Éléments de statistique

Rapport dans le cadre du projet

Pierre Hockers Olivier Moitroux
3e bac ingénieur

14 octobre 2018

Organisation générale du code

Nous avons organisé notre code sous forme de scripts attribués à chaque thématiques abordées dans ce projet. Nous avons donc 4 scripts couvrant l'analyse descriptive, la génération d'échantillon i.i.d., l'estimation et enfin les tests d'hypothèses. Ces différents scripts sont divisés en sections, lesquelles pouvant être exécutées une à une en utilisant le bouton "Run section" de l'environnement MATLAB. Il est bien sûr nécessaire d'avoir importé le fichier .csv au préalable.

1 Analyse descriptive

1.a

Le code nécessaire à cette section est repris dans le script `Q1.m` et est disponible à l'annexe 5.a.
En premier lieu, nous importons les données comprises dans le fichier *db_ stat85.csv* qui nous est attribué à l'aide de notre fonction `import_csv`. Chaque colonne du fichier sera alors importée sous forme de vecteur. Afin de se donner une idée des données contenues dans le fichier, nous avons représenté la consommation annuelle de bière et d'alcool fort de la population. En voici la représentation graphique :

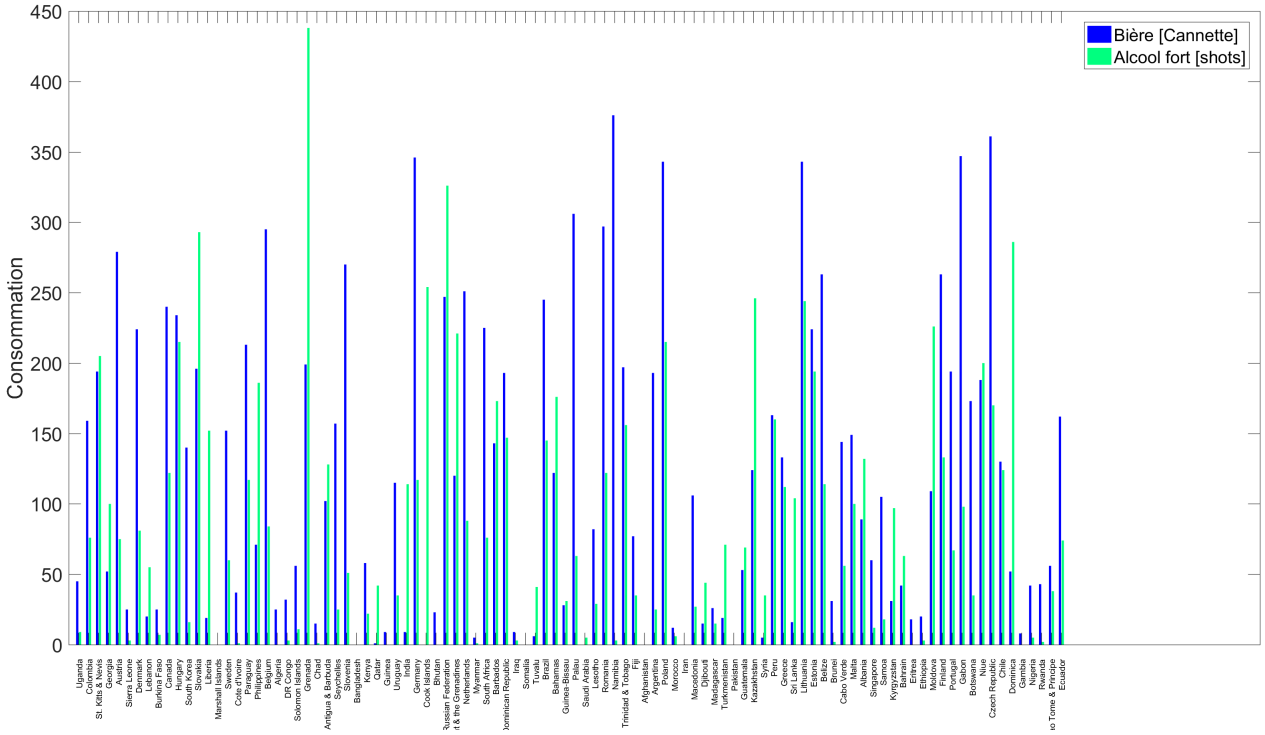

FIGURE 1 – Graphique en bâtonnet - Consommation alcool fort et bière

Au premier coup d'œil, il semble les deux consommations ne soient pas liées. On retrouve en effet des pays qui consomment des quantités semblables des deux boissons mais tout autant qui les consomment de façon disproportionnée. En d'autres termes, certains pays ont une préférence (marquée ou non) pour l'une des deux boissons, d'autres non. Toutefois, les pays ayant une consommation particulièrement faible dans une boisson, ont généralement aussi une consommation faible pour l'autre boisson.

Afin d'exploiter de manière plus pertinente ces données, il nous est proposé d'en dresser un histogramme. Ce dernier, généré à l'aide la fonction `hist` du module statistique de MATLAB, est représenté à la figure 2.

Nous remarquons que la hauteur des bâtonnets décroit de manière plus ou moins constante avec la consommation et que la majorité des pays ont une consommation comprise entre 0 et 50 doses d'alcool.

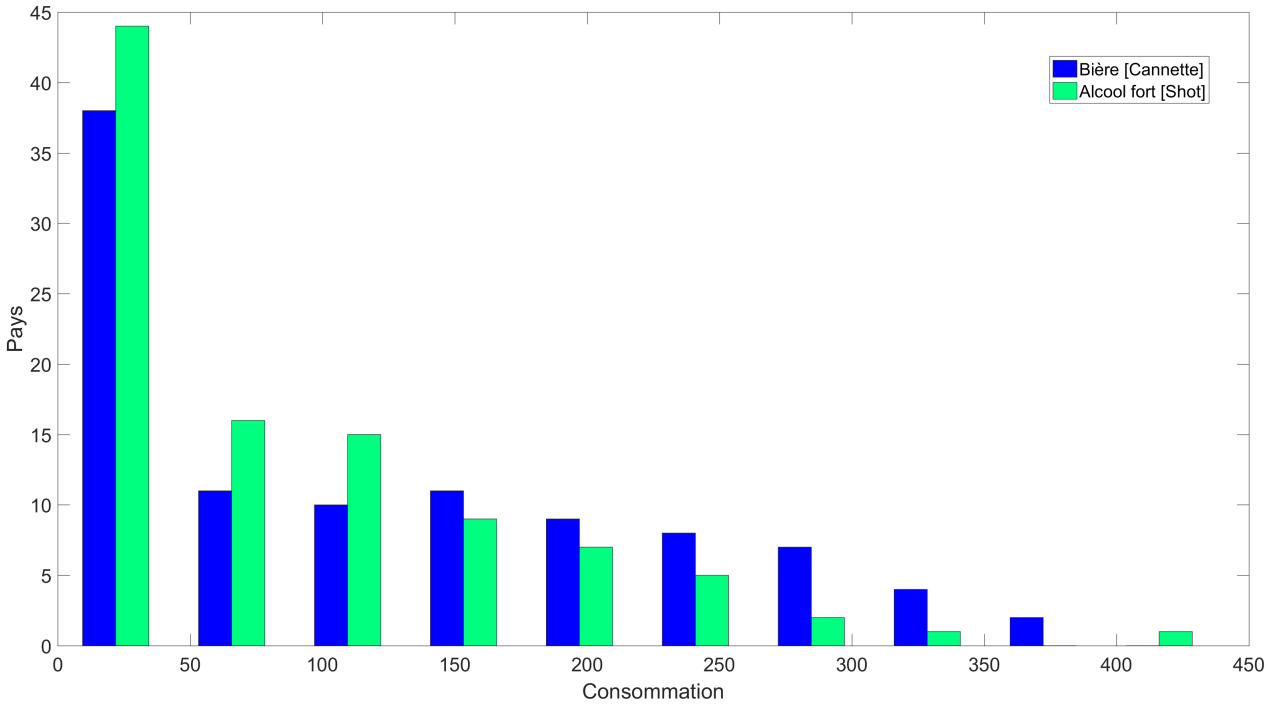

FIGURE 2 – Histogramme - Consommation alcool fort et bière

Plusieurs pays, ne consommant pas du tout (ou peu) d'alcool, sont repris dans le premier couple de bâtonnets et contribuent donc à sa démarcation par rapport aux autres. On remarque aussi que les pays ayant une consommation plus faible ont tendance à préférer l'alcool fort à la bière. En effet, les bâtonnets bleus prennent le dessus seulement à partir d'une consommation supérieure à 100 doses.

1.b

Nous avons calculé les éléments repris dans le tableau suivant (table 1) grâce au script `Q1.m`. Les unités sont les doses d'alcool habituelles, élevées au carré dans le cas de l'écart-type.

