Université de Liège

MATH0487-Eléments de statistiques

Elements de statistiques

Projet de statistiques

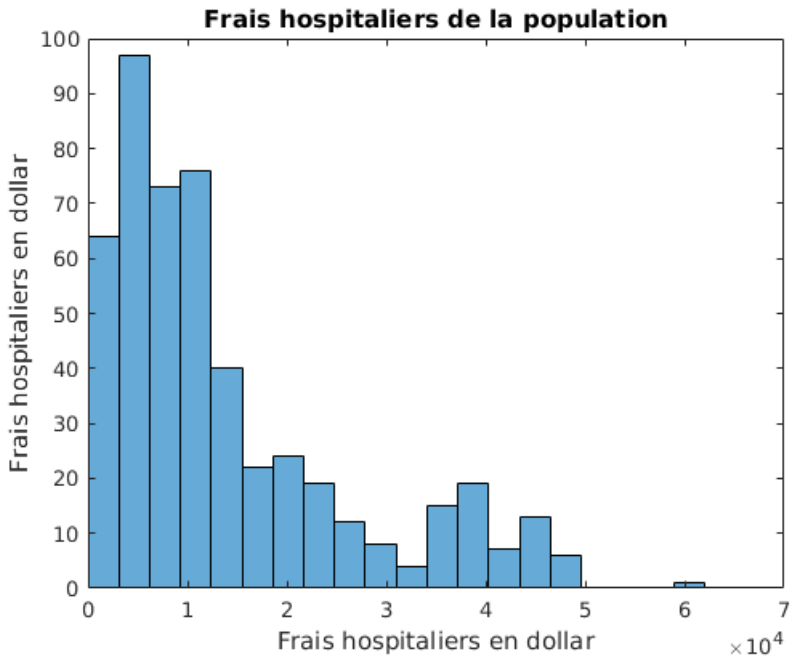
Auteurs:
Bernard Aurélien & Safadi Tasnim

December 6, 2019

Contents

1 Analyse descriptive

1.a Histogrammes



On constate que la valeur modale survient pour les frais compris entre approximativement 3 500$ et 7 000$. De plus, l'histogrammes illustre des données asymétriques à droite. En effet, la majorité des frais hospitaliers sont compris entre 0$ et 10 000$. De plus, on remarque la présence de données abbrérantes. Celles-ci sont reprises dans la barre isolée à l' extrémité droite.

1.b Moyenne, médiane, mode et écart-type

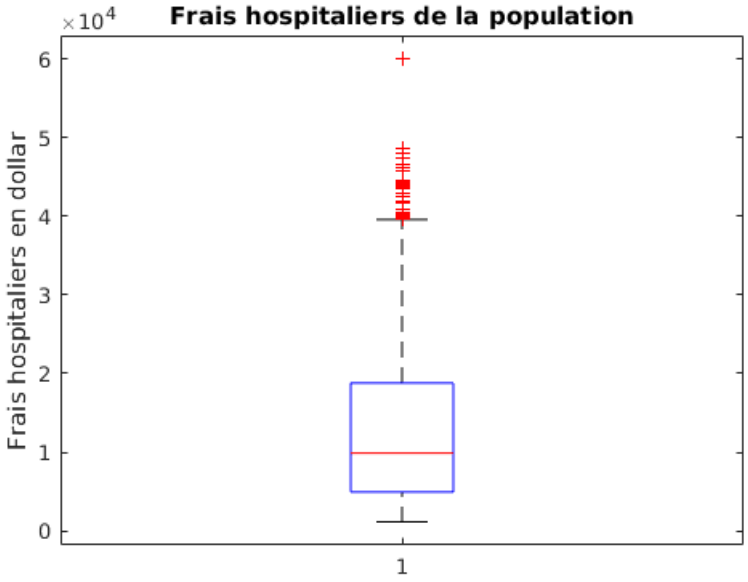
Moyenne	Médiane	Écart-type
13 955$	9 871$	12 253$

On remarque directement que la médiane a une différence assez important avec la moyenne. En effet, la moyenne est bien plus sensible aux données abbérantes que la médiane. Cette différence peut donc s'expliquer par la présence des valeurs abbérantes vues dans l'histogramme; les résultats sont dispersés. Ceux-ci expliquent aussi le fait que l'écart-type soit élevé. Les frais hospitaliers de Ms. Smith s'élèvent à 16 884.924$, ce qui est au dessus de la moyenne. Il s'agit également de frais suppérieurs à la majorité de la population; à savoir 71.8% de la population ont des frais inférieurs ou égals.

