à la question "La distribution des bébés suivant leur poids suit elle une loi uniforme ?". Enfin on construit un **test d'indépendance** pour répondre à la question "Y a t il indépendance entre le poids des bébés et l'âge de la mère ?"
Nous considérons dans tout ce TP que $\alpha = 5\%$.

# 2 Tests parametriques pour un échantillon gaussien :

Ouvrez le fichier de données "BB.csv". Nous nous intéressons au poids des bébés et nous considérons que ces mesures sont indépendantes et suivent une loi normale de moyenne $m$ et de variance $\sigma^2$.

## 2.1 Tests sur la moyenne :

1. Calculez le poids moyen $\bar{x}$ des bébés. Rappelez vous que $\bar{x}$ est une estimation du poids moyen réel des bébés : $m$. Cette estimation est liée à l'échantillon. Sa valeur change avec l'échantillon. On note $\bar{X}_n$ la **variable aléatoire** : moyenne estimée à partir d'un échantillon de taille $n$.

2. Supposez que la variance de votre échantillon $\sigma^2$ est connue et égale à 600000. Construisons ensemble un test d'égalité de la moyenne $m$ avec la valeur $m_0 = 2,5\ kg$ :
   - **Construction du test. Nous rappelons que cette étape ne nécessite pas l'utilisation des observations :**
     - L'hypothèse nulle est $H_0 : m = m_0 = 2500\ g$, l'hypothèse alternative est $H_1 : m \neq m_0$.
     - La statistique de test utilisée est $T = (\bar{X}_n - m_0) * \frac{\sqrt{n}}{\sigma}$. Nous utilisons cette statistique car nous savons que si les données sont gaussiennes et de moyenne $m_0$, c'est à dire si $H_0$ est vraie, alors, $T$ suit **sous** $H_0$ une loi normale centrée réduite. Notez qu'on ne connait pas la loi de $T$ si $H_0$ est fausse.
     - La région critique de $T$ est de la forme $|T| > t_{seuil}$. En reprenant l'expression de $T$, on en déduit que la région critique de $\bar{X}_n$ est de la forme : $\bar{X}_n \in ]-\infty, m_0 - a] \cup [m_0 + a, \infty]$. En effet si $H_0$ est vraie, $\bar{X}_n$ doit être *proche* de $m_0$, $(|\bar{X}_n - m_0| < a)$ : l'ecart entre moyenne estimée $\bar{x}$ et moyenne $m_0$ s'explique par l'erreur d'estimation dûe au nombre fini d'observations $n$ ainsi qu'à la dispersion des données. Si $H_0$ est fausse, on s'attend à ce que $\bar{X}_n$ soit *significativement loin* de $m_0$ : l'ecart entre moyenne estimée $\bar{x}$ et moyenne $m_0$ s'explique par l'erreur de choix d'hypothèse, "$m = m_0$" est une hypothèse fausse. Représentez sur un brouillon la région critique.
     - Afin de calculer les bornes de la région critique, calculez le seuil $t_{seuil}$ tel que : $P(|T| \leq t_{seuil}) = 1-\alpha$. Ceci revient à calculer $t_{seuil}$ tel que $P(T < t_{seuil}) = 1-\frac{\alpha}{2}$. $t_{seuil}$ est donc le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. Sous Open Office, on obtient donc $t_{seuil}$ avec la fonction : LOI.NORMALE.STANDARD.INVERSE$(1 - \alpha/2)$.
     - Ecrivez la région critique sous la forme $\bar{X}_n \in ]-\infty; Inf] \cup [Sup, \infty]$ en calculant les valeurs de $Inf = m_0 - t_{seuil} * \frac{\sigma}{\sqrt{n}}$ et $Sup = m_0 + t_{seuil} * \frac{\sigma}{\sqrt{n}}$
   - **test de $H_0$ à partir des données :**
     - Calculez la valeur de la statistique de test pour l'échantillon : $t_{obs} = (\bar{x} - m_0) * \frac{\sqrt{n}}{\sigma}$. Si $t_{obs}$ est dans la région critique $(|t_{obs}| > t_{seuil})$, on rejette $H_0$ avec un risque de le rejeter à tort inférieur à $\alpha$, sinon on accepte $H_0$. Il existe un risque d'accepter à tort $H_0$, que l'on ne vous demande pas de calculer. Il faut bien comprendre que choisir $H_0$ revient à privilégier à priori cette hypothèse. On construit le test de telle sorte que l'on contrôle le risque de rejeter à tort $H_0$, de rater $H_0$.
     - On peut connaitre exactement la probabilté de rejeter à tort $H_0$. C'est la p-valeur $p_{val}$. Le risque $\alpha$ est tel que $P(|T| \geq t_{seuil}/H_0) = \alpha$, la probabilité p-valeur est telle que : $P(|T| \geq |t_{obs}|/H_0) = p_{val}$, soit encore $P(T \leq -|t_{obs}|/H_0) = P(T \geq |t_{obs}|/H_0) = p_{val}/2$. Sous Open Office on obtient donc la p-valeur avec la fonction : $p_{val}$=2*(1- LOI.NORMALE.STANDARD($|t_{obs}|$).
3. Nous supposons toujours la variance connue et étant égale à 600000. En suivant pas à pas le modèle de la question précédente, construisez maintenant le test $m = 2500\ g$ contre $m < 2500\ g$.
4. Nous supposons toujours la variance connue et étant égale à 600000. En suivant le modèle de la question précédente, construisez maintenant le test $m = 2500\ g$ contre $m > 2500\ g$. Les conclusions de ces 2 derniers tests sont elles cohérentes ?

