Université de Liège

Ingénieur Civil - 3ème BAC

9 décembre 2019

Éléments de statistiques

Projet

LIONE Maxime - s175606
SIBOYABASORE Cédric - s175202

1 Analyse descriptive

Chaque réponse aux sous-questions de cette question est appuyée par une partie de code dans le script **Q1.m**.

1.a

On génère l'histogramme avec la fonction *histogram.m* car elle représente les données en groupe d'intervalles (les pas) de manière automatique et pertinente contrairement à la fonction *hist.m* qui utilise par défaut 10 pas.

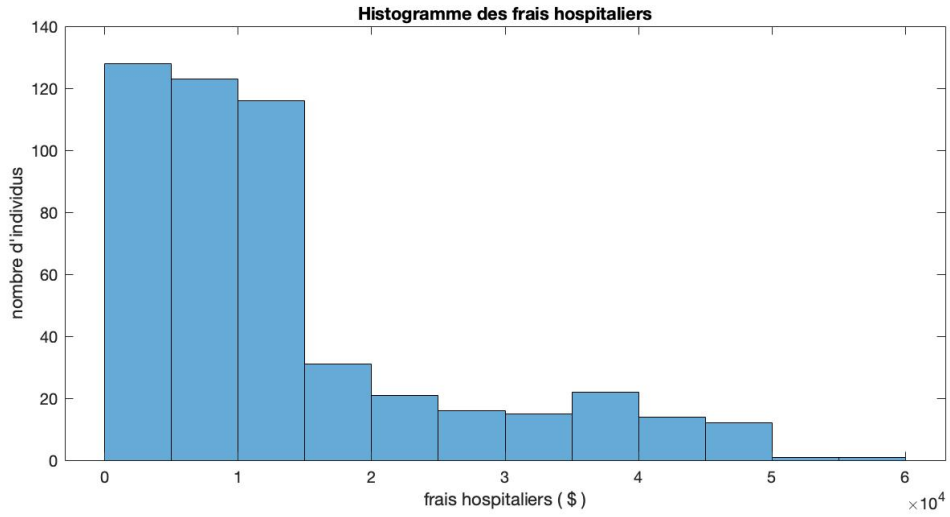

FIGURE 1

Cette figure représente donc les fréquences absolues des frais hospitaliers des 500 individus de notre population. Les données ont été regroupées par pas de 5000 des valeurs des frais. Elle nous renseigne donc le nombre d'individus pour lesquels les frais hospitaliers sont compris dans [0, 5000] , [5000, 10 000], ... , [55 000, 60 000].

On peut facilement interpréter cette figure en énonçant que les frais de la majorité des individus (73,4%) varient entre 0 et 15 000$. De plus, de manière générale, le nombre d'individus diminue lorsque la valeur des frais hospitaliers augmente et il n'existe que 2 individus dans notre population pour lesquels les frais dépassent 50 000$.

1.b

Pour calculer la moyenne, la médiane et l'écart-type des frais hospitaliers de notre population, on utilise respectivement les fonctions *mean*, *median* et *std* de Matlab dans la fonction *sum_up.m* qu'on appelle sur notre population : On obtient une moyenne $\overline{f}$ des frais hospitaliers de 13 534.78 $, une médiane m de 9977.548 $ et un écart-type σ de 12 117.39 $.

L'écart-type de 12 117.39 $ indique la dispersion des frais hospitaliers par rapport à la moyenne. Comme il est grand, beaucoup de patients ont des charges éloignées de la moyenne et donc la moyenne ne reflète pas de manière pertinente la tendance centrale des charges.

Pour représenter la valeur centrale des charges, on peut également calculer la médiane de 9977.548 $ qui départage l'ensemble des charges en deux parts égales.

Ms. Smith a des frais de 16 855 $, situés dans la partie des frais supérieurs à la médiane. Bien que l'écart-type soit élevé, les frais de Ms. Smith sont plutôt proches de la moyenne (et supérieurs).

1.c

On considère comme normaux au sens de la loi normale des frais hospitaliers compris entre $\overline{f} - \sigma = 1417.4\,\$$ et $\overline{f} + \sigma = 25\,652\,\$$.

Notre fonction *normal_interval.m* calcule un intervalle de normalité, au sens d'une loi normale, d'une variable aléatoire X. Elle fait appel à la fonction *sum_up.m* afin d'obtenir la moyenne $\overline{f}$ et l'écart-type σ de X nécessaires au calcul des bornes de l'intervalle [1417.4, 25 562]$.

Pour connaître la proportion de patients ayant des frais normaux (au sens de la loi normale), on parcourt toutes les valeurs des frais hospitaliers dans notre population. On trouve que 83% des patients ont des frais hospitaliers compris dans cet intervalle de normalité.

Ms. Smith a des frais normaux car ses frais de 16 885$ sont compris entre 1417.4 et 25 652 $.

1.d

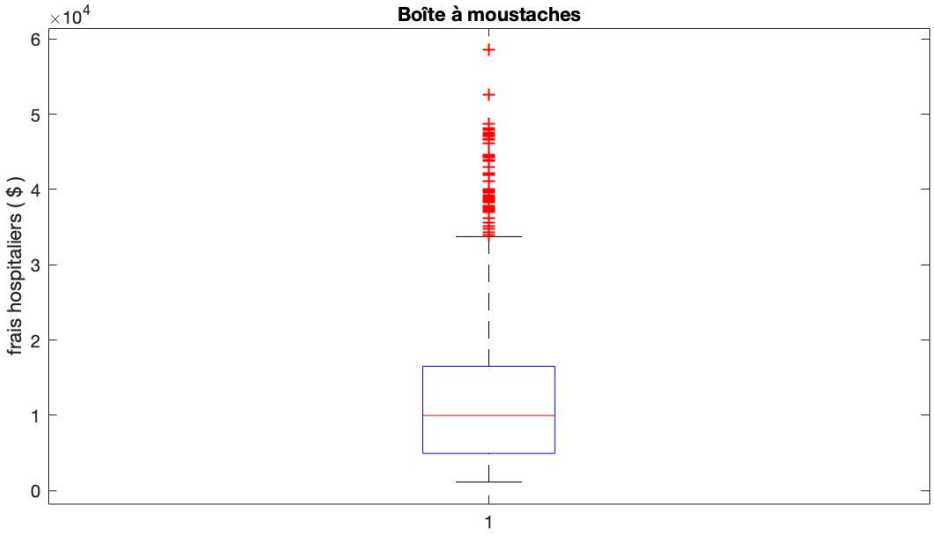

FIGURE 2

La boîte à moustaches résume les indicateurs de position du montant des frais hospitaliers tels que la médiane et les quartiles. La médiane est indiquée par la ligne rouge tandis que la boîte en bleu indique l'intervalle inter-quartile délimité par Q1 et Q3.

