MATH0487-2 - Éléments de statistique

Partie 1 du projet personnel

Mahaux Jonathan

3 ème Bachelier Ingénieur Civil

Année académique 2014-2015

S122157

Question 1 : Analyse descriptive

a. Les trois histogrammes des résultats de questions de théorie sont représentés à la Figure 1 ci-dessous. Les données ont été représentées en 20 bâtonnets différents.

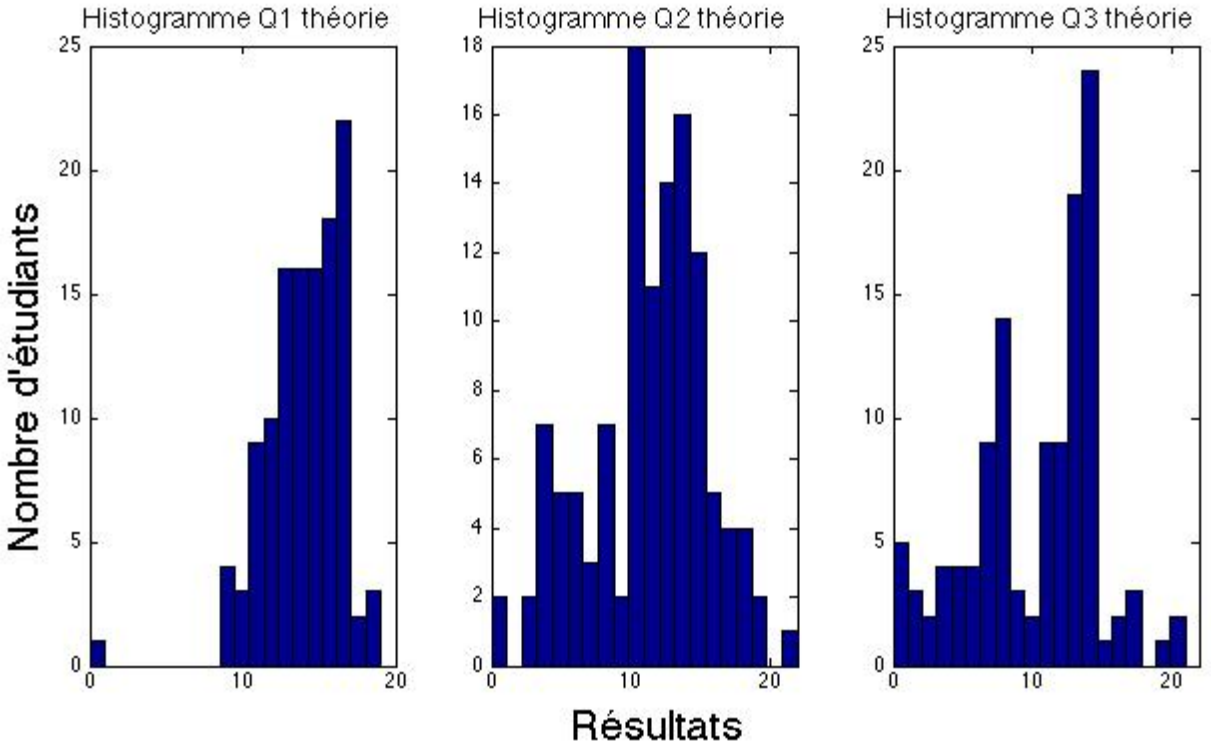


Figure 1. Histogrammes des résultats de théorie

On peut observer logiquement que les trois histogrammes sont différents. Il en ressort que le niveau de difficulté des questions est de plus en plus grand, ou que le temps consacré par les élèves à chaque question est différent : la majorité des élèves ont obtenu une noté supérieure à la moyenne à la première question ; les résultats sont principalement entre 8 et 18, concentrés autour de la moyenne. Tandis que pour la question 2 et 3, on observe une plus grande disparité des résultats ainsi qu'une plus grande proportion d'élèves en échec. La proportion des résultats suivant une loi normale sera donc plus importante à la question 1 et la note de 0 sera une donnée aberrante pour cette question. On constate aussi qu'il y a un nombre croissant d'élèves ayant obtenu une note de zéro, qui peut éventuellement s'expliquer par une mauvaise gestion du temps de la part de ces étudiants.

b. Les moyennes, médianes, modes et écart-types des trois exercices sont présentés au Tableau 1 ci-dessous :

	Exercice 1	Exercice 2	Exercice 3
Moyennes	10,8167	16,8083	7,7333
Médianes	11	18	7,5
Modes	12	20	0
Ecart-types	5,648	3,7866	5,2816

Tableau 1. Statistiques descriptives uni-variées.