	Bière	**Alcool fort**
Moyenne mondiale	118.21	85.6100
Médiane mondiale	95.5	63
Mode mondiale	0.00	0.00
Écart-type mondial	107.4924	88.6401
Consommation belge	295	84

TABLE 1 – Tableau récapitulatif des consommations mondiale et belge en bière et alcool fort

Les médianes nous permettent de savoir que 50% des pays consomment moins de "95,5" cannettes de bière et de 63 shots d'alcool fort par an. Il peut être surprenant de remarquer que le mode mondial des deux boissons est de zéro. Cela s'explique facilement : plusieurs pays, ne consommant pas une goutte d'un certain type d'alcool, ont une consommation nulle dans cette catégorie, là où d'autres pays consommateurs ont peu de chance d'avoir exactement la même consommation qu'un autre pays consommateur pour l'une ou l'autre boisson.

Il ressort également de ces résultats que les belges boivent énormément de bière : leur consommation moyenne par an par personne est plus de deux fois plus élevée que celle mondiale, et est même au delà de la somme de la moyenne et de l'écart type mondial, ce qui signifie que leur consommation est anormale au sens gaussien. En revanche, la consommation belge d'alcool fort est très légèrement inférieure à la moyenne mondiale et donc loin d'être anormale au sens gaussien du terme.

1.c

Une consommation est dite *normale* si sa distance à la moyenne est inférieure à l'écart type. Le script `Q1.m` nous permet également de calculer la proportion de pays ayant une consommation "normale" de bière ou d'alcool fort. Les résultats sont repris à la table 2 :

	Proportion de pays ayant une consommation "normale"	**La consommation belge est-elle "normale" ?**
Bière	66 %	Non
Alcool fort	84 %	Oui

TABLE 2 – Consommations "normales"

On conclue assez vite que, comme dit plus haut, la Belgique a une consommation anormale de bière mais une consommation plus que normale de spiritueux.

1.d

Le script `Q1.m` nous permet toujours d'obtenir les boites à moustaches reprises à la figure 3 :

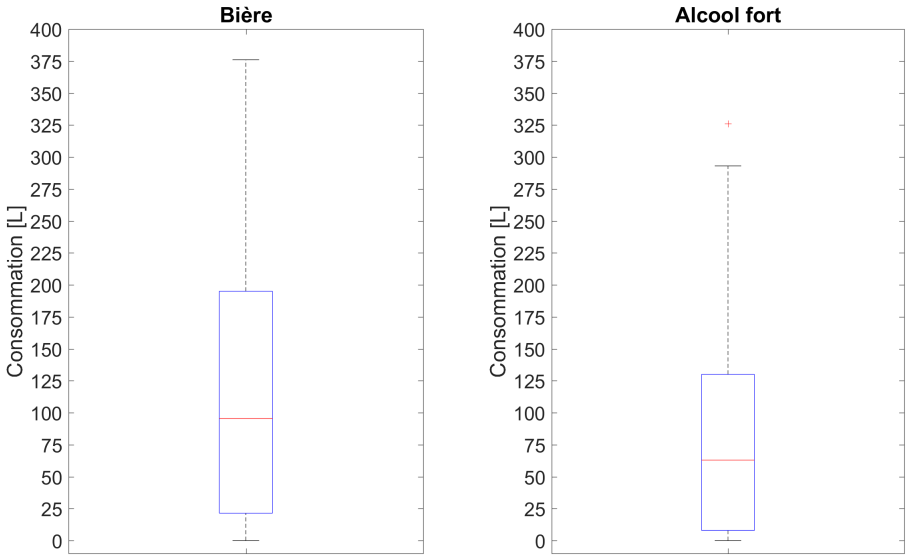

FIGURE 3 – Boîtes à moustaches

Les boites à moustache sont des rectangles délimités par le premier et le troisième quartiles et contenant la médiane (la ligne rouge). On ajoute alors des segments aux extrémités menant jusqu'aux valeurs les plus élevées.

Les valeurs extérieurs à la boite et aux segments qui apparaissent tout de même sont alors appelées valeurs aberrantes.

La croix rouge que l'on peut voir en dehors de la boite de l'alcool fort est une donnée aberrante. Les quartiles sont repris à la table 3 :

	Bière	**Alcool fort**
1er quartile	21.5	8
2ème quartile	95.5	63
3ème quartile	195	130

TABLE 3 – Quartiles

1.e

Il s'agit de tracer les verticales aux graduations des 200 cannettes et à celle de la consommation belge. La projection sur l'axe des ordonnées de l'intersection avec le graphe donne l'intervalle correspondant au critère. Il y a donc approximativement 13% des pays qui appartiennent à cette catégorie. Ce graphe nous a encore une fois été fourni par le script `Q1.m`. Il est visible à la figure 4.

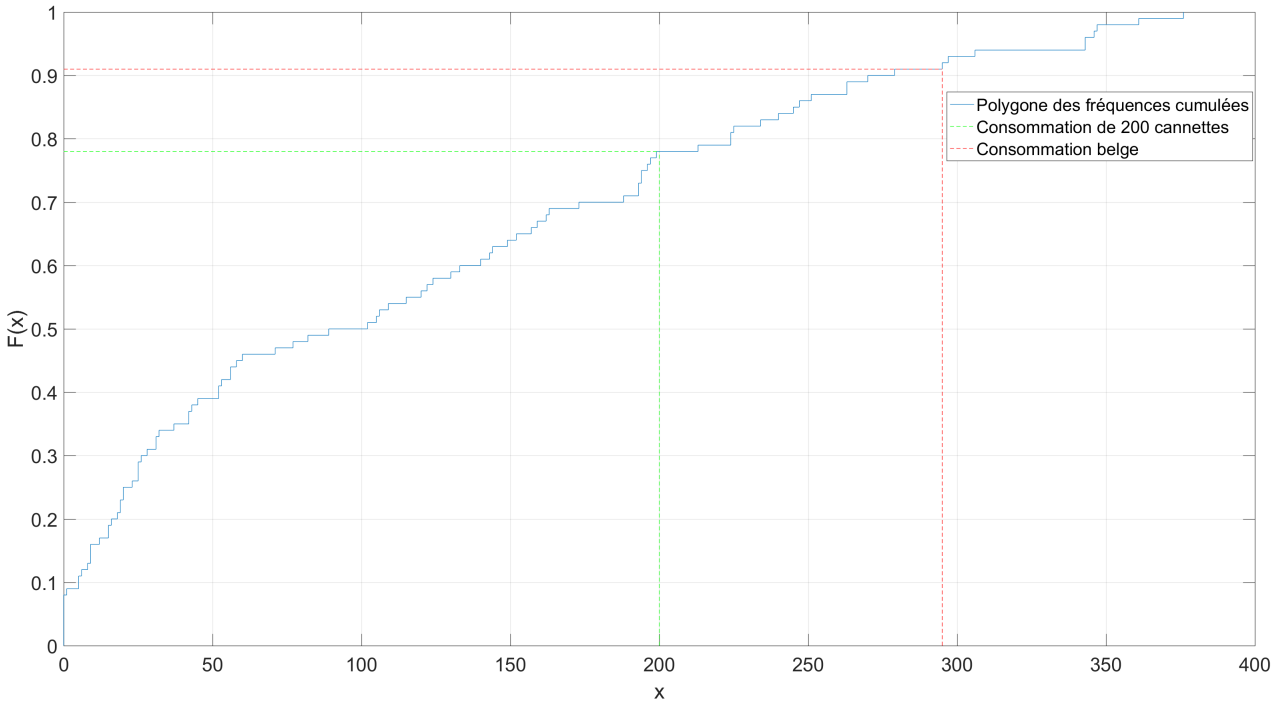

FIGURE 4 – Polygone des fréquences cumulées en bière de la population

1.f

Les résultats finaux de `Q1.m` nous donnent les trois scatterplots demandés repris à la figure 5. Les coefficients de corrélations correspondants sont repris à la table 4.

Il apparait que les consommations d'alcool pur et de bière sont étroitement corrélées : en effet, le coefficient correspondant est proche de 1. C'est moins vrai pour le vin et les spiritueux, mais on reste quand même plus proche du 1 que du 0, qui lui indique une corrélation inexistante. Cela est assez logique : toutes ces boissons contiennent de l'alcool. Donc, peu importe les autres facteurs, si vous consommez l'une d'entre elle, vous consommez du même coup de l'alcool pur. Ces coefficients ont été calculés à l'aide de `corrcoef`.

