TP 2
Tests d'hypothèses

Dans ce TP, on va s'intéresser à la construction de différents tests d'hypothèse, dont on illustrera l'usage sur le jeu de données contenu dans `Tips.csv`. Les variables suivantes y sont présentes

- `TOTBILL` : montant total (en $)
- `TIP` : montant du pourboire (en $)
- `SEX` : sexe de la personne qui a payé
  - `SEX = 0` : homme
  - `SEX = 1` : femme
- `SMOKER` : zone du restaurant
  - `SMOKER = 0` : zone non-fumeur
  - `SMOKER = 1` : zone fumeur
- `DAY` : jours de la semaine
  - `DAY = 3` : mardi
  - `DAY = 4` : mercredi
  - `DAY = 5` : jeudi
  - `DAY = 6` : vendredi
- `TIME` : moment de la journée
  - `TIME = 0` : en journée
  - `TIME = 1` : en soirée
- `SIZE` : nombre de personnes

Traitement et description des données

1. Charger dans une variable nommée `tips` la table de données comprise dans le fichier `Tips.csv`. Vérifier que le type de chaque variable correspond bien à ce qu'elle encode. Changer la classe des colonnes lorsque c'est nécessaire.
2. Dans la suite, on va chercher à comparer la générosité des clients en terme de pourboire, selon par exemple le moment de la journée. Le montant du pourboire est-il vraiment pertinent pour "quantifier" la générosité des clients ? Quelle variable plus pertinente faut-il introduire ?
3. Construire finalement 2 sous-échantillons pour cette nouvelle variable, `TipsJour` pour les commandes effectuées en journée et `TipsSoir` pour celles prises en soirée. Tracer les histogrammes de chaque échantillon et commenter leur distribution.

On va revoir la construction et l'implémentation de 3 tests d'hypothèse classiques : Fisher (test de variance), Student (test de moyenne), Kolmogorov (test d'adéquation).

On se place dans un contexte où

- on dispose de réalisations de 2 échantillons de n et m v.a. $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ où les X_i sont indépendants des Y_j.
- Les X_i (resp. les Y_j) sont iid d'espérance μ_X (resp. μ_Y) et de variance σ_X^2 (resp. σ_Y^2).
- Les paramètres μ_X, σ_X^2, μ_Y, σ_Y^2 sont tous **inconnus**.

1 Test de Fisher

Le test de Fisher permet de comparer les variances de deux échantillons issus de populations **indépendantes**. L'hypothèse nulle s'écrit ici

$$\mathcal{H}_0 \,:\, \{\sigma_X^2 = \sigma_Y^2\}$$

La statistique du test de Fisher est

$$T_{\mathcal{F}} = \frac{S_X^2}{S_Y^2}$$

où S_X^2 et S_Y^2 sont les variances empiriques corrigées des variances de X et Y. **On se place dans le cas d'un test bilatéral.**

1. On va appliquer ce test pour comparer les variances de `TipsJour` et `TipsSoir`. À partir de vos réponses à l'Exercice 1, calculer dans R :
   — l'intervalle de confiance de niveau 95% pour le rapport des variances des 2 échantillons ;
   — la statistique de test ;
   — la zone de rejet au seuil 5% ;
   — la p-valeur du test.

   Quelle décision prendre ? Commenter l'usage de ce test.
2. Le test de Fisher est implémenté dans R via la fonction `var.test`. À partir de l'aide de la fonction, comparer les variances de `TipsJour` et `TipsSoir` et retrouver les résultats calculés ci-dessus.