Les moyennes attestent du niveau de réussite des étudiants sur les exercices. L'exercice 2 a donc été très bien réussi vu sa moyenne assez élevée. L'exercice 1 a été lui

réussi dans une moindre mesure vu sa moyenne de 10,8 ; certains étudiants étant en échec et d'autres non. L'exercice 3 a par contre été plus laborieux avec une moyenne d'échec. Les écart-types ne sont pas très élevés et la répartition des résultats des étudiants est donc concentrée autour de leur moyenne. On devrait donc observer plus de notes élevées dans le cas de l'exercice 2, satisfaisantes pour l'exercice 1 et faibles pour l'exercice3. Les modes le confirment . Beaucoup d'étudiants ont eu la note maximale de 20 à l'exercice 2 alors que beaucoup ont eu une note nulle à l'exercice 3 et la note de satisfaction 12 à l'exercice 1. Les valeurs des médianes sont assez proches des moyennes vu la faible valeur des écart-types.

Un résultat sera normal s'il appartient à l'intervalle $[\bar{X} - s, \bar{X} + s]$ où $\bar{X}$ est la moyenne et s est l'écart-type. Les intervalles de résultats normaux des trois exercices ainsi que la proportion d'étudiants ayant obtenu un tel résultat sont présentés au Tableau 2.

	Exercice 1	Exercice 2	Exercice 3
intervalle	[5.1687 ; 16.4647]	[13.0217 ; 20.5949]	[2.4517 ; 13.015]
Proportion (en %)	65	83,33	65,83

Tableau 2. Résultats normaux

L'intervalle de l'exercice 2 s'étend au-delà de 20, or aucun bonus n'était alloué pour cette question et la note maximale est donc de 20. L'intervalle pourrait être donc limité à 20 dans ce cas de figure.

c. Les trois boîtes à moustaches sont représentées à la Figure 2.