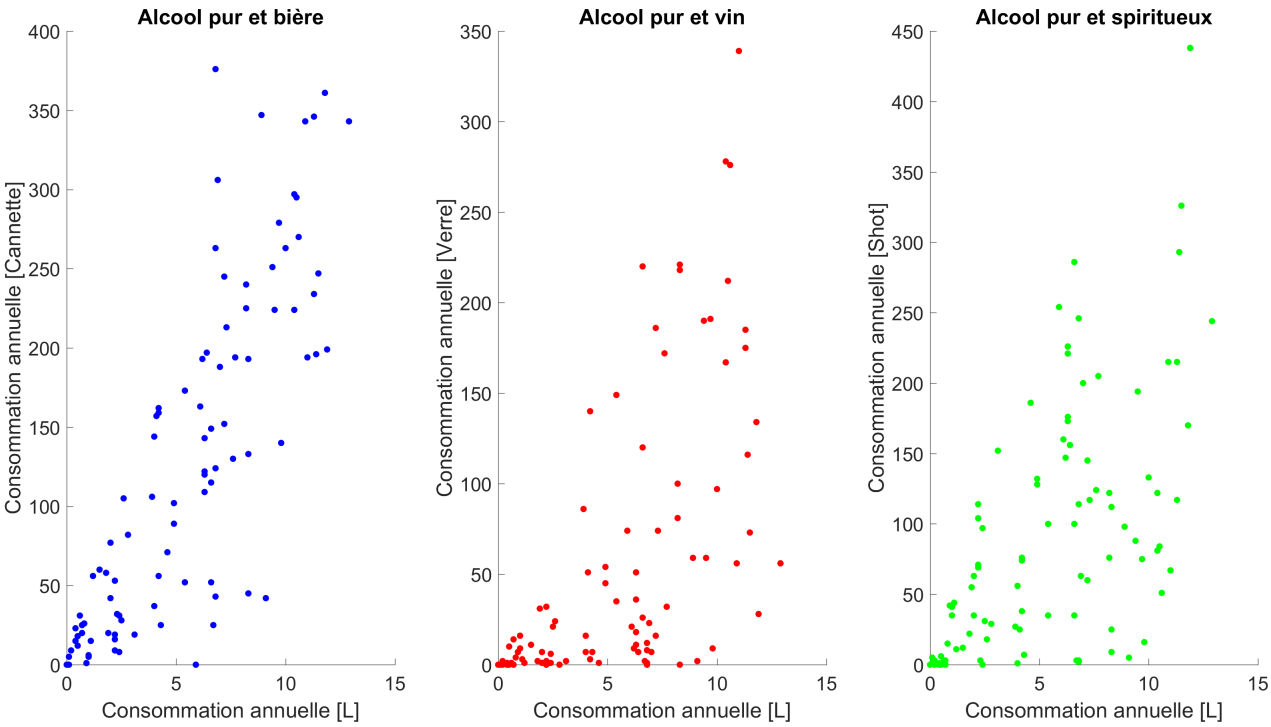

FIGURE 5 – Scatterplots du rapport entre les différentes consommations d'alcool et la consommation totale d'alcool pur

Alcool pur et bière	Alcool pur et vin	Alcool pur et spiritueux
0.833	0.6274	0.6190

TABLE 4 – Coefficients de corrélation

2 Génération d'échantillons i.i.d

2.a

Afin de tirer un échantillon i.i.d. de vingt pays, nous utilisons la fonction `randsample` en lui passant en argument les valeurs d'indexation des pays et le nombre de pays souhaité.

i) Le script `Q2.m` nous permet d'obtenir les résultats suivants, que nous avons mis à coté des résultats de la population pour plus de clarté :

Il est aisé de remarquer que les résultats dérivant de l'échantillon sont relativement proches de ceux de la population.

	Échantillon	
	Bière	Alcool fort
Moyenne	137.7	88.8
Médiane	97.5	86
Écart-type	117.9077	73.6554

TABLE 5 – Données sur l'échantillon

	Population	
	Bière	Alcool fort
Moyenne	118.21	85.61
Médiane	95.5	63
Écart-type	107.4924	88.6401

TABLE 6 – Données sur la population

ii) Les boîtes à moustaches correspondantes à notre échantillon sont reprises à la figure 7. On se rend compte qu'elles sont effectivement assez proches, même si la donnée absurde qui était présente sur la boite d'alcool fort de la population est passée à la trappe. La médiane de l'échantillon dans le cas de l'alcool fort est un peu supérieure à celle de la population tandis que les quartiles sont très proches de la population. A contrario, pour la bière, le troisième quartile a été surestimé.

Au final, les boites restent sensiblement à la même ordonnée et les quartiles sont dans les environs de ceux de la population, de même pour les médianes. Ces différences viennent du fait que l'échantillon est un peu petit pour représenter une population aussi vaste.

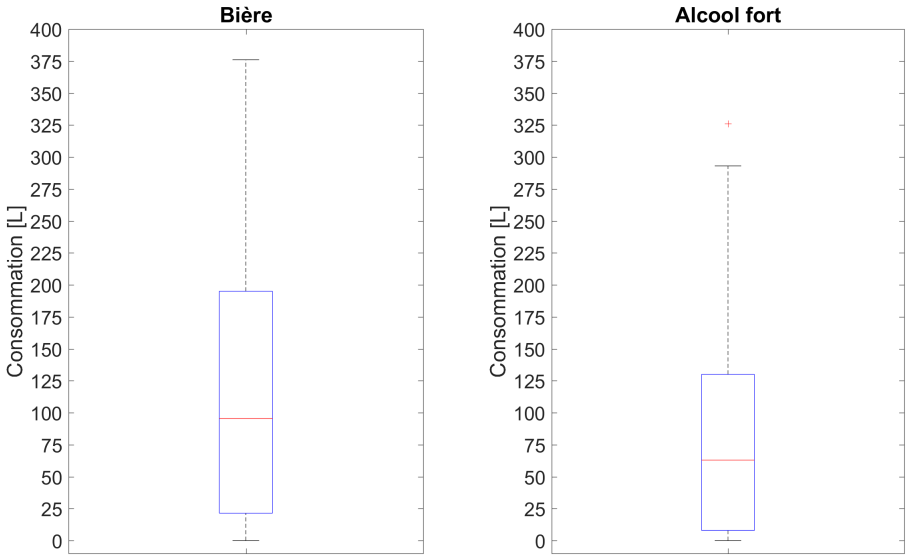

FIGURE 6 – Boîtes à moustaches de la consommation de bière et d'alcool fort de la population

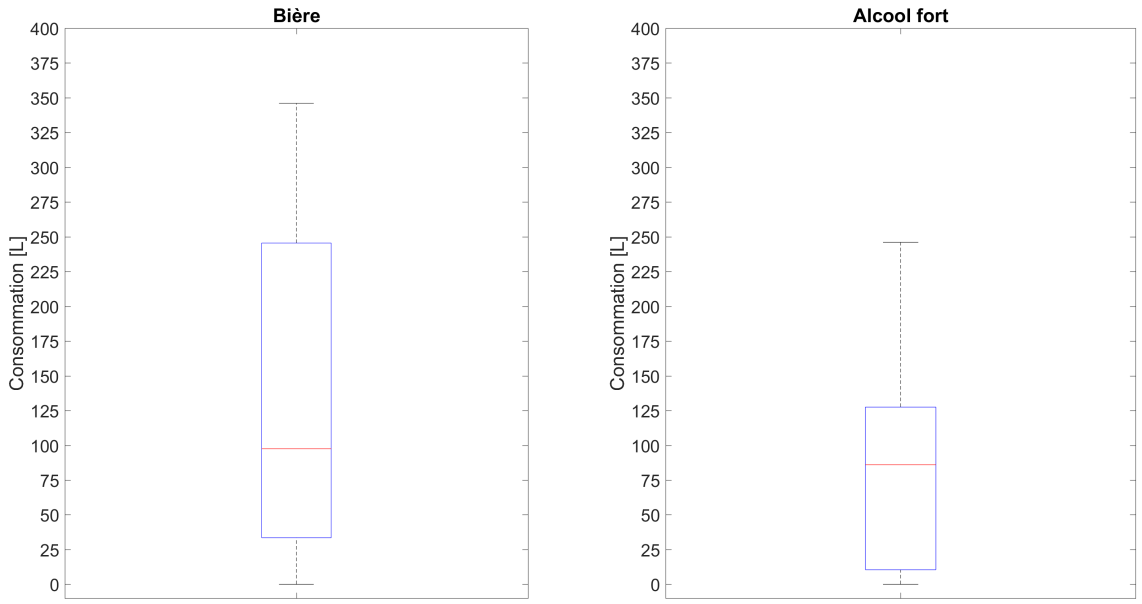

FIGURE 7 – Boites à moustaches de la consommation de bière et d'alcool fort de l'échantillon

iii) Les polygones des fréquences cumulées demandés sont présent à la figure 8. Les polygones se ressemblent assez, même si ceux de l'échantillon sont moins lisses, dû au nombre moindre de pays inclus. Les distances respectives pour la bière et les alcools forts sont de 0.1800 et 0.1700 (assez proches donc).

2.b

i) Les histogrammes (figure 9) ressemblent à des lois normales. La moyenne de la consommation de bière pour ces 100 tirages est de 118.0445, ce qui, comparé au 118.21 de la population est vraiment proche. La moyenne de la consommation de spiritueux basée sur les 100 échantillons est de 88.5925 contre les 85.61 de la population, donc encore une fois pas de grande différence.

ii) Les histogrammes, toujours générés par `Q2.m`, sont disponibles à la figure 10. Ils font penser à des lois normales, même si celui des alcools forts présente des gros sauts de valeurs qui n'ont pas lieu d'être dans une loi normale. La moyenne des médianes pour la consommation de bière est de 92.2000, contre 95.4 pour la population. De même, pour les alcools forts, le résultat est de 66.805 contre 63 pour la population. Les résultats sont donc assez proches.

iii) Les histogrammes des écarts types sont disponibles à la figure 11. Ils ont été générés par le script Q2. Celui qui concerne la bière ressemble à une loi normale. Cependant,encore une fois, son alter ego pour les alcools fort, s'il fait penser par sa forme global à une loi normale, en est beaucoup plus loin, il y a beaucoup d'alternance entre des valeurs très hautes et d'autres fort basses.