Le quartile Q1 (/ Q3) correspond à un seuil tel que 25% (/ 75%) des valeurs des frais dans toute la population soient inférieures à ce seuil. Q2 s'explique de la même manière et correspond donc à la médiane. Ils sont calculés dans la fonction *quartile.m* grâce à la fonction *quantile.m* de Matlab et valent 4927.281$ (Q1), 9977.548$ (Q2, même valeur qu'au point 1.b) et 16 499.137$ (Q3). On peut vérifier ces valeurs sur la boîte à moustaches : la ligne bleue inférieure sur la boîte à moustaches correspond bien à des frais de 4927.281 $, la ligne rouge correspond à 9977.548 $ et la ligne bleue supérieure à 16 499.137 $. On peut donc dire, par exemple, que 25% des individus de notre population ont des frais inférieurs à 4927.281$.

On peut clairement constater la présence de données aberrantes, repérées par des +. Une valeur est considérée comme aberrante si la valeur absolue de l'écart avec Q1 ou Q3 est supérieure à plus de 1,5 x Écart inter-quartile. On en calcule 55. Elles sont toutes au dessus du graphe (car Q1 étant peu élevé, la valeur 'limite' inférieure Q1 - 1.5 * (Q3 - Q1) est négative et les frais sont toujours positifs). La barre horizontale noire (la "moustache") inférieure correspond donc au minimum des frais hospitaliers dans toute la population, soit 1136.399$, tandis que la barre horizontale noire supérieure représente la valeur "limite" Q3 + 1.5(Q3 -Q1) = 33 856.92$ au dessus de laquelle les données sont aberrantes.

1.e

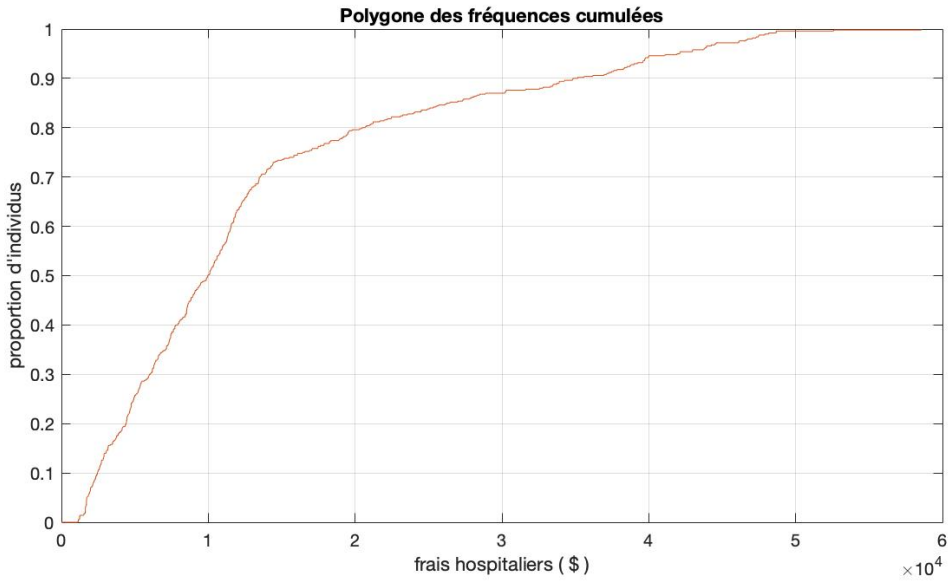

FIGURE 3

Cette figure représente les fréquences relatives cumulées des frais hospitaliers pour les 500 individus de notre population. Les données n'ont pas été regroupées au préalable. On y voit donc, pour chaque valeur de frais x, la proportion d'individus pour lesquels la valeur des frais est inférieure à x.

On pourrait donc estimer la proportion de patients ayant des frais inférieurs ou égaux à 25 000$ et supérieurs à ceux de Ms. Smith (16 884.924$) directement sur base de ce graphique, en soustrayant la proportion qu'on observe pour x = 16 884.924$ à celle qu'on observe pour x = 25 000$. Ces proportions sont calculées de manière précise grâce à la fonction *proportion.m*. Finalement, on trouve 8,4% de patients ayant des frais inférieurs ou égaux à 25 000$ et supérieurs à ceux de Ms. Smith.

1.f

Le coefficient de corrélation linéaire entre les âges des patients masculins et leurs charges hospitalières respectives est de 0.2302 tandis que celui entre les âges des patientes et leurs charges est de 0.3112. On trouve ces valeurs grâce à la fonction $corrcoeff.m$ de Matlab qui retourne une matrice dont les éléments diagonaux sont égaux au coefficient de corrélation linéaire recherché. La fonction *correlation.m* renvoie simplement les éléments diagonaux de la matrice retournée par $corrcoeff.m$.

Les coefficients de corrélation linéaire sont positifs mais même si on peut observer sur les scatterplots (Figure 1 et Figure 2, affichées par la fonction *scatter.m* de Matlab) un alignement de points et donc une droite, un grand nombre de points sont éloignés de cette "droite". Certains frais hospitaliers augmentent linéairement avec l'âge, mais pas tous.