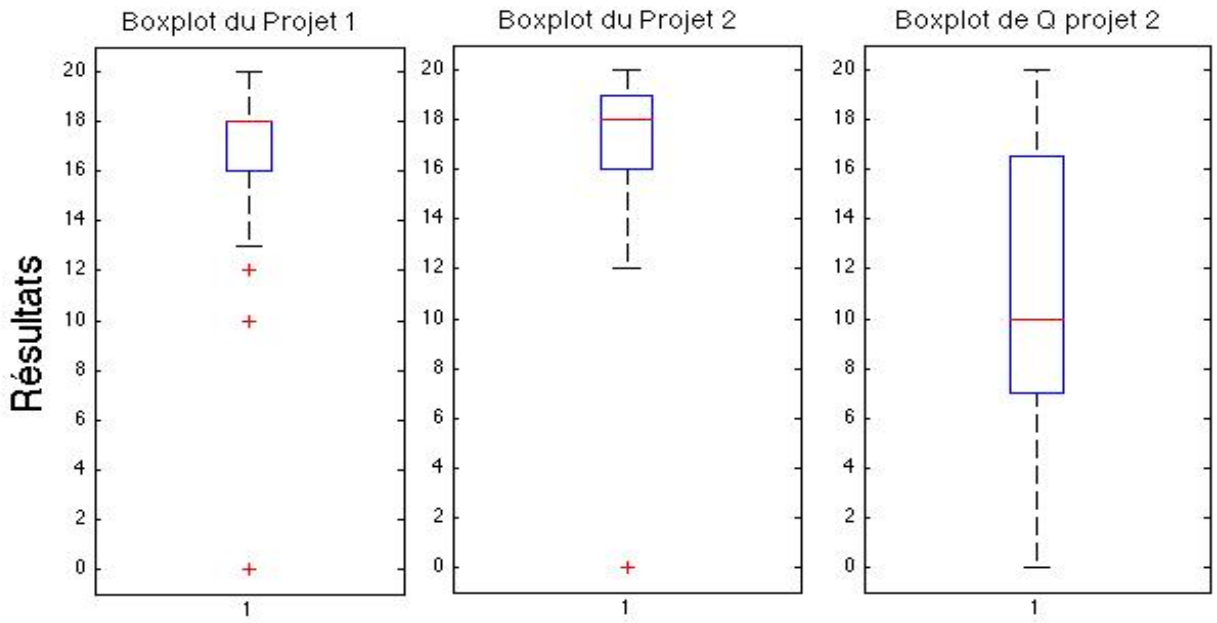


Figure 2. Boxplots relatifs aux projets

Les quartiles sont identifiables par les lignes bleues horizontales pour $Q^3 et\ Q^1$ et par une ligne rouge pour la médiane. Leurs valeurs sont reprises au Tableau 3. Elles permettent de qualifier si des données sont aberrantes ou non, c'est à dire les valeurs inférieures à $Q^1 - 1.5 * (Q^3 - Q^1)$ et supérieures à $Q^3 + 1.5 * (Q^3 - Q^1)$ (absentes ici). Il y en a 3 pour le Projet 1, 1 pour le projet 2 et aucune pour la question sur le projet 2 étant donné que les valeurs limitent sont de 0 et 20. Elles sont repérées par une croix rouge « + ». Deux de ces données aberrantes sont des notes de 0, correspondant sans doute a des

étudiants n'ayant pas réalisé leur projet. Si ces données étaient comptabilisées, elles tireraient clairement la moyenne vers le bas.
A noter que la valeur de la médiane et du percentile 75 sont égales pour le Projet 1, indiquant qu'une grande proportion des étudiants a obtenu la note de 18 à ce projet.

	Projet1	Projet 2	Q projet 2
Percentile 25(Q^1)	16	16	7
médiane	18	18	10
Percentile 75(Q^3)	18	19	16.5

Tableau 3. Quartiles

d. Les polygones des fréquences cumulées de la moyenne de chaque étudiant pour la théorie et les exercices sont illustrés à la Figure 3 ci-dessous.