1.c Caractéristiques d'un taux normal

On considère comme normale au sens de la loi normale des frais hospitaliers compris dans l'intervalle [moyenne - écart-type; moyenne + écart-type]. Dans notre cas cet intervalle est [1 702$; 26 208$]. Dans notre population de 500 personne, 402 personnes se situent dans cet intervalle. Ce qui fait une proportion de 80,4% de patients ayant des frais hospitaliers normaux. On peut en conclure que les frais hospitaliers de Ms. Smith, qui s'élèvent à 16 884.924$, sont considèré comme normaux.

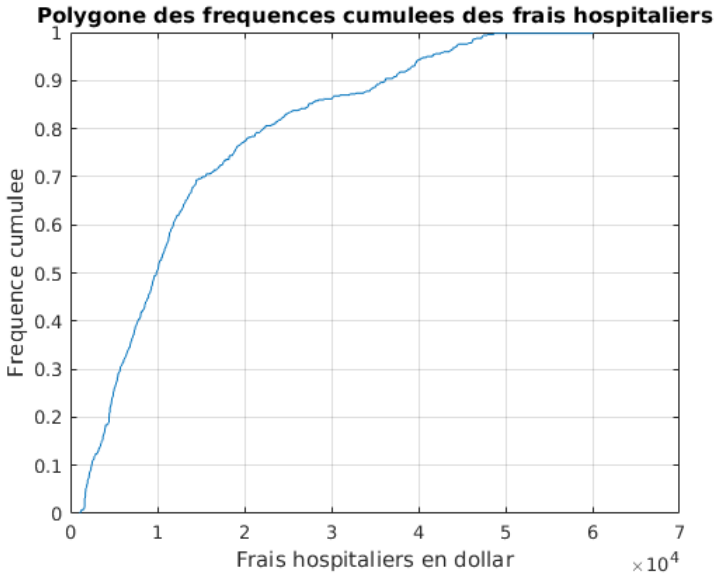
1.d Boîtes à moustache



Quartile 1	Quartile 2 (Médiane)	Quartile 3
4 929$	9 871$	18 786$

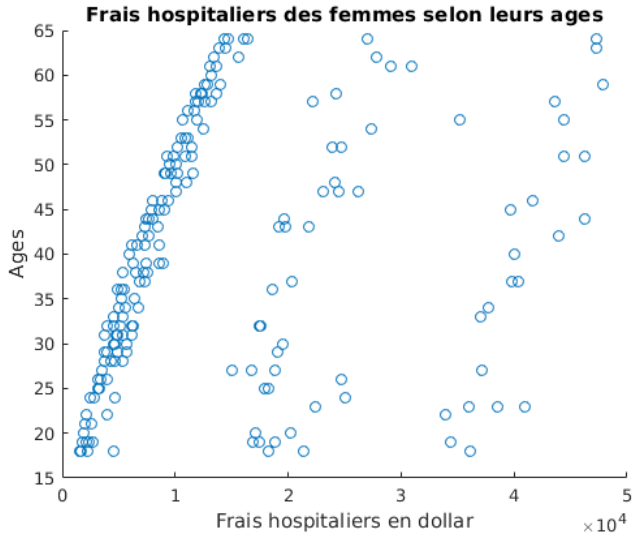
En regardant la boite à moustache, on voit très vite qu'il y a des données abbérantes. Pour qu'une donnée soit abbérante elle doit être soit inférieure à $Q^1 - \frac{1}{2}(Q^3 - Q^1)$ ou supérieure à $Q^3 + \frac{1}{2}(Q^3 - Q^1)$. Dans notre cas, ces valeurs valent respectivement -15 856.5$ et 39 571.5$. Puisqu'il s'agit de frais hospitaliers, il est impossible d'avoir des valeurs négatives. Dès lors, les frais hospitaliers sont au minimum 0$ et il serai impossible d'avoir une valeur égale à -15 856$. Grâce à l'histogramme et la boite à moustache, on sait, en effet, que plusieurs patients ont des frais hospitaliers supérieurs à 39 571.5$. La présence de plusieurs données abbérantes est donc cohérente par rapport à l'histrogramme. De plus, on retrouve aussi la personne ayant des frais supérieurs à 60 000$. Celle-ci était présente sur l'histrogramme dans la barre isolée a l'extrémité droite. En sommant le nombre de données abbérantes, on obtient que 81 personnes ont des frais hospitaliers au dessus de 39 571.5$.