Les coefficients sont faibles (0.2302 et 0.3112) donc il existe une faible relation linéaire entre l'âge d'un patient et ses charges, son âge ne peut pas prédire/expliquer ses frais hospitaliers. Bien que les femmes dépensent plus en fonction de leur âge que les hommes, la différence entre les deux coefficients de corrélation calculés est trop faible pour établir une liaison entre les dépenses d'un patient en fonction de son âge et son sexe.

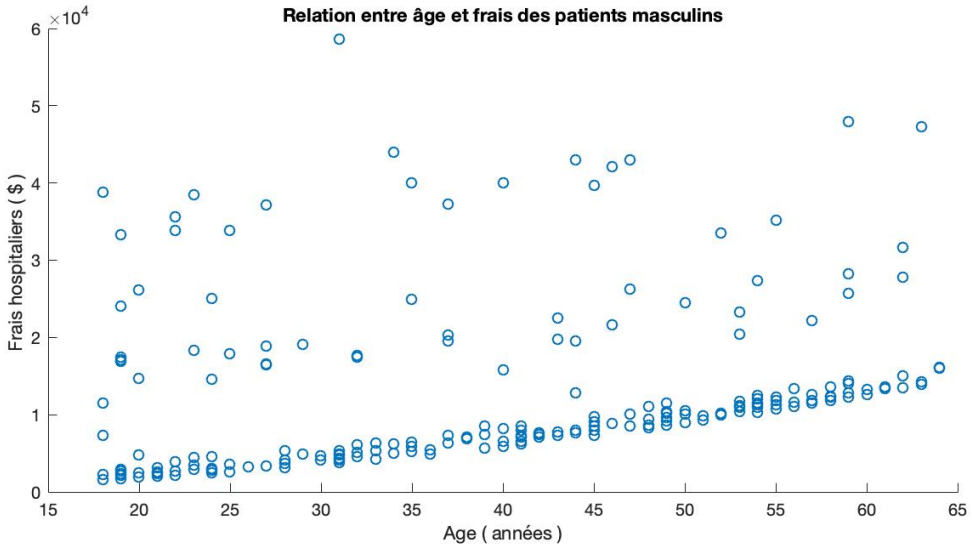

FIGURE 4

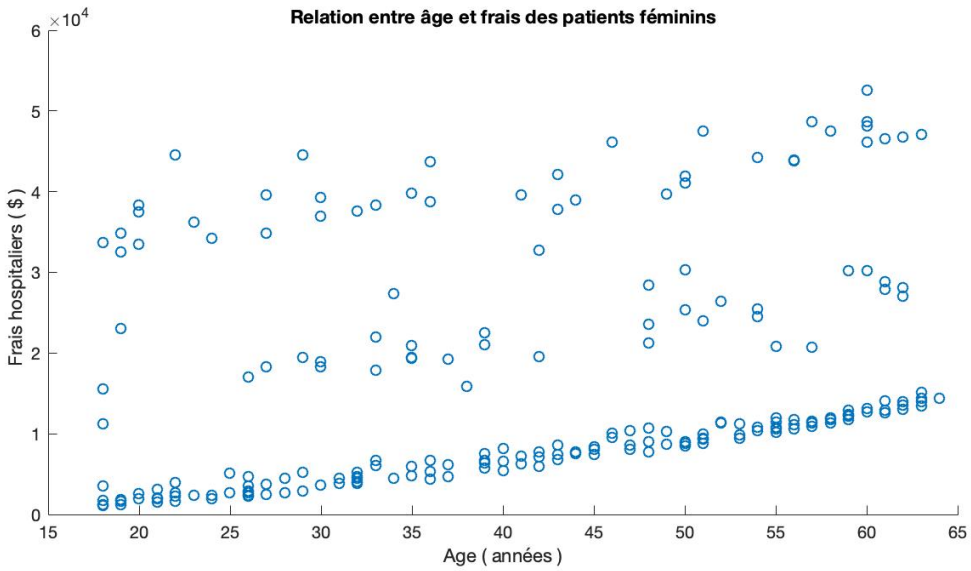
Relation entre âge et frais des patients féminins
×10^4
Frais hospitaliers ($)
Age (années)
0
1
2
3
4
5
6
15
20
25
30
35
40
45
50
55
60
65

FIGURE 5

2 Génération d'échantillons i.i.d.

Chaque réponse aux sous-questions de cette question est appuyée par une partie de code dans le script **Q2.m**

2.a

Pour tirer un échantillon i.i.d. de 50 patients, la fonction *iid_sample.m* établit d'abord un vecteur de 50 nombres aléatoires entre 1 et 500. Ces nombres sont piochés aléatoirement de manière uniforme grâce à la fonction *randsample.m* de Matlab. On spécifie dans les paramètres de cette fonction que le tirage doit être effectué avec remise ($True$). Ce tirage est répété 50 fois et chaque tirage est effectué indépendamment des tirages précédents. Les 50 nombres sont ensuite utilisés comme indices dans notre tableau de données pour récupérer les 50 patients choisis aléatoirement.

i. Pour un échantillon, en utilisant la fonction *sample_sum_up.m* dans laquelle on effectue les mêmes calculs que dans *sum_up.m*, on trouve une moyenne de 15 119.236$, une médiane valant 10 830.253$ et un écart-type de 12 878.544$. Ces valeurs d'échantillon sont chacune fortement proches des valeurs de population calculées au point 1.b, soit 13 534.78 $ pour la moyenne, 9977.548 $ pour la médiane et 12 117.39 $ pour l'écart-type. On peut donc dire que l'échantillon considéré ici représente plutôt justement la population entière.

ii.