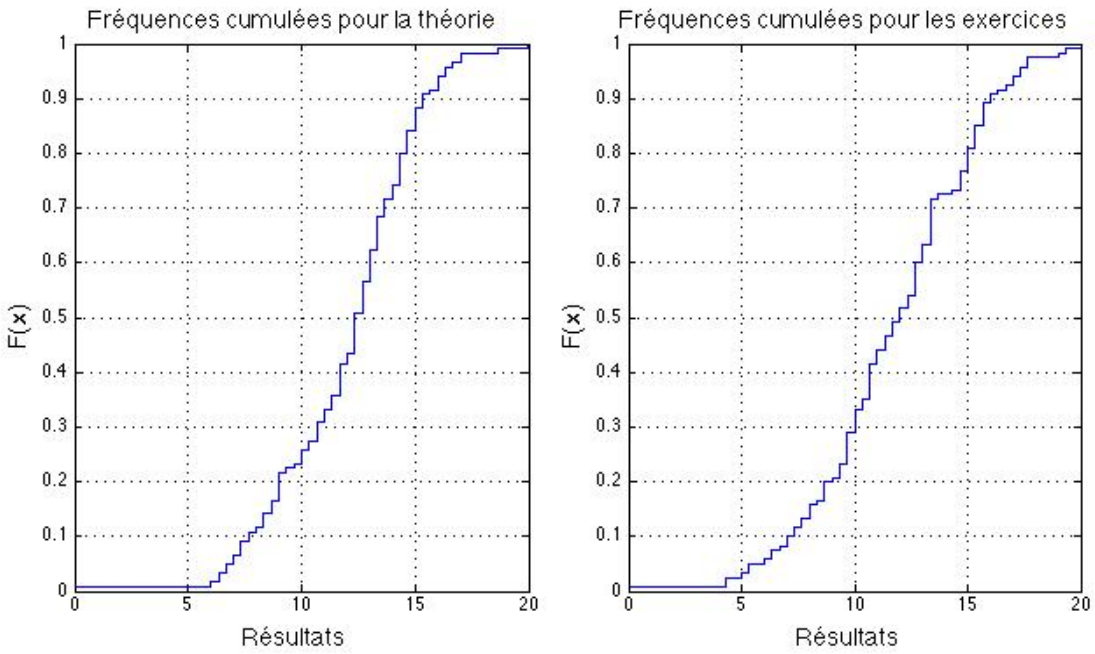


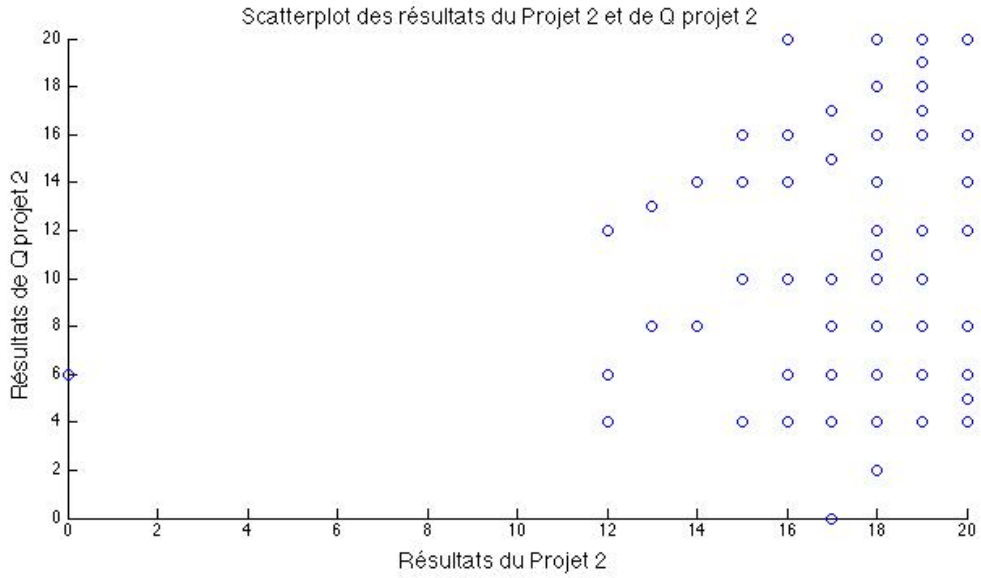
Figure 3. Polygone des fréquences cumulées

La proportion d'étudiants ayant obtenu une note entre 12 et 15 peut se faire de 2 manières. La première est de calculer précisément la fonction de répartition empirique pour des valeurs de y de 12 et 15 : $\widehat{F_x}(y) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i < y)$. La deuxième est une estimation par méthode graphique qui consiste à repérer la valeur de F(x) pour 12 et 15 et à soustraire celles-ci : F(15)-F(12). Les valeurs obtenues par les 2 méthodes sont très proches et donc l'estimation obtenue par le polygone des fréquences cumulées est valable. Les proportions obtenues sont présentées au Tableau 4 ci dessous :

		Théorie	Exercices
Proportion d'étudiants entre 12 et 15 (en %)	Méthode 1	46,67	31,67
	Méthode 2 (graphique)	45	29,17

Tableau 4. Proportion d'étudiants entre 12 et 15.

e. Le scatterplot demandé est illustré à la Figure 4 ci-dessous.



Leur coefficient de corrélation est $\begin{pmatrix} r_{x,x} & r_{x,y} \\ r_{y,x} & r_{y,y} \end{pmatrix} = \begin{pmatrix} 1 & 0{,}1407 \\ 0{,}1407 & 1 \end{pmatrix}$. Le coefficient de corrélation exprime l'intensité du lien, de la dépendance entre les deux variables. On constate donc que la linéarité entre les résultats du Projet 2 et de la question sur ce projet est faible, ils sont peu liés. La question sur le projet était une vérification de la compréhension et de la maîtrise du projet par les étudiants, on peut donc supposer qu'une partie d'entre eux a juste effectué le projet sans pour autant chercher sa valeur pédagogique.