2 Test de Student

Le test de Student permet de comparer les moyennes des deux échantillons. Celui-ci nécessite cependant que les variances théoriques σ_X^2 et σ_Y^2 soient égales ! On peut s'en assurer via un test de Fisher. L'hypothèse nulle s'écrit donc

$$\mathcal{H}_0 : \{\mu_X = \mu_Y\}$$

La statistique du test de Student est

$$T_{St} = \frac{\overline{X} - \overline{Y}}{\sqrt{S_{XY}^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

avec

$$S_{XY}^2 = \frac{n-1}{n+m-2} S_X^2 + \frac{m-1}{n+m-2} S_Y^2$$

où S_X^2 et S_Y^2 sont les variances empiriques corrigées de X et Y. **On se place dans le cas d'un test <u>bilatéral</u>.**

1. On va appliquer ce test pour comparer les moyennes de `TipsJour` et `TipsSoir`. À partir de vos réponses à l'Exercice 2, calculer dans R :
   — l'intervalle de confiance de niveau 95% pour la différence des moyennes des 2 échantillons ;
   — la statistique de test ;
   — la zone de rejet au seuil 5% ;
   — la p-valeur du test.

   Quelle décison prendre ? Commenter l'usage de ce test.
2. Le test de Student est implémenté dans R via la fonction `t.test`. À partir de l'aide de la fonction, comparer les moyennes de `TipsJour` et `TipsSoir` et retrouver les résultats calculés ci-dessus.

3 Test de Kolmogorov

Le test de Kolmogorov est un test non-paramétrique, permettant de tester l'adéquation d'un échantillon a une loi donnée. On teste donc ici

$$\mathcal{H}_0 : \{\mathbb{P} = \mathbb{P}_0\} \quad \text{contre} \quad \mathcal{H}_1 : \{\mathbb{P} \neq \mathbb{P}_0\}$$

où $\mathbb{P}$ est la loi dont sont issues nos données, et $\mathbb{P}_0$ la loi qu'on cherche à tester.

Le test de Kolmogorov effectue cette comparaison en testant la fonction de répartition, c'est-à-dire

$$\mathcal{H}_0 : \{F = F_0\} \quad \text{contre} \quad \mathcal{H}_1 : \{F \neq F_0\}$$

où F est la fonction de répartition de laquelle est issu notre échantillon, et F_0 est une fonction de répartition continue d'une loi donnée. La statistique de test est donnée par

$$T_{\mathcal{K}} = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right|$$

où $\hat{F}_n$ est la fonction de répartition empirique de l'échantillon

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[X_i \leq x\right]$$

On peut en fait montrer que (et c'est cette formule que l'on utilise en pratique)

$$\begin{aligned} T_{\mathcal{K}} &= \max_{i=0,\ldots,n} \Delta_i \\ \text{où } \Delta_i &= \max\{\Delta_i^- \,;\, \Delta_i^+\} \end{aligned} \tag{1}$$

avec

$$\Delta_i^- = \left|\frac{i}{n} - F_0(X_{(i)})\right| \qquad \text{et} \qquad \Delta_i^+ = \left|\frac{i}{n} - F_0(X_{(i+1)})\right|$$

où $X_{(1)} \leq \cdots \leq X_{(n)}$ est l'échantillon initial qu'on a réordonné, et en posant $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$.

La loi de $T_{\mathcal{K}}$ sous $\mathcal{H}_0$ est tabulée (voir Table 1). Ces tables sont obtenues par simulation de $T_{\mathcal{K}}$. On trouve dans ces tables les quantiles $d_{n,1-\alpha}$, permettant ainsi de rejeter $\mathcal{H}_0$ si $T_{\mathcal{K}} \geq d_{n,1-\alpha}$ en se donnant un risque α. En théorie, il faudrait une table pour chaque loi F_0 que l'on souhaite tester. Cependant, le test s'appuie de plus sur le résultat suivant particulièrement remarquable : si $\mathcal{H}_0$ est vraie, alors la loi de $T_{\mathcal{K}}$ ne dépend pas de F_0.

