Statistiques décisionnelles : TP 1

À rendre pour le 12/03/2021 : Vous devez rédiger les réponses à ces trois exercices et me rendre un seul document pdf. Vous pouvez utiliser le traitement de texte de votre choix (Latex, Word, LibreOffice ou bien même R Markdown...).

Pour chaque exercice, vous rédigerez des explications et vous fournirez le code. Le langage utilisé doit impérativement être R. **Le travail à rendre est individuel.** Vous pouvez bien sûr vous entraider mais il ne s'agit pas d'un travail en groupe. Ce TP est noté et comptera dans votre note finale.

Vous trouverez sur Moodle un formulaire des commandes de R qui vous seront utiles.

Exercice 1 (Très bon exercice de révisions)**.** *On dispose de l'échantillon de* 50 *variables indépendantes suivant, les données ont été classées par ordre croissant pour en faciliter le traitement. Les ex-aequo proviennent d'arrondis.*

0.02	*0.02*	*0.07*	*0.07*	*0.07*	*0.09*	*0.11*	*0.13*	*0.16*	*0.16*
0.2	*0.2*	*0.21*	*0.25*	*0.29*	*0.32*	*0.35*	*0.35*	*0.37*	*0.37*
0.38	*0.4*	*0.4*	*0.41*	*0.47*	*0.49*	*0.54*	*0.55*	*0.59*	*0.6*
0.68	*0.68*	*0.7*	*0.73*	*0.75*	*0.79*	*0.79*	*0.8*	*0.81*	*0.81*
0.86	*0.89*	*1.03*	*1.04*	*1.04*	*1.09*	*1.12*	*1.21*	*1.33*	*1.42*

1. *Tester l'adéquation de ces données à une loi uniforme sur* $[0, 1.5]$ *à l'aide d'un test du* χ^2*, au niveau* 90%*. On proposera un découpage en* 5 *intervalles de même longueur. Calculer la p-valeur.*
2. *Tester l'adéquation de ces données à la loi dont la densité est*

$$f(x) = \cos(x) 1_{x \in [0, \pi/2]}$$

   à l'aide d'un test du χ^2*, au niveau* 90%*. On pourra calculer la fonction de répartition de cette fonction et déterminer les probabilités des intervalles construits précédemment, ou bien en construire de nouveaux à l'aide des quantiles. Calculer la p-valeur.*
3. *Reprendre les questions précédentes avec un test de Kolmogorov-Smirnov en modifiant les valeurs pour ne pas avoir d'ex-æquo (il n'y en avait aucun avant l'arrondi). Calculer la p-valeur dans chacun des cas et comparez aux p-valeurs obtenues précédemment.*

Exercice 2 (Découvrez mes week-ends)**.** *C'est à votre tour de créer un exercice pour un TD de L3 MIASHS. Le but est de construire un exercice et sa solution, en ne faisant aucun calcul à la main. Générez des données issues de la loi de votre choix, et construisez un exercice sur un des tests d'ajustement vu en cours.*

Attention, vous devrez aussi faire un énoncé (si possible avec une mise en situation), les questions intermédiaires si nécessaire et la solution finale (où tous les calculs sont faits en R directement).

Pour cet exercice, vous aurez sans doute besoin de générer des variables aléatoires avec la commande `rexp` *ou* `rpois` *(ou r+n'importe quelle classe de variables aléatoires), en faisant bien attention aux paramètres de ces lois.. Vous pouvez aussi les simuler "à la main" à partir de variables uniformes (*`runif`*) à l'aide de la méthode de simulation par inversion de la fonction de répartition.*

À rendre :

1. *Un énoncé d'exercice portant sur le test d'ajustement de votre choix.*
2. *Les codes que vous avez utilisés, avec quelques commentaires explicatifs (vous pouvez faire des captures d'écran si vous le souhaitez).*
3. *La solution rédigée, où tous les calculs sont fait en R, ainsi que les codes qui ont permis de créer le jeu de données.*

Exercice 3 (Créations de tables de lois)**.** *Soit $X_1, \dots, X_n$ un échantillon de variables i.i.d.. Le but de cet exercice est d'obtenir les quantiles de la loi de X_1 à partir de l'échantillon, pourvu que n soit assez grand. On note $X_{(1)}, \dots, X_{(n)}$ l'échantillon ordonné par ordre croissant.*