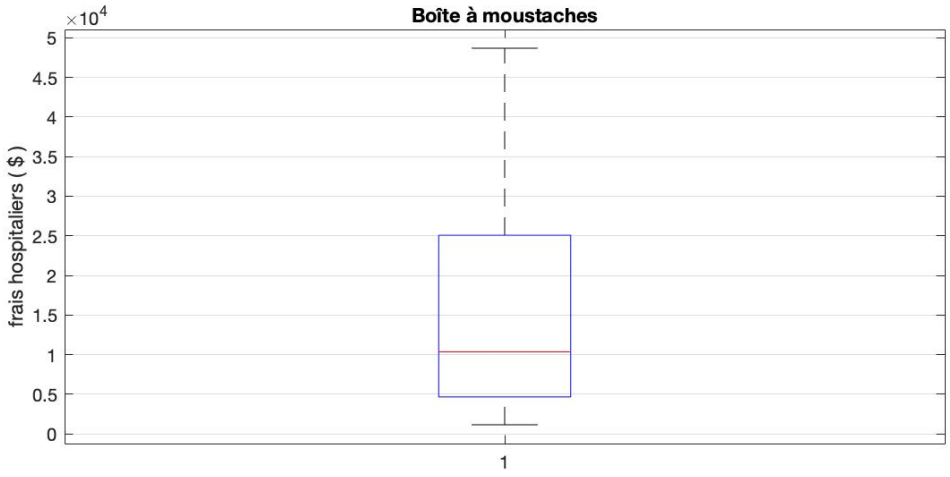

FIGURE 6

Dans le graphique de la boite à moustache de cet échantillon, il n'y a pas de valeurs aberrantes. Les moustaches supérieure et inférieure correspondent donc respectivement au maximum et au minimum des valeurs des frais présentes dans l'échantillon. Elles valent 46 599.108$ et 1 621.340$, alors qu'elles valaient 33 856.92$ et 1 136.399$ dans la boîte à moustache de la population. Bien que leurs quartiles Q1 soient semblables, on peut clairement observer que le quartile Q3 de l'échantillon est plus élevé que celui de la population.

iii.

Pour réaliser cette question, nous avons implémenté une fonction *functionFreqCum.m*, qui se charge de calculer la fréquence relative cumulée pour un y donné, soit

$$\widehat{F}_x(y) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i < y)$$

Cette fonction permet de construire, dans *ks_distance.m*, un vecteur contenant les fréquences relatives cumulées de la population, avec un pas de 0.5 pour les frais, un autre vecteur construit de la même manière pour l'échantillon et un dernier vecteur contenant la valeur absolue de la différence entre les valeurs des deux vecteurs précédents 2 à 2.

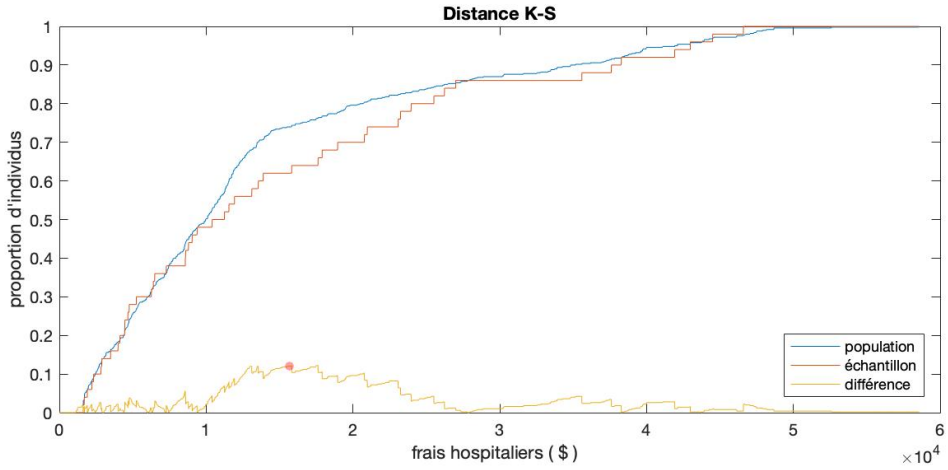

FIGURE 7

Cette figure ci-dessous présente donc, en orange, le polygone des fréquences cumulées des frais hospitaliers de l'échantillon, en bleu, celui de la population et en jaune, la valeur absolue de la différence entre les deux. On remarque que les 2 premières courbes sont fort similaires et que la courbe jaune est plutôt stable et basse, ce qui signifie à nouveau que l'échantillon suit la même loi que la population, et qu'il la représente plutôt fidèlement.

La valeur de la distance de Kolmogorov Smirnov est représentée par un point rouge. Elle vaut 0.1220 et correspond à la plus grande différence entre les polygones des fréquences cumulées, c'est la valeur retournée par *ks_distance.m*.

2.b

i. On crée un échantillon i.i.d. de patients de taille 50 grâce à *iid_sample.m* et on sauvegarde la moyenne des charges de cet échantillon retournée par *sample_sum_up.m* dans une nouvelle variable *mean_charge_samples*. Après avoir répété ce processus 499 autre fois grâce à une boucle et concaténé les moyennes des 499 échantillons à la nouvelle variable *mean_charge_sample* , on génère l'histogramme (grâce à *histogram.m*) de la nouvelle variable contenant à présent les moyennes des 500 échantillons. L'allure de cet histogramme ressemble à celle d'une loi normale. Cela a du sens car selon le théorème central-limite, la moyenne d'échantillon suit une loi

proche de la loi normale centrée en la moyenne de la population lorsque la taille de l'échantillon est grande (ce qui est le cas ici).
La moyenne des 500 moyennes contenues dans la nouvelle variable est de 13 619.789$. Elle est proche de la moyenne des frais de la population qui vaut 13 534.78.