Les moyennes obtenues pour l'écart type de la consommation de bière et d'alcool fort sont respectivement de 106.2033 et de 88.5925. Les comparer aux 107.4924 et 88.6401 respectifs de la population nous montre que ces résultats sont très semblables.

iv) Les histogrammes des distances entre les polygones de fréquences cumulées de la population et des 100 échantillons sont repris à la figure 12.

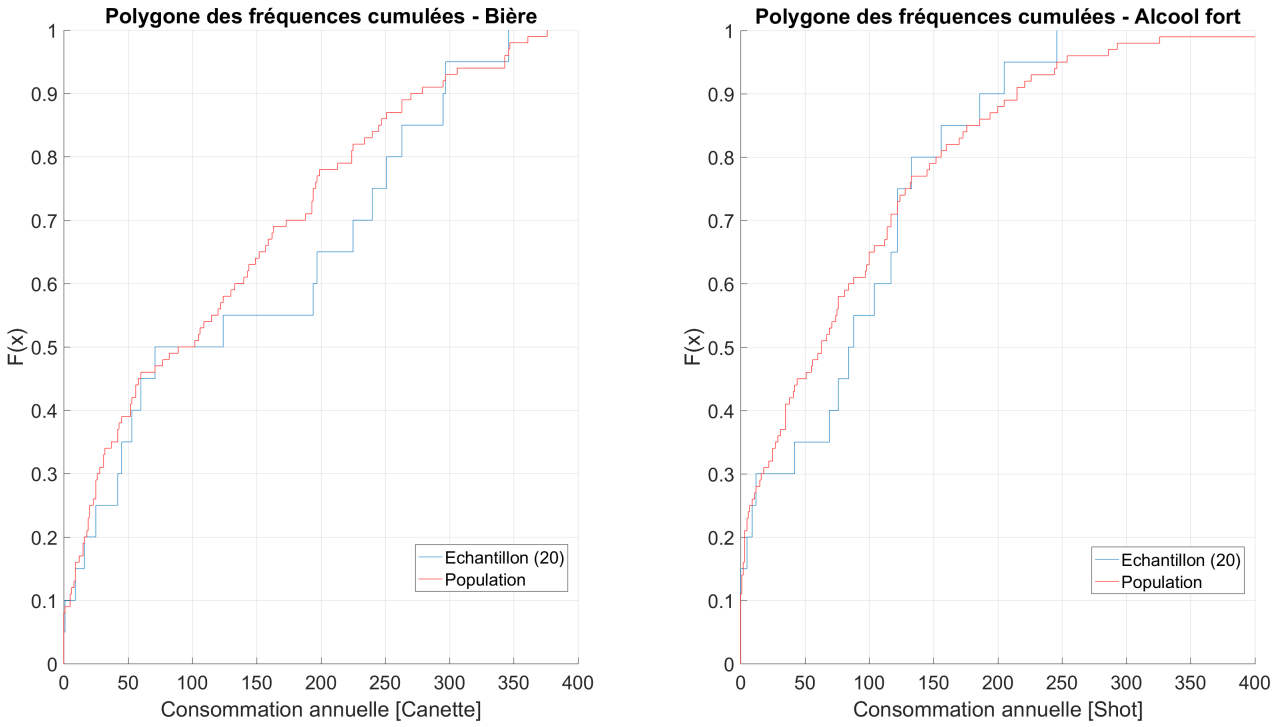

FIGURE 8 – Polygones des fréquences cumulées de la consommation de bière et d'alcool fort de l'échantillon et de la population

v) Les histogrammes des distances entre les polygones de fréquences cumulées de la population et des 100 échantillons pour chacun des alcools sont repris à la figure 13.

3 Estimation

3.a Échantillon de taille 20

Le script `Q3.m` nous a permis d'obtenir les résultats suivants :

— l'estimation du biais de l'estimateur m_x de la consommation moyenne de vin de la population vaut 0.7055

— l'estimation de la variance de l'estimateur $m_x$de la consommation moyenne de vin de la population vaut 238.9249

3.b

Nous obtenons comme estimations :

— le biais de l'estimateur $median_x$ de la consommation moyenne de vin de la population vaut 5.91

— la variance de l'estimateur $median_x$de la consommation moyenne de vin de la population vaut 190.8807

3.c Échantillon de taille 50

Nous trouvons, pour des échantillons de taille 50 :