On admet qu'on peut estimer $Q(p)$, le quantile d'ordre $p \in]0,1[$ de X_1 par

$$Q_n(p) = \begin{cases} X_{(np)} & \text{si } np \text{ est entier} \\ X_{\lfloor (np) \rfloor + 1} & \text{sinon.} \end{cases}$$

$Q_n(p)$ est un estimateur convergent du quantile d'ordre p de la loi de X. On admet de plus que la précision de cette estimation est d'ordre $1/\sqrt{n}$. En réalité, la largeur de l'intervalle de confiance autour de la vraie valeur du quantile $Q(p)$ est de la forme $c/\sqrt{n}$ où la constante c dépend de la loi des X_i et de p et est donc inconnue. On se contentera ici de dire que l'approximation est d'ordre $1/\sqrt{n}$, ce qui est une affirmation très imprécise.

Remarque 1. *Cet exercice vous fait découvrir une méthode d'estimation des quantiles. Ayez conscience que ce n'est en général pas la meilleure et que les tables que je vous distribue n'ont pas été obtenues ainsi. Par exemple, pour les lois à densité, on préfèrera calculer de manière approchée la fonction de répartition (avec une très grande précision pour le cas des gaussiennes), et ensuite en déduire les quantiles.*

La méthode d'estimation proposée donne des résultats corrects pour les quantiles "pas trop proches de 0 *ni de* 1*". En effet, la précision s'effondre dès lors qu'on cherche à estimer un quantile trop proche de* 1 *(avec* 10000 *données, il serait déraisonnable de vouloir estimer le quantile d'ordre* 0.999 *et espérer avoir de bons résultats).*

1. *Avec cette définition, si $n = 1000$, quelle variable faut-il regarder si on veut le quantile d'ordre* 0.95 *? Quelle variable faut-il regarder si on veut le quantile d'ordre* 0.975 *?*
2. *Quelle valeur de n faut-il choisir pour garantir que le quantile estimé a ses deux premiers chiffres après la virgule corrects ? Dans toute la suite, on utilisera au minimum cette valeur de n.*
3. *Utilisez la fonction* `rnorm` *pour recréer la table des quantiles de la loi normale entre* 0.8 *et* 0.95*, échantillonnée tous les* 0.1*. Comparez aux vrais quantiles de la loi normale. Quelle valeur de n faut-il prendre pour obtenir une précision de 10^{-2} ?*

4. *On cherche à obtenir la table de la loi d'une somme de deux variables uniformes sur* $[0,1]$ *et indépendantes. Donner les quantiles de cette loi entre* 0.8 *et* 0.95*, échantillonnée tous les* 0.1*. Vous tracerez aussi l'histogramme des valeurs obtenues et la fonction de répartition empirique observée.*

5. *Écrire un programme qui calcule la statistique de Kolmogorov-Smirnov pour* 10 *variables observées*

$$h_{10} = \sup_{t\in\mathbb{R}}\{\frac{1}{10}\sum_{i=1}^{10} 1_{U_i\leq t} - t\}$$

où $U_1,\ldots,U_{10}$ *sont des variables indépendantes, uniformes sur* $[0,1]$ *et s'en servir pour obtenir la table des quantiles entre* 0.85 *et* 0.95*, échantillonnée tous les* 0.1*. Comparez la table obtenue avec celle distribuée en cours.*

6. *Obtenir la table des quantiles entre* 0.85 *et* 0.95*, échantillonnée tous les* 0.1 *de la loi de la statistique de Lilliefors pour* 10 *variables observées*

$$L_{10} = \sup_{t\in\mathbb{R}}\{\frac{1}{10}\sum_{i=1}^{10} 1_{X_i\leq t} - F_{(\overline{X_{10}},V_{10})}(t)\}$$

où $X_1,\ldots,X_{10}$ *sont des variables indépendantes,* $\mathcal{N}(0,1)$*, où* $\overline{X_{10}}$ *est la moyenne empirique, où* V_{10} *est la variance empirique de l'échantillon, et où* $F_{(m,\sigma^2)}$ *est la fonction de répartition d'une variable de loi* $\mathcal{N}(m,\sigma^2)$*.*