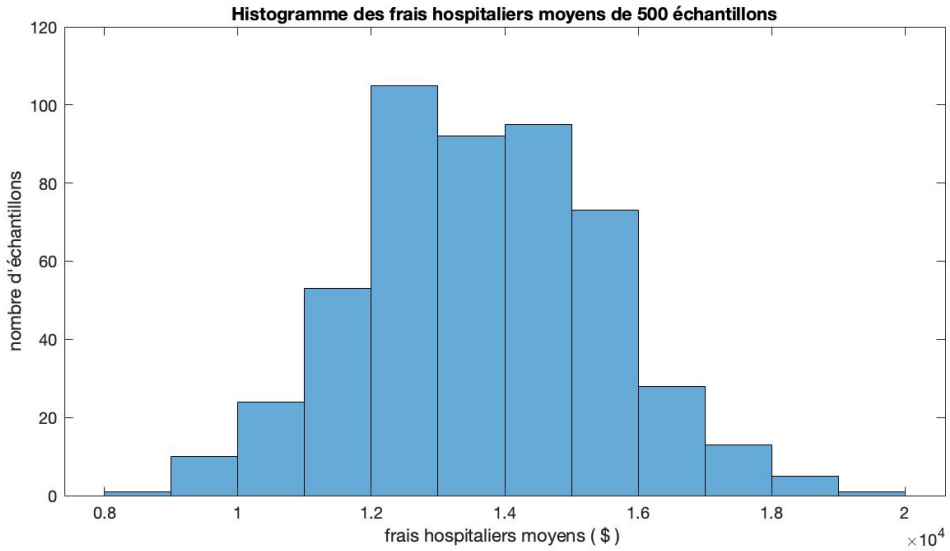

FIGURE 8

ii. On calcule la médiane des charges de la même manière qu'au point précédent, en appelant *sample_sum_up.m* à chaque échantillon créé. On sauvegarde la médiane de chacun des 500 échantillons dans une nouvelle variable *median_charge_samples* dont l'histogramme ne nous fait pas penser à une loi théorique connue, mais la loi à laquelle l'histogramme ressemble le plus serait une loi normale (il ressemble moins à une loi normale que l'histogramme des moyennes des 500 échantillons).

La moyenne de cette nouvelle variable vaut 9880.0784$ tandis que la médiane des frais hospitaliers de la population vaut 9977.548 $. La moyenne des médianes des 500 échantillons est plutôt éloignée de la médiane des frais de la population comparée à la moyenne calculée au point précédent qui est assez proche de la moyenne de la population. Cela a du sens car, bien qu'elles soient toutes deux des estimateurs non-biaisés, la moyenne d'un échantillon est plus précise que la médiane d'un échantillon car elle est de plus faible variance.

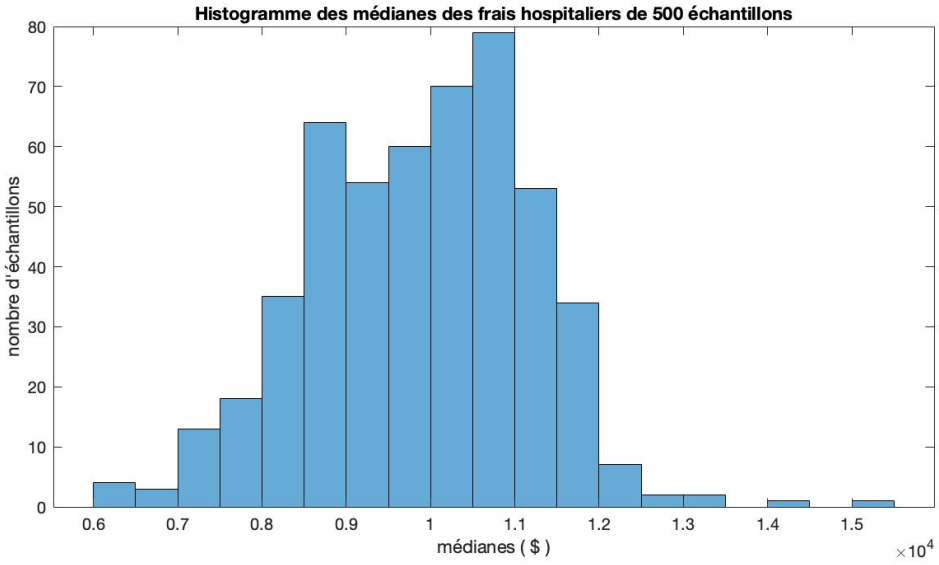

FIGURE 9

iii. On sauvegarde dans une nouvelle variable les écarts-types retournés par *sample_sum_up.m* appelée sur chacun des 500 échantillons de 50 patients créés. L'histogramme de cette nouvelle variable a l'allure d'une loi normale et sa moyenne vaut 12 111.3816$. Cette moyenne est proche de l'écart-type des frais de la population qui vaut 12 117.39 $. La moyenne des écarts-types, celle des médianes et celle des moyennes des 500 échantillons sont toutes plus proches des écart-type, médiane et moyenne de la population que l'écart-type, la médiane et la moyenne des frais d'un seul échantillon (cfr 2.*a*). Étudier plusieurs échantillons d'une population permet de mieux se rapprocher des véritables écart-type, médiane et moyenne de la population.

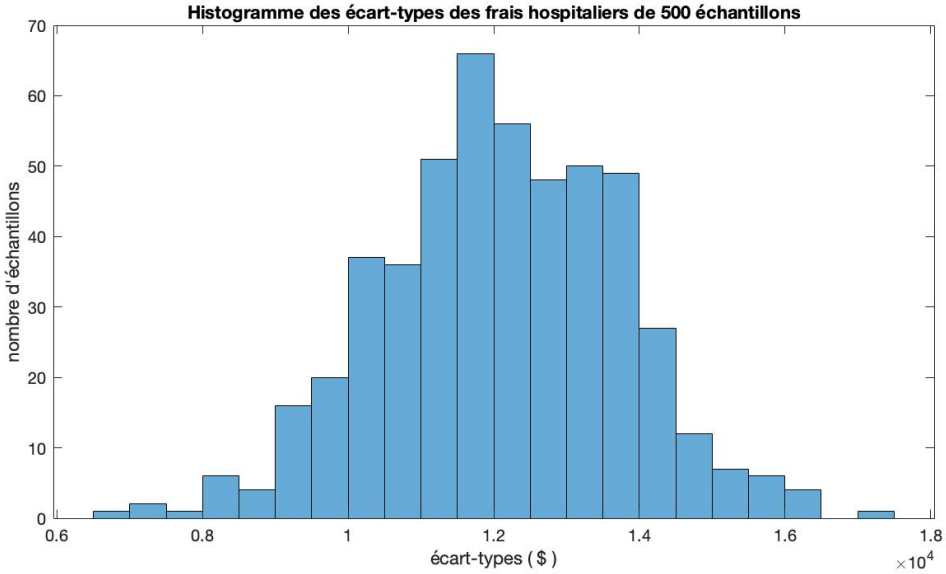

FIGURE 10

iv. On calcule la distance de Kolmogorov entre les polygones des fréquences cumulées de la population et un échantillon grâce à la fonction *ks_distance.m*. On l'appelle sur chacun des 500 échantillons et on conserve chaque résultat dans une nouvelle variable.