5. Construisez maintenant le test $m = 3000\ g$ contre $m > 3000\ g.$, puis le test $m = 3000\ g$ contre $m < 3000\ g$. Les conclusions de ces 2 derniers tests sont elles cohérentes ?
6. **Nous ne supposons plus la variance connue**. Calculez l'estimation non biaisée de la variance $S'^2$. En suivant le modèle de la question précédente, construisez un test d'égalité de la moyenne $m$ avec la valeur $m_0 = 2,5\ kg$. Inteprétez et comparez vos résultats avec ceux obtenus à la question 2. En particulier, quel est le test le plus précis ?

### 2.2 Tests sur la proportion :

1. Estimez à partir des données brutes la proportion de bébés pesant moins de 2,5 Kg.
2. Construisez un test paramétrique vous permettant de répondre à la question "Peut-on considérer que la moitié des bébés issus de mère fumeuse pèsent moins de 2,5 Kg ?" donner la région critique du test. Donner la p-valeur. Interpréter.

## 3 Test de comparaison de 2 échantillons :

### 3.1 Echantillon non apparié :

Construisez le test de comparaison entre le poids moyen des bébés issus de mère fumeuse et le poids moyen des bébés issus de mère non fumeuse. Détaillez toutes les étapes et conclure. Rappelons qu'il s'agit d'un test de comparaison dans le cas d'échantillons **non appariés** : chaque individu est décrit par une seule variable (ici le poids) et nous comparons le comportement (ici la moyenne) de cette variable dans 2 populations différentes.

### 3.2 Echantillon apparié :

1. Ouvrez le fichier "Humerus.csv". Il contient les longueurs d'humerus droit et gauche, prises sur 10 squelettes feminins. nous considérons que ces longueurs suivent une loi normale. Le test que vous allez construire répondra à la question "Y a t il une différence de longueur entre humerus droit et humerus gauche chez les femmes ?" Rappelons qu'il s'agit d'un test de comparaison dans le cas d'échantillons **appariés** : chaque individu est décrit par 2 variables (ici la longueur de l'humerus gauche et celle de l'humerus droit) et nous comparons le comportement (ici la moyenne) de ces 2 variables au sein de la même population.
2. Pour chaque squelette, calculez la différence de longueur entre humerus droit et humerus gauche. Soit $D$ cette variable. $D$ suit une loi normale de moyenne $m_D$ et de variance $\sigma_D^2$
3. Calculez la moyenne et la variance estimées de $D$,
4. Construisez le test $H_0$ : il n' y a pas de différence de longueur entre l'humérus droit et l'humérus gauche chez les femmes contre $H_1$ : il y a une différence de longueur entre l'humérus droit et l'humérus gauche chez les femmes. Pour cela, remarquez que le test peut s'écrire sous la forme $H_0 :\ m_D = 0$ contre $H_1 :\ m_D \neq 0$.
5. Concluez !!

# 4 Test d'adéquation :

Vous allez réaliser un test d'adéquation de la distribution empirique des poids à une distribution théorique de loi normale dont les paramètres sont la moyenne $\bar{x}$ et la variance $s'^2$ estimées à partir des données brutes.
Pour cela, vous allez calculer la distance du $\chi^2$ entre la distribution observée des bébés suivant leur poids et la distribution théorique des bébés suivant leur poids que l'on aurait si ces poids suivaient bien une loi normale de parametres $\bar{x}$ et $s'^2$. Cette distance suit une loi du $\chi^2$ **sous l'hypothèse d'adéquation** et sous les conditions : l'effectif théorique dans chaque classe est supérieur ou égal à 5. Le nombre de degrés de libertés $d$ du $\chi^2$ est égale à : $d = n_c - n_p - 1$ où $n_c$ est le nombre de classes utilisé dans le tableau de distribution, $n_p$ le nombre de parametres de la loi théoriques que vous avez estimé à partir des données ( ici $n_p = 2$ puisque on utilise une loi normale dont la moyenne et la variance ont été estimées à partir des données).

1. Construisez le tableau de distribution observé des poids,
2. Construisez le tableau de distribution théorique des poids, en utilisant la loi normale dont les paramètres sont la moyenne $\bar{x}$ et $s'^2$,
3. Regrouper les classes de telle sorte que l'effectif théorique dans chaque classe soit supérieur à 5.
4. Calculer la distance du chi deux entre les 2 distributions,
5. Réalisez le test d'adéquation. Détaillez toutes les étapes.

# 5 Test d'indépendance :

1. Y a t il indépendance entre l'âge et le poids des bébés observés ?. il vous faut pour construire ce test construire le tableau de la loi jointe, le tableau des effectifs théoriques sous l'hypothèse d'indépendance, procéder au regroupement de classes dans les 2 tableaux de telle sorte que les effectifs théoriques soient dans chaque classe supérieur à 5, puis calculer la distance du chi deux et la p valeur associée. Conclure.
2. On souhaite maintenant comparer la distribution des poids des bébés suivant que leur mère est fumeuse ou non fumeuse. Ce test peut être vu comme un test d'indépendance entre les variables "poids" et "Etat Fumeur ou Non Fumeur"