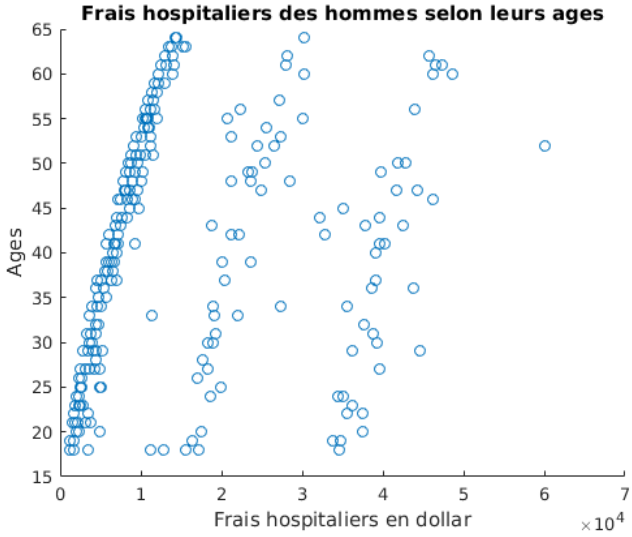
1.e Polygône des fréquences cumulées



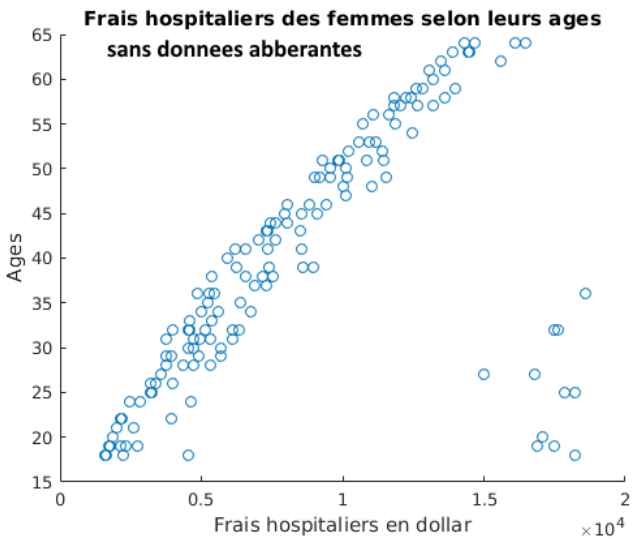
Parmi les patients, seuls 57 ont des frais hospitaliers inférieurs ou égaux à 25 000$ et supérieurs à ceux de Ms. Smith. Ce qui fait une proportion de 11.4% de la population. Grâce à la fonction proprotion, nous avons pu obtenir ce nombre simplement en faisant une différence entre le résultat de porportion sur les charges et 25 000$ et le résultat de proportion sur les charges et les frais hospitaliers de Ms. Smith.