On peut apercevoir que l'histogramme de cette nouvelle variable a l'allure d'une loi normale non-symétrique, contrairement aux histogrammes des moyennes et des écarts-types qui étaient symétriques. Les distances d'occurrence les plus élevées sont comprises entre 0.1 et 0.12, ce qui veut dire que la majorité des échantillons ont leurs distance maximale entres leur polygone des fréquences cumulées et celui de la population comprise entre 0.1 et 0.12.

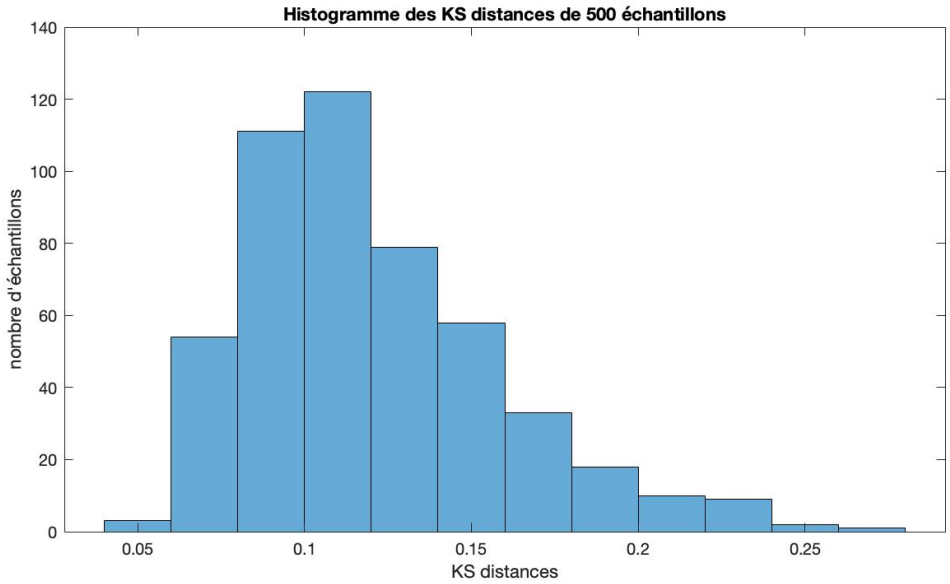

FIGURE 11

3 Estimation

Chaque réponse aux sous-questions de cette question est appuyée par une partie de code dans le script **Q3.m**

3.a

Dans le script **Q3.m** qui correspond à cette question, on tire 100 échantillons de 50 patients grâce à *iid_sample.m*. Pour chaque échantillon, on ne retient que la valeur qui nous intéresse ici, soit la valeur x du BMI. On calcule ensuite la moyenne de ces BMI (avec *sample_sum_up.m*), qui représente notre estimateur m_x et on conserve ces 100 estimateurs dans une nouvelle variable. Alors, pour calculer le biais de m_x, on calcule la différence entre la moyenne de cette nouvelle variable et la moyenne $\mu_{\mathcal{X}}$ du BMI pour la population. On calcule la variance de cette nouvelle variable avec la fonction *var.m*. Pour un tirage particulier [1] de 100 échantillons, on trouve donc :

$$Biais_{mean} = E\{m_x\} - \mu_{\mathcal{X}} = 30.6382 - 30.6998 = 0.0616$$

$$V\{m_x\} = 0.5714$$

3.b

De la même manière, après avoir tiré 100 échantillons de 50 patients et retenu les valeurs des BMI, on a calculé la médiane, $median_x$, de ces BMI pour chaque échantillon. On a ensuite calculé la moyenne de ces 100 médianes et on l'a comparée à la vraie médiane de la population, $median_{\mathcal{X}}$. On calcule la variance de cette nouvelle variable avec la fonction *var.m*. On obtient, pour un certain tirage de 100 échantillons :

$$Biais_{median} = E\{median_x\} - median_{\mathcal{X}} = 30.3974 - 30.4975 = 0.1001$$

$$V\{median_x\} = 1.1527$$

3.c

En répétant les deux points précédents avec des échantillons de 100 patients, on trouve désormais un biais pour la moyenne qui vaut 0.0360 et une variance correspondante qui vaut 0.3189 et un biais pour la médiane qui vaut 0.0474 et une variance correspondant qui vaut 0.5759.

En répétant l'expérience pour 100 autres échantillons de taille 100, on obtient des résultats similaires. Cependant, on constate que les biais ne deviennent pas systématiquement inférieurs à ceux qu'on trouve pour des échantillons de taille 50. On a donc décidé d'effectuer plusieurs fois chaque cas pour pouvoir mieux interpréter nos résultats. On trouve, au cours de 10 expériences :