Question 2 : Génération d'échantillons i.i.d

(a). i. Les statistiques uni-variées demandées de l'échantillon sont reprises à la Tableau 45ci-dessous :

	Exercice 1	Exercice 2	Exercice 3
Moyennes échantillon	11	17.1	8.25
Médianes échantillon	12	18.5	8.5
Ecart-types échantillon	6.3578	3.4473	4. 0115

Tableau 5. Statistiques uni-variées de l'échantillon.

Tous les résultats obtenus sont proches de ceux de la population, l'échantillon i.i.d. de 20 étudiants est donc un bon estimateur et peut représenter la population d'un point de vue statistique. Sa précision sera d'autant meilleure qu'il y aura d'étudiants dans l'échantillon.

ii. Les boîtes à moustache sont représentées à la Figure 5 ci-dessous. Ils ont une allure similaire à ceux de la population et l'échantillon représente celle-ci correctement. Si plusieurs échantillons sont tirés, on remarque qu'il y a moins de données aberrantes dû au plus petit nombre d'individus dans l'échantillon. Cependant, vu le nombre réduit d'étudiants, la médiane peut avoir la même valeur que le percentile 75 et même du maximum.

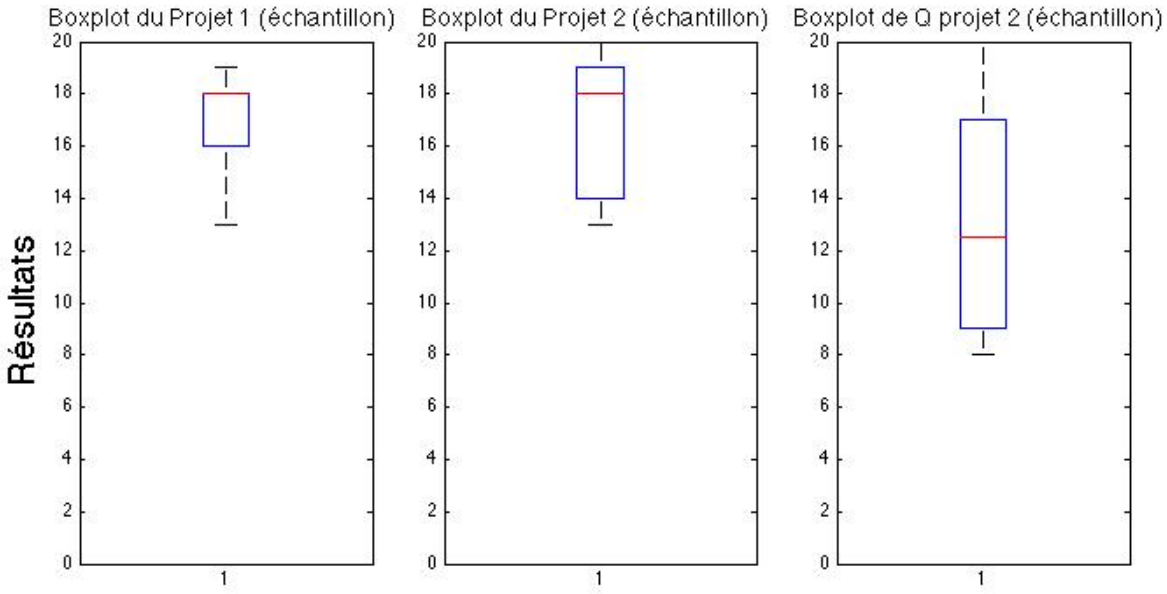


Figure 4. Boxplots de l'échantillon.

iii. Le polygone des fréquences cumulées pour la théorie de l'échantillon est représenté à la Figure 5 ci-dessous.