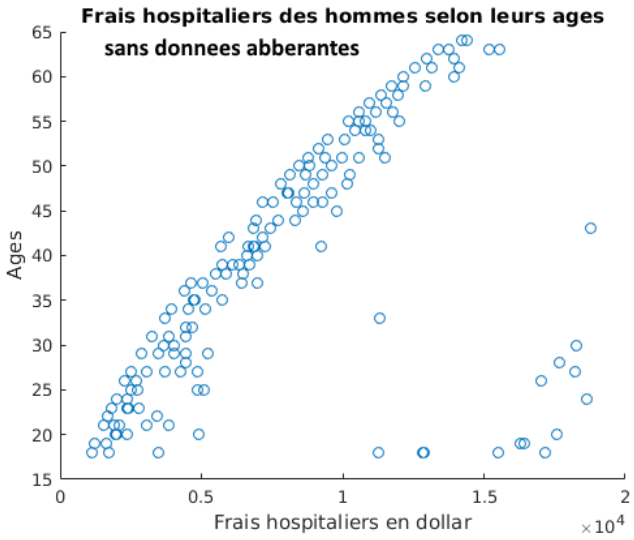
1.f Scatterplot





Le coefficient de correlation du scatter plot des frais hospitaliers des femmes et des hommes valent respectivement 0.2948 et 0.2752. On peut donc dire que la correlation est faible entre les frais et l'âge pour femmes comme les hômmes. Cependant ceci est peut-être dûe à la présentation de données abbérantes car normalement, ces deux variables ne devraient pas être si indépendantes. Nous avons donc décidé de retirer tous les frais hospitaliers supérieurs à 18 786$. On obtient les nuages de points suivants:





Cette fois le coefficient de correlation pour les femmes et les hommes valent respectivement 0.6672

et 0.6051. On peut donc en déduire qu'hormis les personnes ayant des frais hospitaliers supérieurs à 18 786$, généralement les frais augmentent avec l'âge et ce pour les hômmes et les femmes. En ce qui concernent les personnes dont les frais hospitalier dépassent 18 786$, elles souffrent probablement de plus gros soucis de santé.

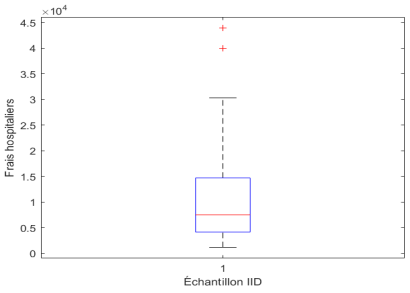
2 Génération d'échantillons i.i.d. de taille 50

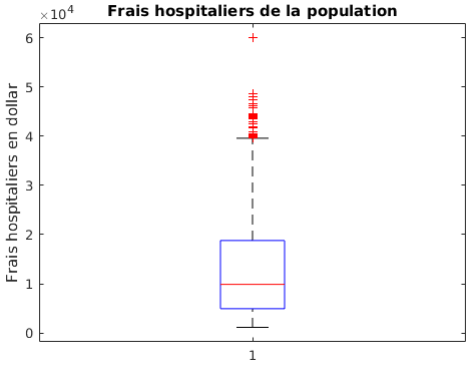
2.a Moyenne, medianne et écart type d'un échantillon

	Moyenne	Médiane	Écart-type
Population totale	13 955	9 871	12 253
Échantillon IID	11 149	7 493	10 018

Comme il peut être constaté, les trois statistiques calculées sont systematiquements plus petites pour l'échantillon iid que pour la population totale. Ceci peut être expliqué par la réduction dans la proportion de valeurs abbérantes par rapport aux valeurs normales. La plus part des valeurs se trouvant dans l'écart interquartille de la population, il est plus probable de tirer au hasard une valeur comprise dans cet écart. Ceci réduit effectivement la probabilité d'avoir des données abberantes aux quelles la moyenne, la mediane et l'écart type sont sensibles.