1. Chaque tirage de 100 échantillons donne des résultats qui diffèrent légèrement

taille n = 50				taille n = 100			
$Biais_{mean}$	V_{mean}	$Biais_{median}$	V_{mean}	$Biais_{mean}$	V_{mean}	$Biais_{median}$	V_{mean}
0.0616	0.5714	0.1001	1.1527	0.0360	0.3189	0.0474	0.5759
0.1014	0.6256	0.1449	0.7022	0.0385	0.3202	0.0499	0.6553
0.0950	0.7071	0.0127	1.4262	0.0537	0.3905	0.0125	0.5127
0.1280	0.6200	0.1261	0.9761	0.0289	0.4320	0.0013	0.5952
0.0460	0.6936	0.1119	1.1349	0.0176	0.3367	0.0633	0.5379
0.0673	0.7721	0.0944	1.0675	0.0713	0.4342	0.1097	0.6830
0.0081	0.5998	0.0938	1.0796	0.0473	0.3609	0.0637	0.5652
0.0485	0.8507	0.0550	1.3358	0.0280	0.3637	0.0113	0.5197
0.0869	0.7098	0.0478	1.4713	0.0413	0.3857	7.5e-04	0.6512
0.0248	0.5790	0.0078	0.8732	0.0969	0.4452	0.0720	0.7376
moyennes				**moyennes**			
0.06676	0.67291	0.07945	1.12195	0.04595	0.3788	0.04318	0.55154

On voit que nos résultats moyens au cours de 10 expériences pour des échantillons de taille 100 sont bien moins élevés que ceux pour des échantillons de taille 50. Ceci confirme bien la théorie de l'estimation statistique qui vise essentiellement à "construire des estimateurs de faible biais, de faible variance, et tels que les deux diminuent aussi rapidement que possible lorsque la taille n de l'échantillon augmente." En effet, lorsque les estimateurs sont calculés sur des échantillons de plus grande taille, ils sont plus précis.

3.d

À l'intérieur d'une boucle qui itère 100 fois, on tire un échantillon de 20 patients grâce à *iid_sample.m*. Pour chaque échantillon tiré, on calcule deux intervalles de confiance à l'aide des fonctions *student_interval.m* et *gauss_interval.m* :

La première fonction, *student_interval.m*, construit un intervalle de confiance à 95% pour la variable parente de x à partir d'un vecteur de valeurs de la variable x pour chaque individu d'un échantillon. On y calcule la moyenne de ce vecteur (avec *sample_sum_up.m*), soit dans notre cas m_x, la moyenne des BMI sur un échantillon. On calcule ensuite les bornes de l'intervalle de confiance à partir de la formule suivante :

$$m_x - t_{1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq m_x + t_{1-\frac{\alpha}{2}} \frac{s_{n-1}}{\sqrt{n}}$$

où

— $\alpha = 1 - 0.95 = 5\%$, la probabilité que la variable parente se trouve dans l'intervalle ;

— σ = écart-type des BMI de la population, considéré comme inconnu ;

— $\mu = 30.6998$ kgm^{-2} est la véritable moyenne des BMI de la population ;

— $n = 20$, la taille de l'échantillon ;

— s_{n-1} est l'écart-type corrigé de l'échantillon, soit $\sqrt{\frac{n}{n-1}} s_n$ où s_n est renvoyé par *sample_sum_up* ;

— $t_{1-\frac{\alpha}{2}} = t_{0.975}$ dépend du nombre de degrés de liberté. Comme $n = 20$, on a $n - 1 = 19$ degrés de liberté.
On peut déterminer $t_{0.975}$ grâce à la fonction *icdf.m* de Matlab dans laquelle nous spécifions la loi de probabilité 'Gauss' en paramètre. Elle nous renvoie 2.093. On peut vérifier cette valeur grâce à la Table 5 du formulaire d'examen du cours : on

regarde à la ligne correspond à 19 degrés de liberté et on trouve bien $t_{0.975} = 2.093$

La deuxième fonction, *gauss_interval.m*, construit également un intervalle de confiance à 95%, mais en utilisant la loi de Gauss. On y calcule aussi la moyenne m_x du vecteur x contenant les BMI de l'échantillon considéré et les bornes de l'intervalle de confiance qui, vu que la taille des échantillons est de 20 (< 30), sont donnés par :

$$m_x - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq m_x + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

où

— α, n et μ sont les mêmes que dans le calcul des bornes selon la loi de Student ;

— $\sigma = 5.9075$ kgm^{-2} est l'écart-type des BMI de la population, considéré comme connu (calculé par *sum_up.m*) ;

— $u_{1-\frac{\alpha}{2}} = u_{0.975}$ est déterminé grâce à la fonction *icdf.m* de Matlab dans laquelle nous spécifions la loi de probabilité de Student ('T') en paramètre. Elle nous renvoie 1.96. On peut vérifier cette valeur grâce à la table en dessous de la Table 3 du formulaire d'examen qui nous donne les valeurs de $u_{1-\frac{\alpha}{2}}$ pour les α les plus utilisés. On trouve bien $u_{1-\frac{\alpha}{2}} = 1.96$ pour α=5%

Nos proportions d'intervalle contenant la vraie moyenne varient selon les 100 échantillons sélectionnés par *iid_sample.m*. On construit plusieurs fois 100 échantillons de taille 20 et on obtient les proportions suivantes :

Student	**Gauss**
94%	94%
95%	96%
92%	91%
97%	97%
97%	95%

Proportions selon Loi de Student et Loi de Gauss

On voit aussi bien pour Student que pour Gauss que les proportions d'intervalles contenant la moyenne $\mu = 30.6998$ kgm^{-2} des BMI de la population oscillent autour de 95%. Comme on observe la même proportion (95%) pour les deux lois, on peut dire que faire l'hypothèse que l'écart-type σ de la population est connu (pour Gauss) ou inconnu (pour Student) n'a pas d'importance. On peut très bien utiliser l'écart-type corrigé s_{n-1} de l'échantillon au lieu de calculer σ pour déterminer les intervalles.
Après avoir construit des intervalles de confiance à *95%* du BMI en utilisant la Loi de Gauss, on trouve également une proportion de *95%* d'intervalles contenant la moyenne μ, ce qui veut dire que faire l'hypothèse que la parente est Gaussienne était raisonnable.