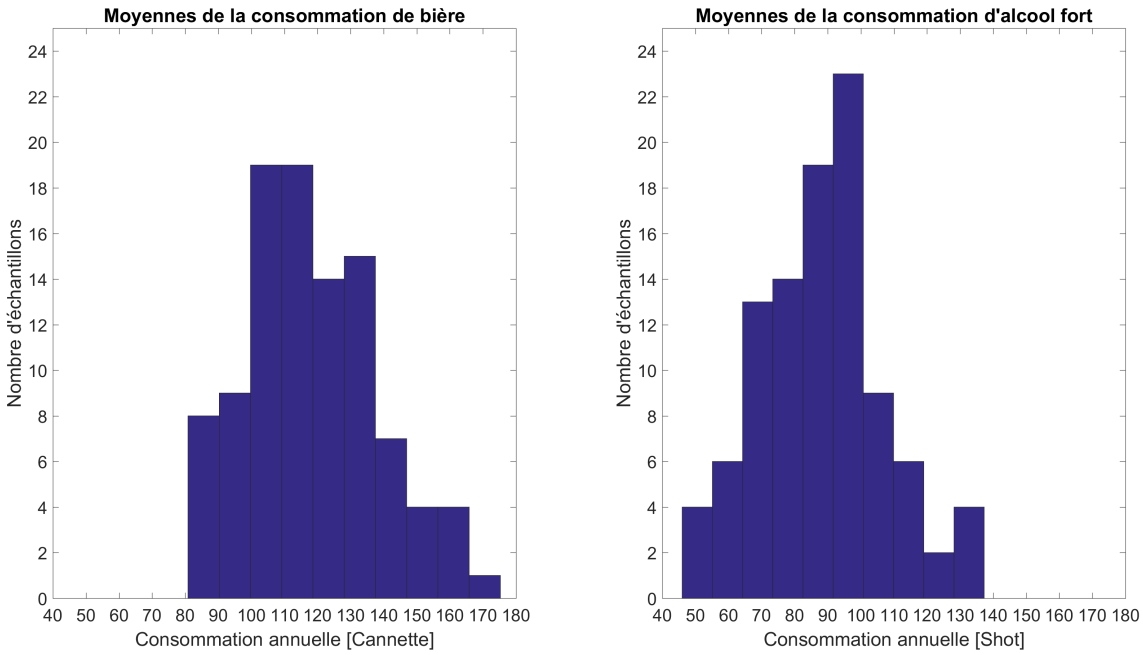

FIGURE 9 – Moyenne des 100 échantillons de 20 pays

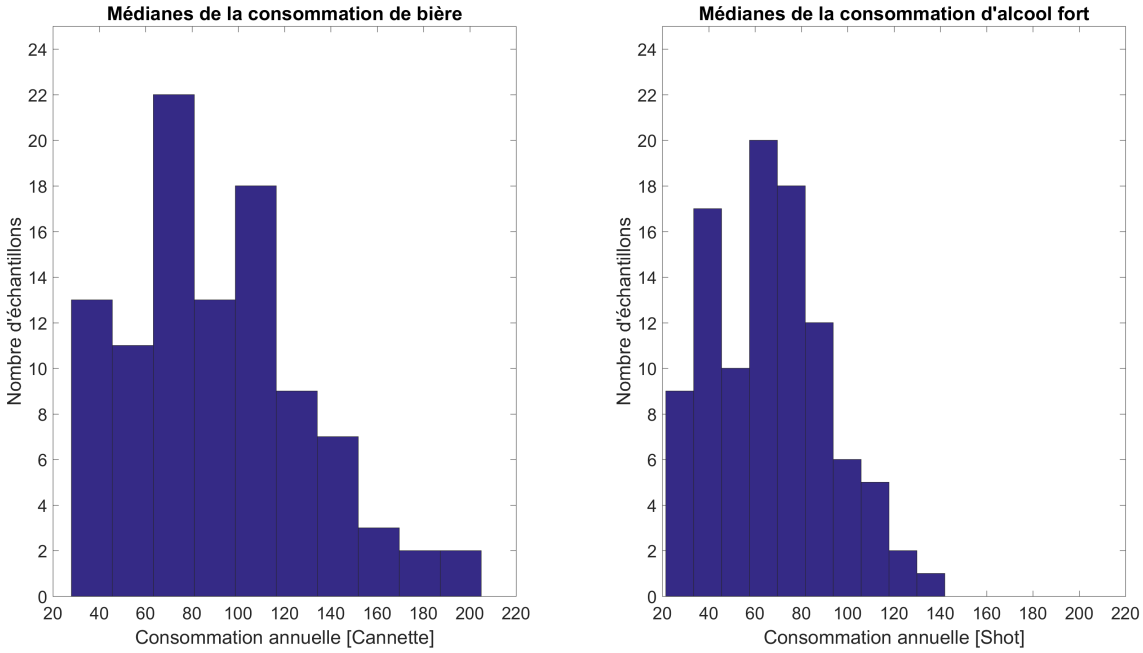

FIGURE 10 – Moyenne des 100 échantillons de 20 pays

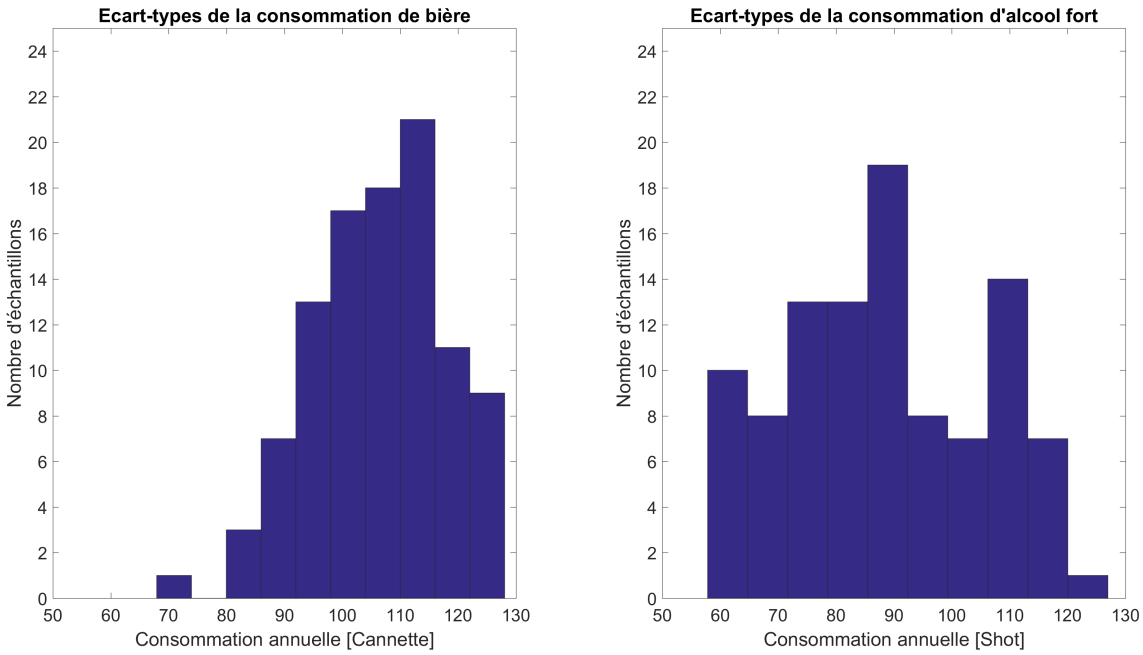

FIGURE 11 – Moyenne des 100 échantillons de 20 pays

— le biais de l'estimateur m_x de la consommation moyenne de vin de la population vaut 0.2526
— la variance de l'estimateur m_x de la consommation moyenne de vin de la population vaut 62.1311
— le biais de l'estimateur $median_x$ de la consommation moyenne de vin de la population vaut 0.63
— la variance de l'estimateur $median_x$de la consommation moyenne de vin de la population vaut 22.6445

On se rend compte que l'estimateur m_x pour un échantillon de taille 50 représente le mieux la population : en effet, son biais est le plus faible de tous. Toutefois, il arrive de temps à autres que ce soit l'estimateur $median_x$ qui soit le plus faible. Étant donné que sa variance est plus faible, les valeurs obtenues pour cet estimateur semble moins s'égarer. Cet estimateur peut donc être aussi considéré comme un estimateur correct.

En conclusion, les valeurs trouvées à l'aide de la médiane sont généralement plus élevées et traduisent une moins bonne représentation de la population étudiée, à fortiori quand un échantillon plus petit est considéré.

3.d

Les deux intervalles de confiances ont été construits par le script `Q3.m`.
Pour rappel, la loi de *student* donne l'intervalle de confiance de la façon suivante :

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

où s_{n-1} est la racine carrée de la variance corrigée, $t_{\frac{\alpha}{2}}$ est un coefficient (2.093) déterminé via la table de student, n est le nombre de pays par échantillon. En utilisant cette loi, nous trouvons que sur 100 échantillons, seulement 57 correspondent à notre critère.

La loi de Gauss, quant à elle, donne l'intervalle de confiance via :

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

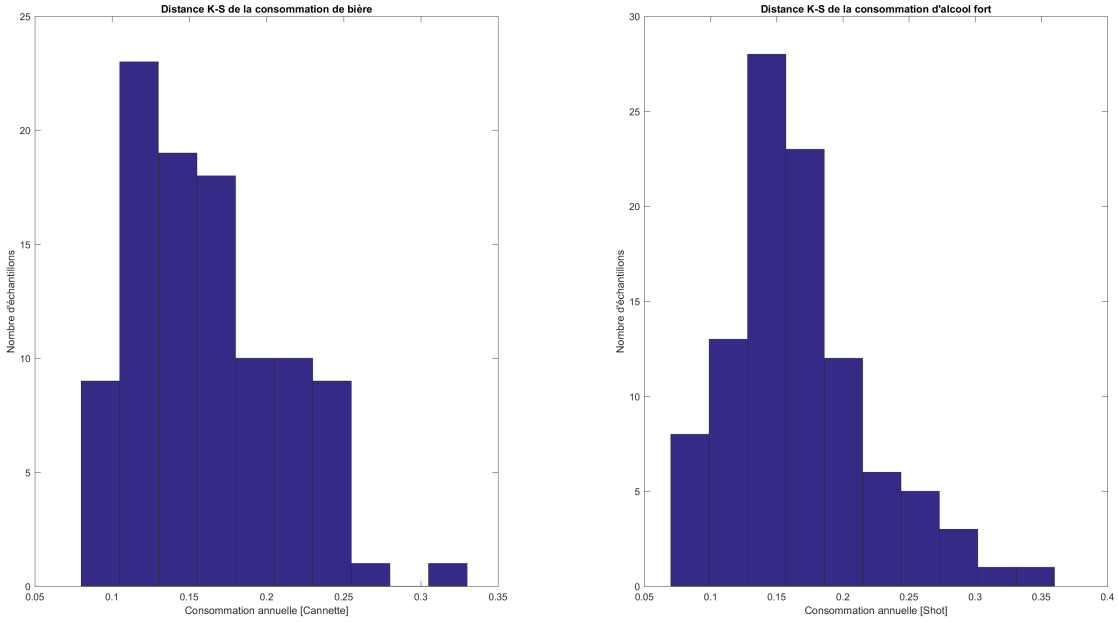

FIGURE 12 – Distances entre les polygones de fréquences cumulées de la population et des 100 échantillons de 20 pays

avec σ l'écart type, et $u_{\frac{\alpha}{2}}$ un coefficient trouvé dans la table de Gauss. Lorsque $1-\alpha = 0.95$, $u_{\frac{\alpha}{2}} = 1.96$. Ici, ce sont 54 échantillons qui ont été comptabilisés.

Il y a donc légèrement plus d'échantillons contenant la valeur de la population quand on utilise la loi de student. Cela est assez cohérent : l'intervalle de confiance de Gauss prend en compte l'écart type, ce qui a pour effet de le rendre plus fin. Mais la faible différence entre les deux nombres nous permet d'affirmer que supposer que la variable parente était gaussienne n'était pas déraisonnable.

4

4.a

Le script `Q4.` nous permet de savoir que l'état belge a rejeté l'hypothèse 4 fois, soit dans 4% des cas, ce qui est juste en dessous de α.

4.b

Il vient que dans 10 cas sur 100, l'OMS a considéré que les belges faisait partie des plus gros consommateurs de bière au monde. C'est plus élevé que le résultat du point précédent. Cela signifie que les instituts indépendants condamnent la Belgique plus vite. Cela est dû au fait que la Belgique pour rendre son jugement n'a eu accès qu'à un seul échantillon pour son test, là où les instituts sont plusieurs et donc ont généré plusieurs échantillons. Et comme il suffit d'un seul institut concluant que l'hypothèse est à rejeter, cela arrive beaucoup plus fréquemment. Si l'on avait pris un x (cfr. énoncé) plus élevé, nous aurions eu moins de rejets de la part de l'état et des instituts.