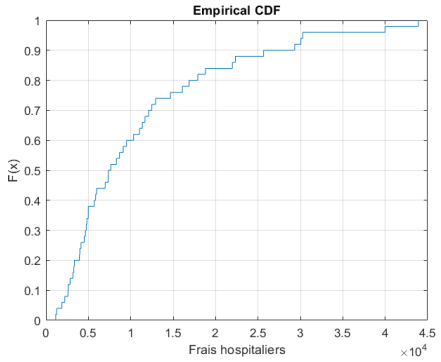
2.b Boite à moustache d'un echantillon





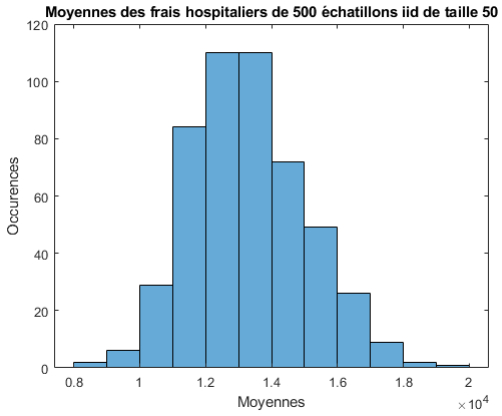
Les deux boites à moustache ci dessus ont à première vue une répartition similaire. On constate que les valeurs de la boîte correspondant à l'échantillon iid sont moins dispersées que la boîte de droite. De plus, ces boîtes montrent plus clairement ce qui a été dit au point précédant. La boîte à gauche contient un écart interquartille plus petit et une proportion réduite de données abbérantes comparé à la boîte de la population totale. Sur un échantillon de 50 observations nous observons deux données abbérantes, ce qui est proportionnellement inférieur à ce qui peut être observé dans la population totale. En fait, dans la population totale (contenant 500 observations), on retrouve plus de 20 données abbérantes.

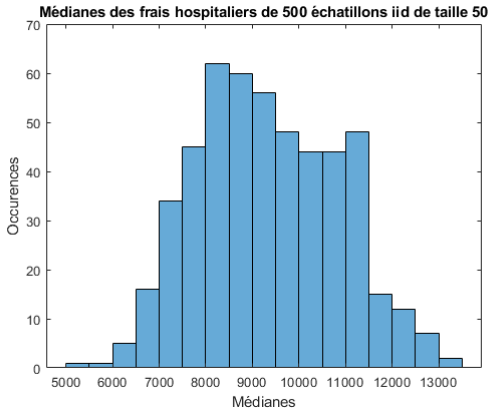
2.c Polygone des fréquences cumulées d'un echantillon

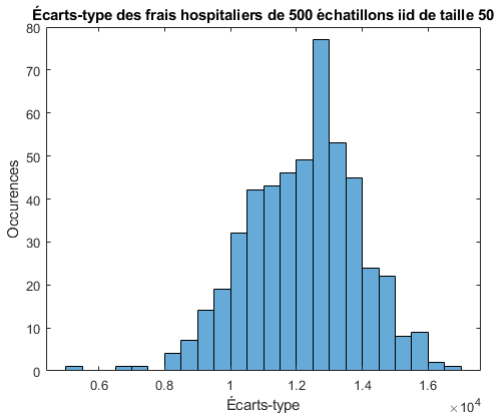


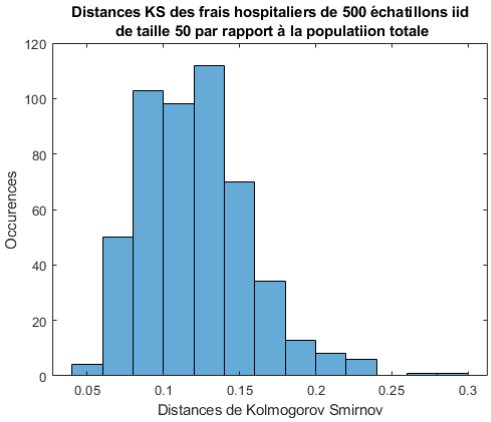
Le polygone des fréquences cumulées ci dessus est fort similaire au polygone de la population totale. La principale différence que l'on peut observer est le fait que, une fois de plus, un grand nombre de valeurs abbérantes sont absantes. Ceci a pour résultat que la courbe soit légérement plus aplatie horizontalement. La distance de Kolmogorov Smirnov entre la population et l'échantillon étant de 0.1060, il peut en être conlut que l'aspect des deux courbes sont fortement similaires.