4 Tests d'hypothèse

4.a

Pour cette question, nous sommes amenés à tester l'hypothèse suivante :

H0 = « *Les frais hospitaliers des fumeurs sont en moyenne supérieurs de* $x = 23\ 948.2079\ \$$ [2] *aux frais hospitaliers des non-fumeurs.* »

L'hypothèse H1 n'ayant pas été précisée explicitement dans l'énoncé, nous avions le choix de considérer un test unilatéral ou bilatéral. Nous avons choisi d'effectuer un test bilatéral. L'hypothèse alternative H1 s'écrit donc :

H1 = « *Les frais hospitaliers des fumeurs* ***ne sont pas*** *en moyenne supérieurs de* $x = 23\ 948.2079\ \$$ *aux frais hospitaliers des non-fumeurs.* »

Pour effectuer le test avec un seuil de signification α qui vaut 5%, nous suivons la règle de décision suivante :

On suppose H0 vraie. On détermine un intervalle de probabilité IP_{H0} au niveau $1 - \alpha = 95\%$ autour de x , qui ne contient que des valeurs d'échantillon $x_{éch}$ favorables à H0 (c'est-à-dire des valeurs qui restent "proches" de 23 948.2079 $.) A l'extérieur de IP_{H0}, on ne trouve que des valeurs d'échantillon $x_{éch}$ favorables à H1, c'est-à-dire éloignées de 23 948.2079 $.

Pour implémenter ce test (dans le script **Q4.m**), on divise d'abord notre population en deux catégories : les fumeurs, dont les frais hospitaliers sont conservés dans la variable *charge_smoker_pop*, et les non-fumeurs, dont les frais hospitaliers sont conservés dans la variable *charge_nonSmoker_pop*.

Ensuite, on calcule la vraie valeur de x sur l'ensemble de notre population. On construit ensuite, l'intervalle de probabilité IP_{H0} autour de cette valeur selon :

$$x - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq x_{éch} \leq x + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

où :

— α, le seuil de signification vaut 5%

— $n = 50$, la taille d'un échantillon ;

— $u_{1-\frac{\alpha}{2}}$ est déterminé à partir des tables de Gauss :

On cherche :

$$1 - \frac{\alpha}{2} = P(\mathcal{Z} < u_{1-\frac{\alpha}{2}})$$

On trouve, grâce à la Table :

$$0.975 = P(\mathcal{Z} < 1.96)$$

2. Cette valeur correspond à la la différence entre la moyenne des frais hospitaliers des fumeurs et la moyenne des frais hospitaliers des non fumeurs.

On utilise en fait, dans notre code, la fonction *icdf* qui se charge de renvoyer la valeur de u recherchée.

— σ, l'écart type, est calculé à partir de la variance de la manière suivante :

$$\sigma\{F_{fum} - F_{nonFum}\} = \sqrt{V\{F_{fum} - F_{nonFum}\}}$$

On développe donc le membre de droite :

$$\begin{aligned}
&V\{F_{fum} - F_{nonFum}\} \\
&= E\{(F_{fum} - F_{nonFum})^2\} - E\{F_{fum} - F_{nonFum}\}^2 \\
&= E\{F_{fum}^2 - 2F_{fum}F_{nonFum} + F_{nonFum}^2\} - (E\{F_{fum}\} - E\{F_{nonFum}\})^2 \\
&= E\{F_{fum}^2\} - E\{2F_{fum}F_{nonFum}\} + E\{F_{nonFum}^2\} - E\{F_{fum}\}^2 - E\{F_{nonFum}\}^2 \\
&\quad +2E\{F_{fum}\}E\{F_{nonFum}\} \\
&= E\{F_{fum}^2\} - E\{F_{fum}\}^2 + E\{F_{nonFum}^2\} - E\{F_{nonFum}\}^2 \\
&= V\{F_{fum}\} + V\{F_{nonFum}\}
\end{aligned}$$

On calcule donc, dans notre code :

$$\sigma = \sqrt{V\{F_{fum}\} + V\{F_{nonFum}\}}$$

Une fois l'intervalle construit, on réalise 100 fois le test d'hypothèse, avec 100 échantillons différents (et le même x) :

Pour se faire, on utilise la fonction *iid.sample.m* (cfr. 2.a). On obtient, pour chaque test d'hypothèse, 1 échantillon i.i.d. de 50 patients fumeurs et 1 échantillon i.i.d. de 50 patients non-fumeurs. On calcule ensuite le $x_{éch}$, soit la différence entre la moyenne des frais pour l'échantillon des fumeurs et la moyenne des frais pour l'échantillon des non-fumeurs, et on vérifie enfin s'il appartient à l'intervalle IP_{H0}.

Le nombre de fois que l'hypothèse est rejetée, sur 100 tests, tourne autour de 5. Ce résultat correspond bien à la valeur de α qui nous indiquait qu'il y avait une probabilité de 5% de rejeter H0 alors qu'elle est vraie. On peut donc conclure que le pourcentage d'échantillons qu'on rejette correspond certainement à l'erreur de type-1 et que d'une manière générale, les résultats sur les 100 échantillons renforcent notre conviction que H0 est vraie.

4.b

Pour cette question, on opère de manière similaire à la précédente, sauf qu'ici on cible les patients de plus de 50 ans. On redivise donc notre population initiale en deux nouvelles sous-populations : les fumeurs de plus de 50 ans et les non-fumeurs de plus de 50 ans. On recalcule donc l'écart-type σ_{50} sur base des variances calculées sur les deux nouvelles sous-populations et on reconstruit un intervalle de probabilité autour du même x. On obtient un intervalle $IP_{50,H0} = [21168.06\,;\ 27258.23]$, qui ne contient que des valeurs des frais favorables à H0.

On construit donc des échantillons i.i.d. de 50 patients **de plus de 50 ans** et on calcule le $x_{éch}$ pour chacun d'entre eux. On obtient désormais environ 40 cas pour lesquels l'hypothèse est rejetée.

On comprend donc que les résultats des échantillons sont souvent trop éloignés de x que pour être dûs au seul hasard de l'échantillonnage. Ces résultats renforcent notre conviction que, dans le cas des patients de plus de 50 ans, l'hypothèse H0 n'est pas vraie, mais que H1 l'est.