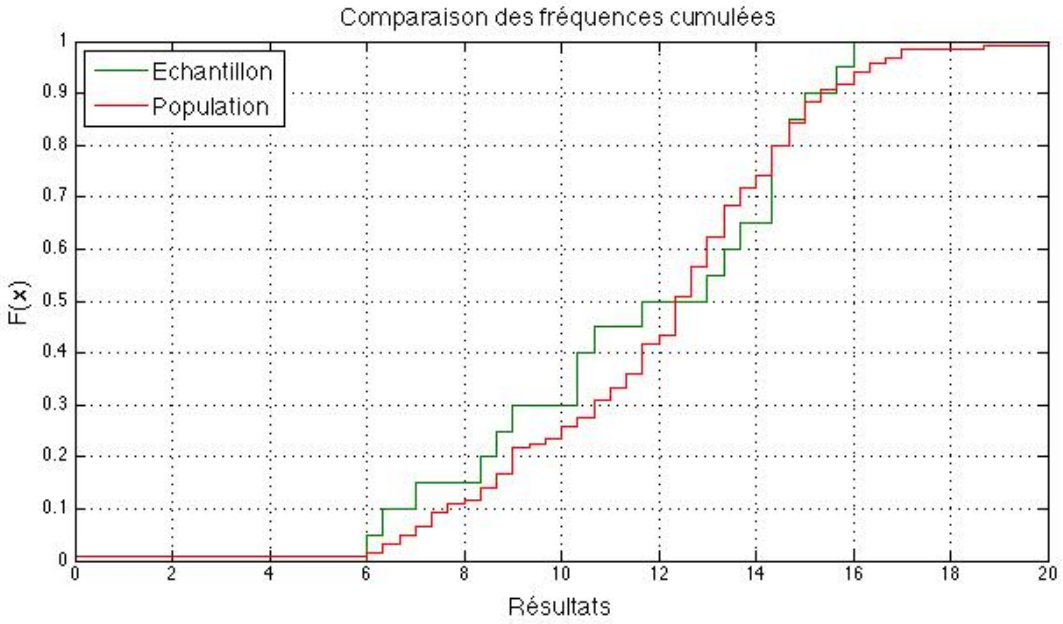


Figure 5. Comparaison des fréquences cumulées

L'allure de ces deux diagrammes est similaire, celui de l'échantillon ne comportant qu'un nombre de marches moins important. La distance de Kolgomorov-Smirnov calculée est 0.1111 qui est bien ≤ 1.

(b). i. L'histogramme des 100 moyennes de l'exercice 1 des échantillons est présenté à la Figure 6 ci-dessous. Son allure laisse supposer que la variable suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. Sa moyenne $E(m_x)$est de 10,7825, ce qui est très proche de la moyenne de la population. Ceci est confirmé par la théorie, qui affirme que $E(m_x) = \mu_x$.