2.d Statistiques sur 500 échantillons IID de taille 50









Moyenne des moyennes	Moyenne des médianes	Moyenne des écarts-type
13 303$	9 340.5$	12 228$

Comme nous pouvons le constater, les 4 histogrammes genrérés par les 500 échatillons iid ont une allure fortement similaire. Dans les 4 histogrammes nous pouvons voir la densité des valeurs distribuées autour d'une valeur centrale en faisant une forme de cloche. Cette forme est la forme distinctive d'une loi normale. Ceci est un comportement attendu car dans la population totale les valeurs se trouvent concentrées autour de la moyenne et la médiane. La probabilité de tirer aléatoirement un échantillon dont les valeurs sont loin des valeurs centrales est petite. Cette probabilité est réduite plus la valeur est loin des valeurs centrales de la population totale. Ce qui est une description de la loi normale.

Les moyenne des données utilisées afin de générer ces histogrammes sont trés proches des valeurs de la poulation totale. En particulier la valeur de l'écart-type. Cette dérnière ne s'éloigne que de 3$ de l'écart-type de la population totale. Ceci est probablement dû au fait que l'effet de tirer aléatoirement une grande quantité d'échantillons iid réduit l'effect de "ne pas avoir de chance" et tirer un échantillon abbérant par rapport aux valeurs de la population totale.

3 Estimation

3.a Moyenne d'échantillons de taille 50

Bias m_e	Variance m_e
0.0479	0.9369

Le bias est donné par la formule $\frac{1}{n}(\sum_{i=1}^{50} m_i) - m_x$ où les différents m_i sont la moyenne de chaque échantillon i. La variance de l'estimateur de m_x est donnée par $var(m_i)$ où var une fonction qui renvoie la variance des différents m_i.

3.b Médiane d'échantillons de taille 50

Bias $median_e$	Variance $median_e$
0.1236	1.1994

Le bias est donné par la formule $\frac{1}{n}(\sum_{i=1}^{50} median_i) - median_x$ où les différents $median_i$ sont la médiane de chaque échantillon i. La variance de l'estimateur $median_x$ est donnée par $var(median_i)$ où var une fonction qui renvoie la variance des différents $median_i$.

3.c Moyenne et médiane d'échantillons de taille 100

Bias m_e	Variance m_e	Bias $median_e$	Variance $median_e$
0.0449	0.3588	0.164	0.6468

Une première observation flagrante est la baisse significative des valeurs de la variance m_e et de la variance $median_e$. Ceci est directement lié à l'augmentation du nombre d'échantillons. En effet, plus le nombre d'échantillons augmente, plus la moyenne de ces échantillons se rapproche de celle des frais hospitaliers dans notre population. Nous obtenons donc des résultats plus concentrés et par conséquent des variances plus petites que celles obtenues pour la questions précédentes.

De plus, il est important de constater que la variance m_e a changé d'ordre de grandeur, se rapprochant plus de 0. Il en est de même pour la variance $median_e$ qui se rapproche également de 0. Par contre, les bias m_e et $median_e$ ne suivent pas ce même comportement, gardant le même ordre de grandeur qu'avant, indépendamment du nombre d'échantillons considérés.

En conclusion, nous venons d'observer que m_e et $median_e$ sont des estimateurs sans bias.

3.d Intervalle de confiance

Comme on s'intéresse à un intervalle de confiance à 95%, on a $1 - \alpha = 0.95$. On fixe donc $\alpha = 0.05$. De plus on utilise des échantillon de taille 20, d'où n = 20 et le nombre de ddl vaut $n - 1$, d'où 19.

3.d.i Loi de Student

Comme σ est inconnu et que $n \leq 30$, on peut utiliser la formule suivante:

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

En regardant dans la table des valeurs de t student, on trouve que pour une valeur 19 du nombre ddl et $t_{0.975}$ [1], on a la valeur 2.093. Cette valeur peut être obtenue grâce à la fonction tinv de MatLab. On sait aussi que:

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_x)^2$$

On peut donc calculer les bornes de l'intervalle de confiance pour chaque échantillon. On constate que 95 échantillons parmis nos 100 ont un BMI compris dans l'intervalle de confiance.