De fait, la Table 1 a été calculée en simulant $T_{\mathcal{K}}$ sous l'hypothèse que $(X_1, \ldots, X_n)$ était issu d'une loi uniforme sur [0;1].

n	$\alpha = 0.20$	$\alpha = 0.15$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404

TABLE 1 – Valeurs critiques pour le test de Kolmogorov selon la taille d'échantillon n et le risque α

Pour des échantillons plus grand, la valeur critique pour le seuil α est approximée par

$$\sqrt{-\frac{1}{2n}\log\left(\frac{\alpha}{2}\right)}$$

On va appliquer à présent le test de Kolmogorov pour vérifier la normalité des échantillon `TipsJour` et `TipsSoir`. Pour chaque échantillon :

1. Tracer sur le même graphique la fonction empirique de l'échantillon ainsi que la fonction de répartition que l'on teste (vous pouvez vous servir de la fonction `ecdf` ou implémenter vous même le calcul de $\hat{F}_n$) Commenter.
2. À l'aide de (1) et la Table 1, prenez une décision sur la normaluté des données en se donnant un risque de première espèce $\alpha = 5\%$.

3. Le test de Kolmogorov est implémenté dans R via la fonction `ks.test`. À partir de l'aide de cette fonction, tester la normalité de `TipsJour` et `TipsSoir`.
4. Dans le cas spécifique où l'on veut tester l'adéquation à une loi Normale (comme ici), il est préférable d'utiliser le test de Shapiro-Wilk. Ce test éprouve la normalité d'un échantillon sans nécessiter de renseigner les paramètres. Appliquer le test de Shapiro-Wilk à l'aide de la fonction `shapiro.test` pour éprouver la normalité de `TipsJour` et `TipsSoir`.

4 Distribution de la p-valeur

On se place dans le contexte où on teste une hypothèse nulle $\mathcal{H}_0$ et on s'intéresse ici au comportement théorique de la p-valeur d'un test d'hypothèse sous l'hypothèse nulle.

1. Justifier que la p-valeur d'un test d'hypothèse est une variable aléatoire.
2. Donner la loi de la p-valeur sous $\mathcal{H}_0$.
3. Proposer une étude par simulation pour illustrer ce résultat sur le test de votre choix.
4. Proposer une étude par simulation pour illustrer ce qu'il se passe lorsqu'on utilise un test de Student sur des données non-gaussiennes, en faisant varier la taille de l'échantillon.

5 Étude théorique

Exercice 1 Test de Fisher

On se place dans le contexte de la Section 1.

1. On rappelle que dans le cas de deux échantillons issus de $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ alors
$$\frac{S_X^2}{\sigma_X^2}\frac{\sigma_Y^2}{S_Y^2} \sim \mathcal{F}(n-1, m-1)$$
En déduire un intervalle de confiance de niveau $1-\alpha$ pour le rapport $\frac{\sigma_X^2}{\sigma_Y^2}$.
2. Construire la zone de rejet de $\mathcal{H}_0$, pour un seuil α donné.
3. Exprimer le calcul de la p-valeur associée à ce test.

Exercice 2 Test de Student

On se place dans le contexte de la Section 2.

1. Sachant que dans le cas de deux échantillons issus de $X \sim \mathcal{N}(\mu_X, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$, si on a
$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_{XY}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim \mathcal{S}t(n+m-2)$$
donner un intervalle de confiance de niveau $1-\alpha$ pour la différence $\mu_X - \mu_Y$.
2. Construire la zone de rejet de $\mathcal{H}_0$, pour un seuil α donné.
3. Exprimer le calcul de la p-valeur associée à ce test.

Exercice 3 Statistique du test de Kolmogorov

1. On va démontrer l'équation (1) par double inégalité
   (a) Montrer que
   $$\left|\hat{F}_n(x) - F_0(x)\right| \leq \max\left\{\Delta_k^-;\Delta_k^+\right\}$$
   en étudiant les intervalles
   $$I_k = \left[X_{(k)}, X_{(k+1)}\right[, \qquad I_0 = \left]X_{(0)}, X_{(1)}\right[$$

pour $k = 0, \dots, n-1$, où $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$.

(b) En déduire (1) en exploitant l'aspect *càdlàg* de $\hat{F}_n$.

2. Montrer que sous $\mathcal{H}_0$ la loi de $T_{\mathcal{K}}$ ne dépend pas de F_0. *Indication : montrer que sous $\mathcal{H}_0$, on se ramène au cas où F_0 est la fonction de réparition de la loi Uniforme $\mathcal{U}(0, 1)$.*