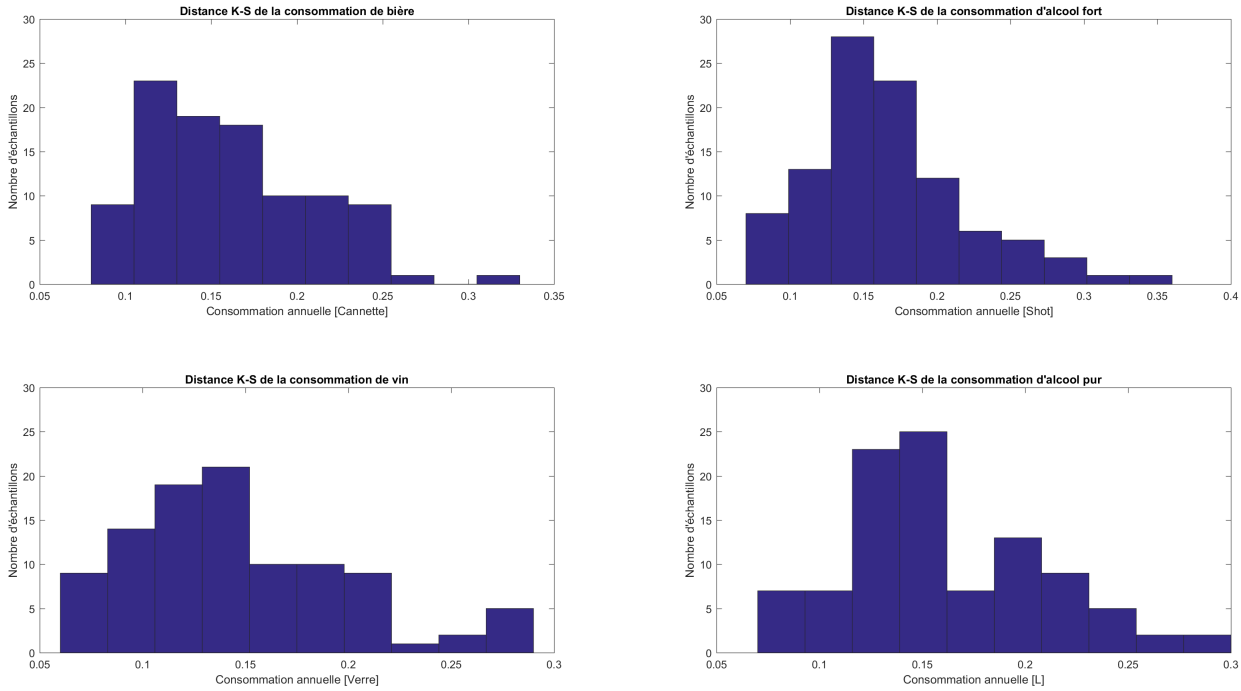

FIGURE 13 – Distances entre les polygones de fréquences cumulées de la population et des 100 échantillons de 20 pays pour chaque alcool

4.c

Comme expliqué plus haut, les instituts sont avantagés parce qu'il suffit qu'il y en ait un seul qui trouve que la consommation belge est problématique pour que ce jugement soit validé. Comme tous les instituts génèrent leur propre échantillon, plus il y a d'instituts plus les chances de trouver un échantillon rejetant l'hypothèse augmentent.

On peut donc agir à plusieurs niveaux :

- Une première solution serait de ne garder qu'un seul institut. Cela signifierait un seul échantillon en plus de celui généré par l'état Belge et donc la fin du déséquilibre.
- Une seconde solution serait d'imposer le même échantillon pour tous les participants.
- Enfin une dernière solution équitable serait de changer la règle de base et de dire que l'OMS doit faire ses recommandations à la Belgique seulement si une majorité des retours rejettent l'hypothèse.

5 Annexes

Voici les codes nécessaires à la réalisation de ce projet.

5.a Q1.m

```
%% Q1 - Analyse descriptive
% @AUTHOR Olivier MOITROUX
% @AUTHOR Pierre HOCKERS

close all;
clc;
clear all;

%% Importing DATA
filename = 'C:\Users\Philippe\Documents\MATLAB\db_stat85.csv';
[countries, beer_servings, wine_servings, spirit_servings, tot_lit_pure_alcohol] = import_c

%% a) Histogramme consommation biere et alcool fort
    % Plot data
    figure('Name','1a) Histogramme bire et alcool fort','NumberTitle', 'off');
    bar(1:100, [beer_servings, spirit_servings]);
    legend('Bire [Cannette]', 'Alcool fort [shots]');
    % title('Histogramme de la consommation de bire et d''alcool fort');
    ylabel('Consommation');
    set(gca, 'XTickLabel', countries, 'XTick', 1:numel(countries), 'fontsize', 18);
    rotateticklabel(gca); % @AUTHOR : Andy Bliss
    colormap winter
    %colormap([0 0 1; 1 0 0]);

    % plot histogramme
    figure;
    hist([beer_servings, spirit_servings]);
    legend('Bire [Cannette]', 'Alcool fort [Shot]');
    xlabel('Consommation');
    ylabel('Pays');
    set(gca, 'fontsize', 18);
    colormap winter

%% b) Moyenne - Mdiane - Mode - cart-type
    disp('b)');
    beerAvrg   = mean(beer_servings)
    spiritAvrg = mean(spirit_servings)

    beerMedian   = median(beer_servings)
    spiritMedian = median(spirit_servings)

    beerMode   = mode(beer_servings)
    spiritMode = mode(spirit_servings)

    beerStDev   = std(beer_servings) %sqrt(var())
    spiritStDev = std(spirit_servings)

%% c) Consommation normale
```

```
beerProp = 0;
lowerBound = beerAvrg - beerStDev;
upperBound = beerAvrg + beerStDev;
for i = 1:length(beer_servings)
        if beer_servings(i) > lowerBound && beer_servings(i) < upperBound
            beerProp = beerProp + 1;
        end
end

beerProp = beerProp / length(countries);

disp(['c-i) ', num2str(beerProp*100), ' % des pays ont une consommation de bire "norma
if beer_servings(20) > lowerBound && beer_servings(20) < upperBound
    disp('La belgique a une consommation de bire "normale"');
else disp('La Belgique a une consommation de bire "anormale"');
end

lowerBound = spiritAvrg - spiritStDev;
upperBound = spiritAvrg + spiritStDev;
spiritProp = 0;
for i = 1:length(spirit_servings)
        if spirit_servings(i) > lowerBound && spirit_servings(i) < upperBound
            spiritProp = spiritProp + 1;
        end
end

spiritProp = spiritProp / length(countries);

disp(['c-ii) ', num2str(spiritProp*100), ' % des pays ont une consommation de spiritueu
if beer_servings(20) > lowerBound && beer_servings(20) < upperBound
    disp('La belgique a une consommation de spiritueux "normale"');
else disp('La Belgique a une consommation de spiritueux "anormale"');
end

%% d) i)Boites a moustaches
figure('Name','1d) Bote  moustache bire et alcool fort','NumberTitle', 'off');
subplot(1,2,1); % 1x2 grid first graph
boxplot(beer_servings);
title('Bire');
ylabel('Consommation [L]');
set(gca, 'XTickLabel', '', 'YTick', 0:25:400, 'fontsize', 18);
ylim([-10 400]);

subplot(1,2,2);
boxplot(spirit_servings);
title('Alcool fort');
ylabel('Consommation [L]');
set(gca, 'XTickLabel', '', 'YTick', 0:25:400, 'fontsize', 18);
ylim([-10 400]);

% ii) quartiles
disp('d) Quartiles :');
beerQuart   = quantile(beer_servings, [.25, .50, .75])
spiritQuart = quantile(spirit_servings, [.25, .50, .75])
```

```
%% e) Polygone de frquence cumule de la consommation de bire
    figure('Name','1e) Frquence cumule consommation de bire','NumberTitle', 'off');
    p = cdfplot(beer_servings); % empirical cumulative distribution function
    % On aurait pu exploiter la structure STATS [H,STATS] renvoye par
    % cdfplot pour b)
    hold on;
    l1 = line([200 200], [0 .78], 'Color', 'g', 'LineStyle','--');
    l2 = line([0 200], [.78 .78], 'Color', 'g', 'LineStyle','--');
    l3 = line([beer_servings(20) beer_servings(20)], [0 .91], 'Color', 'r', 'LineStyle','--
    l4 = line([0 beer_servings(20)], [.91 .91], 'Color', 'r', 'LineStyle','--');
    hold off;
    legend([p, l1, l3], 'Polygone des frquences cumules', 'Consommation de 200 cannettes',
    set(gca, 'fontsize', 18);
    title(''); % turn off auto title

%% f) Scatterplot et cofficients de corrlation linaire

    % $$Cor(X,Y) = \frac{Cov{X,Y}}{\sigma_X \sigma_Y}$$
    disp('f) Cofficients de corrlation linaire :');

    figure('Name','1f) ScatterPlots','NumberTitle', 'off');
    subplot(1,3,1);
    scatter(tot_lit_pure_alcohol, beer_servings,'filled', 'b');
    title('Alcool pur et bire');
    set(gca, 'fontsize', 18);
    xlabel('Consommation annuelle [L]');
    ylabel('Consommation annuelle [Cannette]');

    disp('Alcool pur - Bire');
    r1 = corrcoef(tot_lit_pure_alcohol, beer_servings);
    r1 = r1(1,2)

    subplot(1,3, 2);
    scatter(tot_lit_pure_alcohol, wine_servings, 'filled', 'r');
    title('Alcool pur et vin');
    set(gca, 'fontsize', 18);
    xlabel('Consommation annuelle [L]');
    ylabel('Consommation annuelle [Verre]');

    disp('Alcool pur - Vin');
    r2 = corrcoef(tot_lit_pure_alcohol, wine_servings);
    r2 = r2(1,2)

    subplot(1,3, 3);
    scatter(tot_lit_pure_alcohol, spirit_servings, 'filled', 'g');
    title('Alcool pur et spiritueux');
    set(gca, 'fontsize', 18);
    xlabel('Consommation annuelle [L]');
    ylabel('Consommation annuelle [Shot]');

    disp('Alcool pur - Spiritueux');
    r3 = corrcoef(tot_lit_pure_alcohol, spirit_servings);
    r3 = r3(1,2)
```

```
    clearvars l1 l2 l3 l4 p filename i
```