[1] 0.975 est simplement le résultat de $1 - 0.05/2$

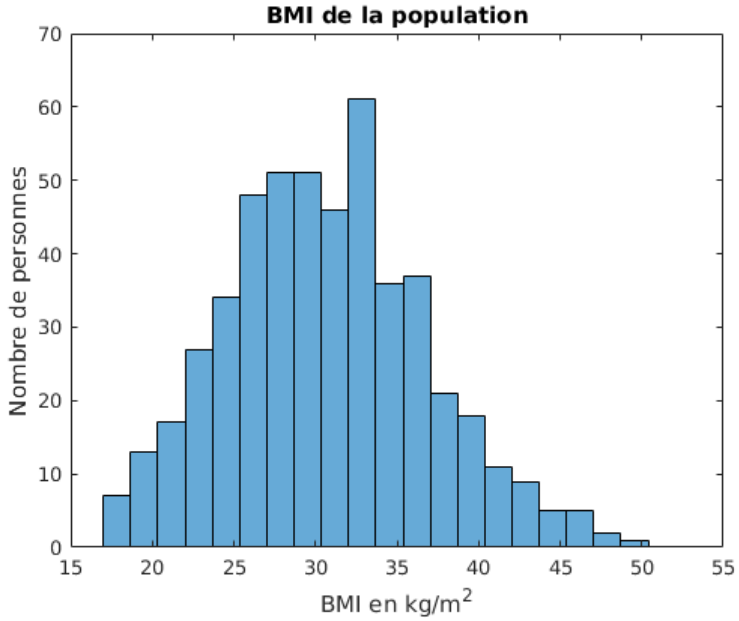
3.d.ii Loi de Gauss

Appliquons cette fois la formule suivante:

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

En regardant la table de Gauss on a $u_{1-\frac{\alpha}{2}} = 1.96$. Encore une fois, il est possible d'obtenir cette valeur grâce à une fonction fournie par MatLap; norminv. En appliquant donc la formule, on trouve que 97% des observations sont dans l'intervalle de confiance.

On peut donc conclure qu'il était raisonnable de supposer que la variable parente était Gaussienne. Cette supposition se confirme à l'aide de l'histogramme ci-dessous représentant le BMI de la population. En effet, il est possible de visualiser la forme de "cloche" typique d'une loi normale.



4 Tests d'hypothèse

Afin d'éféctuer le test d'hypothèse, nous faisons l'hypothèse que les frais hospitaliers sont répartis suivant une loi normale. De plus, la statistique dans le cas d'échantillons d'éparaillés $T_{\Delta x} = \frac{m_{\Delta x}}{s_{\Delta x}/\sqrt{n-1}}$ est utilisé.

Les hypothèses que l'on souhaite tester sont les suivantes:
Soit x la différence entre la moyenne des frais hospitaliers des fumeurs et la moyenne des frais hospitaliers des non-fumeurs dans notre population.
Soit l'hypothése null $h_0 =$ << Les frais hospitaliers des fumeurs sont en moyenne supérieurs de x aux frais hospitaliers des non-fumeurs>>
Soit l'hypothese alternative $h_1 =$ << Les frais hospitaliers des fumeurs sont en moyenne inférieurs ou égale a la somme des frais hospitaliers des non-fumeurs et x >>

4.a Population totale

Pour 100 échantillons iid de taille 50 h_0 est rejetée 75 fois. L'hypothése est donc rejetée 3 fois sur 4. Ceci ne permet donc pas de confirmer l'hypothèse posée par la companie d'assurance. Le choix de $\alpha = 5\%$ est fort stricte mais garanti que l'hypothèse ne soit pas rejetée à tort. Augmenter α de 5 à 10 ou à 15 pourcent augmente aussi le nombre de fois où l'hypothése null est acceptée. Cependant, cette augmentation n'est pas assez importante pour considérer que l'hypothèse null est verifiée.

4.b Population de plus de 50 ans

Pour 100 échantillons iid de taille 50 dans la population regroupant les personnes agées d'au moins 50 ans, h_0 est rejetée qu'une seul fois. Contrairement à la valeur précédente, celle-ci permet de confirmer l'hypothèse posée par l'assurance. Les effects nocifs de la cigarette sont beaucoup plus facile à apercevoir dans une population plus agée. Ceci est probablement dû à une exposition prolongée au tabac et ses effects au long terme.