5.b import_csv.m

```
function [country, beer_servings, wine_servings, spirit_servings, total_litres_of_pure_alco
% IMPORT_CSV Gnre les vecteurs colonnes prsents dans un fichier csv

% Auto-generated by MATLAB on 2017/11/15 21:11:44 and modified by Olivier
% Moitroux and Pierre Hockers

% Initialize variables.
if nargin == 1
    delimiter = ',';
    startRow = 2;
end


% Format for each line of text:
%   column1: text (%s)
%       column2: double (%f)
%   column3: double (%f)
%       column4: double (%f)
%   column5: double (%f)
formatSpec = '%s%f%f%f%f%[^\n\r]';

% Open the text file.
fileID = fopen(filename,'r');

% Read columns of data according to the format.
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter, 'HeaderLines' ,startRow-1

% Close the text file.
fclose(fileID);

% Post processing for unimportable data.
% No unimportable data rules were applied during the import, so no post
% processing code is included.

% Allocate imported array to column variable names
country = dataArray{:, 1};
beer_servings = dataArray{:, 2};
spirit_servings = dataArray{:, 3};
wine_servings = dataArray{:, 4};
total_litres_of_pure_alcohol = dataArray{:, 5};

% Clear temporary variables
% clearvars filename delimiter startRow formatSpec fileID dataArray ans;

end
```

5.c Q2.m

```
%% Q2 - Gnration d'chantillons indpendants et identiquement distribus
% @AUTHOR Olivier MOITROUX
```

```
% @AUTHOR Pierre HOCKERS
clear all;
close all;
clc;

%% Importing DATA
filename = 'C:\Users\Philippe\Documents\MATLAB\db_stat85.csv';
[~, beer_servings, wine_servings, spirit_servings, tot_lit_pure_alcohol] = import_csv(filena
clear filename;

%% a) Tirage d'chantillon i.i.d. de 20 pays
    rng(1);
    % Mme seed pour garder la mme gnration alatoire (
    % reproductibilit)
    sample = randsample(1:length(beer_servings), 20, true);

%% i)
    disp('a-i)');
    beerAvrgSamp    = mean(beer_servings(sample))
    spiritAvrgSamp = mean(spirit_servings(sample))

    beerMedianSamp    = median(beer_servings(sample))
    spiritMedianSamp = median(spirit_servings(sample))

    beerModeSamp      = mode(beer_servings(sample))
    spiritModeSamp    = mode(spirit_servings(sample))

    beerStDevSamp     = std(beer_servings(sample)) %sqrt(var())
    spiritStDevSamp   = std(spirit_servings(sample))

 %% ii)
    disp('a-ii)');
    figure('Name','2a-ii) Bote  moustache bire et alcool fort','NumberTitle', 'off');
    subplot(1,2,1); % 1x2 grid first graph
    boxplot(beer_servings(sample));
    title('Bire');
    ylabel('Consommation [L]');
    set(gca, 'XTickLabel', '', 'YTick', 0:25:400, 'fontsize', 18);
    ylim([-10 400]);

    subplot(1,2,2);
    boxplot(spirit_servings(sample));
    title('Alcool fort');
    ylabel('Consommation [L]');
    set(gca, 'XTickLabel', '', 'YTick', 0:25:400, 'fontsize', 18);
    ylim([-10 400]);

%% iii)
    figure('Name','2a-iii) Frquence cumule consommation de bire et d''alcool fort','Number
    % Fonction de distribution empirique cumulative
    grid;
    subplot(1,2,1);
    hold on;
    cdfplot(beer_servings(sample));
```

```
    cdf2 = cdfplot(beer_servings);
    hold off;
    set(cdf2, 'Color', 'r');
    set(gca,'XTick', 0:50:400, 'fontsize', 18);
    xlim([0, 400]);
    xlabel('Consommation annuelle [Canette]');
    title('Polygone des frquences cumules - Bire');
    legend('Echantillon (20)', 'Population');

    subplot(1,2,2);
    hold on;
    cdfplot(spirit_servings(sample));
    cdf4 = cdfplot(spirit_servings);
    hold off;
    xlim([0, 400]);% max(spirit_servings)]
    xlabel('Consommation annuelle [Shot]');
    set(cdf4, 'Color', 'r');
    set(gca,'XTick', 0:50:400, 'fontsize', 18);
    title('Polygone des frquences cumules - Alcool fort');
    legend('Echantillon (20)', 'Population');

    clearvars cdf2 cdf4;

    % distance de Kolmogorov - Smirnov
    disp('a-iii) Distance de Kolmogorov-Smirnov :')

    [~,~,beerKSD]    = kstest2(beer_servings, beer_servings(sample))
    [~,~,spiritKSD] = kstest2(spirit_servings, spirit_servings(sample))

%% b) Tirage d'chantillon 100 i.i.d. de 20 pays

    rng(2);
    % Mme seed pour garder la mme gnration alatoire (
    % reproductibilit)
    size = length(beer_servings);

%%   i)
    beerAvrg100   = zeros(1, size);
    spiritAvrg100 = zeros(1, size);
    for i = 1:size
        rand_countries   = randsample(1:size, 20);
        beerAvrg100(i)   = mean(beer_servings(rand_countries));
        spiritAvrg100(i) = mean(spirit_servings(rand_countries));
    end

    figure('Name','2b-i) Histogramme moyenne consommation de bire et d''alcool fort d''cha
    subplot(1,2,1);
    hist(beerAvrg100);
    title('Moyennes de la consommation de bire');
    xlabel('Consommation annuelle [Cannette]');
    ylabel('Nombre d''chantillons');
    xlim([40, 180]);
    ylim([0, 25]);
    set(gca, 'fontsize', 18);
```

```
set(gca, 'XTick', 40:10:180,'YTick', 0:2:25, 'fontsize', 18);
% bar(beerAvrg100)

subplot(1,2,2);
hist(spiritAvrg100);
title('Moyennes de la consommation d''alcool fort');
xlabel('Consommation annuelle [Shot]');
ylabel('Nombre d''chantillons');
xlim([40, 180]);
ylim([0, 25]);
set(gca, 'fontsize', 18);
set(gca, 'XTick', 40:10:180, 'YTick', 0:2:25, 'fontsize', 18);

disp('b-i) Comparaisons des moyennes avec la population');
%stdOfBeerAvrg100  = std(beerAvrg100)
meanOfBeerAvrg100 = mean(beerAvrg100)

%stdOfSpiritAvrg100  = std(spiritAvrg100)
meanOfSpiritAvrg100 = mean(spiritAvrg100)

%%  ii)
rng(2);
size          = length(beer_servings);
beerMed100    = zeros(1, size);
spiritMed100 = zeros(1, size);
for i = 1:size
    rand_countries  = randsample(1:size, 20);
    beerMed100(i)   = median(beer_servings(rand_countries));
    spiritMed100(i) = median(spirit_servings(rand_countries));
end

figure('Name','2b-ii) Histogramme mdianes consommation de bire et d''alcool fort d''cha
subplot(1,2,1);
hist(beerMed100);
title('Mdianes de la consommation de bire');
xlabel('Consommation annuelle [Cannette]');
ylabel('Nombre d''chantillons');
xlim([20, 220]);
ylim([0, 25]);
set(gca, 'fontsize', 18);
set(gca, 'XTick', 20:20:220,'YTick', 0:2:25, 'fontsize', 18);
% bar(beerAvrg100)

subplot(1,2,2);
hist(spiritMed100);
title('Mdianes de la consommation d''alcool fort');
xlabel('Consommation annuelle [Shot]');
ylabel('Nombre d''chantillons');
xlim([20, 220]);
ylim([0, 25]);
set(gca, 'fontsize', 18);
set(gca, 'XTick', 20:20:220, 'YTick', 0:2:25, 'fontsize', 18);
% bar(beerAvrg100)
```

```
disp('b-ii) Comparaisons des mdianes avec la population et avec b-i)');
%stdOfBeerMed100  = std(beerMed100)
meanOfBeerMed100 = mean(beerMed100)

%stdOfSpiritMed100  = std(spiritMed100)
meanOfSpiritMed100 = mean(spiritMed100)

%%  iii)
rng(2);
size          = length(beer_servings);
beerStd100    = zeros(1, size);
spiritStd100 = zeros(1, size);
for i = 1:size
    rand_countries  = randsample(1:size, 20);
    beerStd100(i)   = std(beer_servings(rand_countries));
    spiritStd100(i) = std(spirit_servings(rand_countries));
end

figure('Name','2b-iii) Histogramme cart-types consommation de bire et d''alcool fort d'
subplot(1,2,1);
hist(beerStd100);
title('Ecart-types de la consommation de bire');
xlabel('Consommation annuelle [Cannette]');
ylabel('Nombre d''chantillons');
xlim([50, 130]);
ylim([0, 25]);
set(gca, 'fontsize', 18);
set(gca, 'XTick', 50:10:130,'YTick', 0:2:25, 'fontsize', 18);

subplot(1,2,2);
hist(spiritStd100);
title('Ecart-types de la consommation d''alcool fort');
xlabel('Consommation annuelle [Shot]');
ylabel('Nombre d''chantillons');
xlim([50, 130]);
ylim([0, 25]);
set(gca, 'fontsize', 18);
set(gca, 'XTick', 50:10:130,'YTick', 0:2:25, 'fontsize', 18);

disp('b-iii) Comparaisons des cart-types avec la population');
meanOfBeerStd100   = mean(beerStd100)
meanOfSpiritStd100 = mean(spiritStd100)

%%  iv)

rng(2);
size          = length(beer_servings);
beerKSD100    = zeros(1, size);
spiritKSD100 = zeros(1, size);
for i = 1:size
    rand_countries        = randsample(1:size, 20);
    [~,~,beerKSD100(i)]   = kstest2(beer_servings, beer_servings(rand_countries));
    [~,~,spiritKSD100(i)] = kstest2(spirit_servings, spirit_servings(rand_countries));
end
```

```
figure('Name','2b-iv) Histogramme distance K-S consommation de bire et d''alcool fort d
subplot(1,2,1);
hist(beerKSD100);
title('Distance K-S de la consommation de bire');
xlabel('Consommation annuelle [Cannette]');
ylabel('Nombre d''chantillons');

subplot(1,2,2);
hist(spiritKSD100);
title('Distance K-S de la consommation d''alcool fort');
xlabel('Consommation annuelle [Shot]');
ylabel('Nombre d''chantillons');

%%  v)
figure('Name','2b-v) Histogramme distance K-S consommation des diffrents boissons','Nu
subplot(2,2,1);
hist(beerKSD100);
title('Distance K-S de la consommation de bire');
xlabel('Consommation annuelle [Cannette]');
ylabel('Nombre d''chantillons');
ylim([0, 30]);

subplot(2,2,2);
hist(spiritKSD100);
title('Distance K-S de la consommation d''alcool fort');
xlabel('Consommation annuelle [Shot]');
ylabel('Nombre d''chantillons');
ylim([0, 30]);

rng(2);
size               = length(beer_servings);
wineKSD100         = zeros(1, size);
pureAlcoholKSD100 = zeros(1, size);
for i = 1:size
    rand_countries                = randsample(1:size, 20);
    [~,~,wineKSD100(i)]          = kstest2(wine_servings, wine_servings(rand_countries));
    [~,~,pureAlcoholKSD100(i)] = kstest2(tot_lit_pure_alcohol, tot_lit_pure_alcohol(ran
end

subplot(2,2,3);
hist(wineKSD100);
title('Distance K-S de la consommation de vin');
xlabel('Consommation annuelle [Verre]');
ylabel('Nombre d''chantillons');
ylim([0, 30]);

subplot(2,2,4);
hist(pureAlcoholKSD100);
title('Distance K-S de la consommation d''alcool pur');
xlabel('Consommation annuelle [L]');
ylabel('Nombre d''chantillons');
ylim([0, 30]);
```

```
    clear i;
```

5.d Q3.m

```
%% Q3 - Estimation
% @AUTHOR Olivier MOITROUX
% @AUTHOR Pierre HOCKERS

close all;
clc;
clear all;

%% Importing DATA
filename = 'C:\Users\Philippe\Documents\MATLAB\db_stat85.csv';
[~,~, wine_servings, ~,~] = import_csv(filename);

%% a) moyenne, biais et variance
    disp('a)');
    [biaisMx20, varMx20, mx20] = mean_estim(wine_servings, 20);
    biaisMx20, varMx20
%% b) mdiane
    disp('b)');
    [biaisMedx20, varMedx20] = med_estim(wine_servings, 20)

 %% c) Idem avec 50

    disp('c-i)');
    [biaisMx50, varMx50] = mean_estim(wine_servings, 50)

    disp('c-ii)');
    [biaisMedx50, varMedx50] = med_estim(wine_servings, 50)

 %% d) Intervalle de confiance  95 %

    size = length(wine_servings);
    cntStudent = 0;
    cntGauss = 0;
    studLowerBound  = zeros(1,100); studUpperBound  = zeros(1,100);
    gaussLowerBound = zeros(1,100); gaussUpperBound = zeros(1,100);

    for i=1:size
     %% i) Student
        stdStud = std(randsample(1:size, 20))/sqrt(19);
        studLowerBound(i) = mx20(i) - 2.093*stdStud;
        studUpperBound(i) = mx20(i) + 2.093*stdStud;

        if mean(wine_servings) >= studLowerBound(i) && mean(wine_servings) <= studUpperBoun
            cntStudent = cntStudent+1;
        end
     %% ii) Gauss
        stdGauss = std(randsample(1:size, 20))/sqrt(20);
        gaussLowerBound(i) = mx20(i) - 1.960*stdGauss;
        gaussUpperBound(i) = mx20(i) + 1.960*stdGauss;
        if mean(wine_servings) >= gaussLowerBound(i) && mean(wine_servings) <= gaussUpperBo
            cntGauss = cntGauss+1;
```

```
        end
    end

    disp('d)');
    cntStudent
    cntGauss

    clearvars i  filename;
```

5.e mean_estim.m

```
function [estBiaisMx, varMx, mx] = mean_estim( servings, nbCountry)
%Mean_estim estime le biais et la variance de l'estimateur m_x de la
%consommation moyenne de servings.

size    = length(servings);
mx = zeros(1, size);
%rng(3);
for i = 1:size
    rand_country = randsample(1:size, nbCountry, true);
    mx(i)        = mean(servings(rand_country));
end

% Biais
estBiaisMx = mean(mx) - mean(servings);
% Variance
varMx      = var(mx);

end
```

5.f med_estim.m

```
function [ estBiaisMedx, varMedx ] = med_estim( servings, nbCountry)
%Med_estim estime le biais et la variance de l'estimateur median_x de la
%consommation moyenne de servings.

size    = length(servings);
medx = zeros(1, size);
%rng(3);
for i = 1:size
    rand_country = randsample(1:size, nbCountry);
    medx(i)      = median(servings(rand_country));
end

% Biais
estBiaisMedx = mean(medx) - median(servings);
% Variance
varMedx      = var(medx);
end
```

5.g Q4.m

```
%% Q4 - Test d'hypothses
% @AUTHOR Olivier MOITROUX
% @AUTHOR Pierre HOCKERS
```

```
close all;
clc;
clear all;

%% Importing DATA
filename = 'C:\Users\Philippe\Documents\MATLAB\db_stat85.csv';
[countries ,beer_servings,~, ~,~] = import_csv(filename);
clear filename;

%% Tirage de 100 fois 6 chantillons i.i.d. de 50 pays
    % Initialisation1
    size = length(beer_servings); % 100
    belgiumIndex = strmatch('Belgium', countries);

    disp('Pourcentage des pays ayant une plus grande cons. de bire que la Belgique:');
    x = compute_x(beer_servings, belgiumIndex)
    % x = 0.10, ...
    % Initialisation2
    u_alpha = 1.645; % cfr. table de Gauss, alpha = 0.05;
    var = sqrt((1-x)*x/size);
    rejectedState = 0; rejectedOMS = 0;
    for i = 1:100
        %% a) L'tat belge
        randCountries = randsample([1:belgiumIndex-1, belgiumIndex+1:size], 49);
        randCountries(50) = belgiumIndex;
        boolRejState    = test_hyp0(beer_servings, randCountries,x, u_alpha, var);
        if(boolRejState)
                rejectedState = rejectedState + 1;
        end
        %% b) 5 instituts de statistique indpendants
        for j = 1:5
            randCountries = randsample([1:belgiumIndex-1, belgiumIndex+1:size], 49);
            randCountries(50) = belgiumIndex;
            boolOMS     = test_hyp0(beer_servings, randCountries,x, u_alpha, var);
            if(boolOMS)
                rejectedOMS = rejectedOMS + 1;
                break;
            end
        end
    end
    disp('a)');
    rejectedState
    disp('b)');
    rejectedOMS
```

5.h compute_x.m

```
function [x] = compute_x(beer_servings, countryIndex)
%Calcule le pourcentage 'x' de pays ayant une consommation plus grande que
% l'index du pays donn en argument.

n = length(beer_servings); tmp = 0;

for i = [1:countryIndex-1, countryIndex+1:n]
```

```
    if(beer_servings(i) > beer_servings(countryIndex))
        tmp = tmp + 1;
    end
end

x = tmp/n;

end
```

5.i test_hyp0.m

```
function [ bool ] = test_hyp0(beer_servings, randCountries,x, u_alpha, var)
%test_hyp0 value si l'hypothse H0 est rejete.

n = length(randCountries); tmp = 0;

for i = 1:n-1
    if beer_servings(randCountries(i)) > beer_servings(randCountries(n))
        tmp = tmp + 1;
    end
end

sampleFreq = tmp/n;
if(sampleFreq >= x+u_alpha*var)
    bool = 1; % Rejet de l'hypothse H0
else bool = 0;
end
bool = logical(bool);
end
```