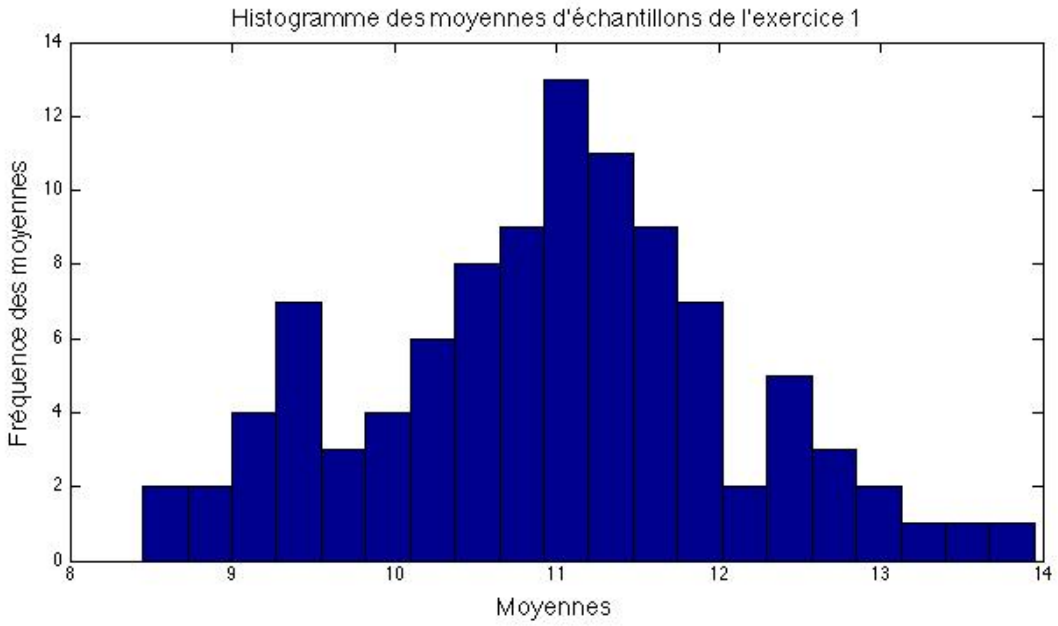


Figure 6. Histogramme des moyennes d'échantillons de l'exercice 1.

Au point 1.b. , nous avions calculé que 65% des résultats des étudiants suivent une loi gaussienne. Lorsque qu'une variable χ suit une loi gaussienne, on peut dire que m_χ suit une loi gaussienne $\mathcal{N}(\mu_\chi; \frac{\sigma_\chi}{\sqrt{n}})$, quel que soit n. Vu qu'une grande proportion des résultats suivait une loi normale, il est normal d'observer une allure de loi gaussienne pour l'histogramme.

ii. L'histogramme des 100 médianes d'échantillons de l'exercice 1 est présenté à la Figure 7 ci-dessous.

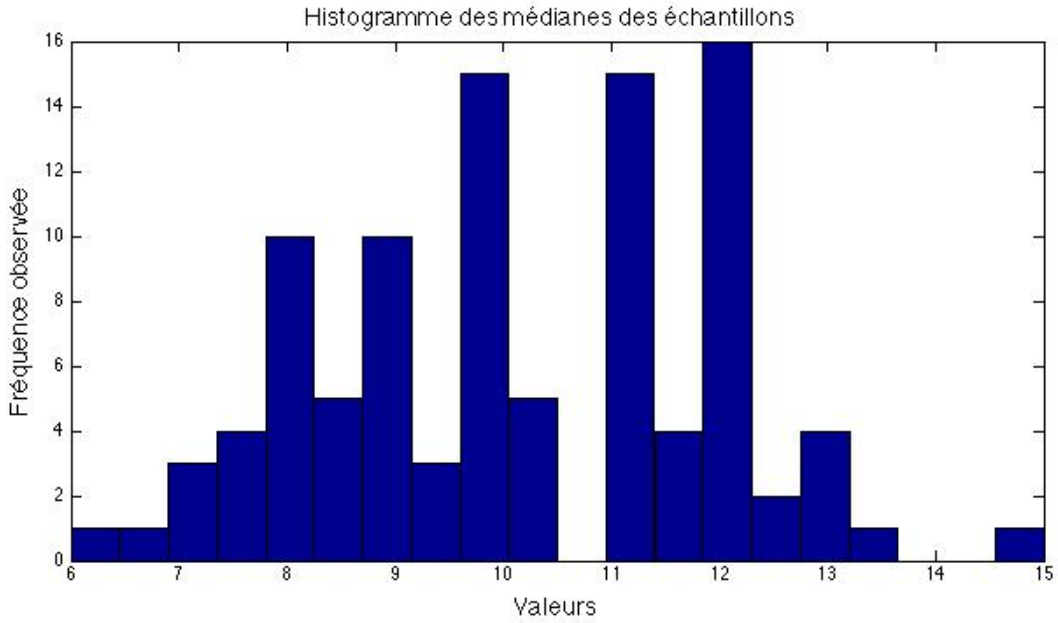


Figure 7. Histogramme des médianes

Bien qu'il existe une lacune de valeurs des médianes entre 10 et 11, l'allure de cet histogramme fait penser à celui d'une loi normale comme au point précédent. La moyenne des médianes est de 10.515 et est donc plus éloignée de la moyenne de la population que celle calculée au point précédent. La moyenne des moyennes sera donc dans ce cas un meilleur estimateur de la moyenne de la population. Si un grand nombre de données aberrantes était présent, la moyenne des médianes aurait certainement été un meilleur estimateur que la moyenne des moyennes.

iii. L'histogramme des 100 écart-types des échantillons est illustré à la Figure 8 ci-dessous.

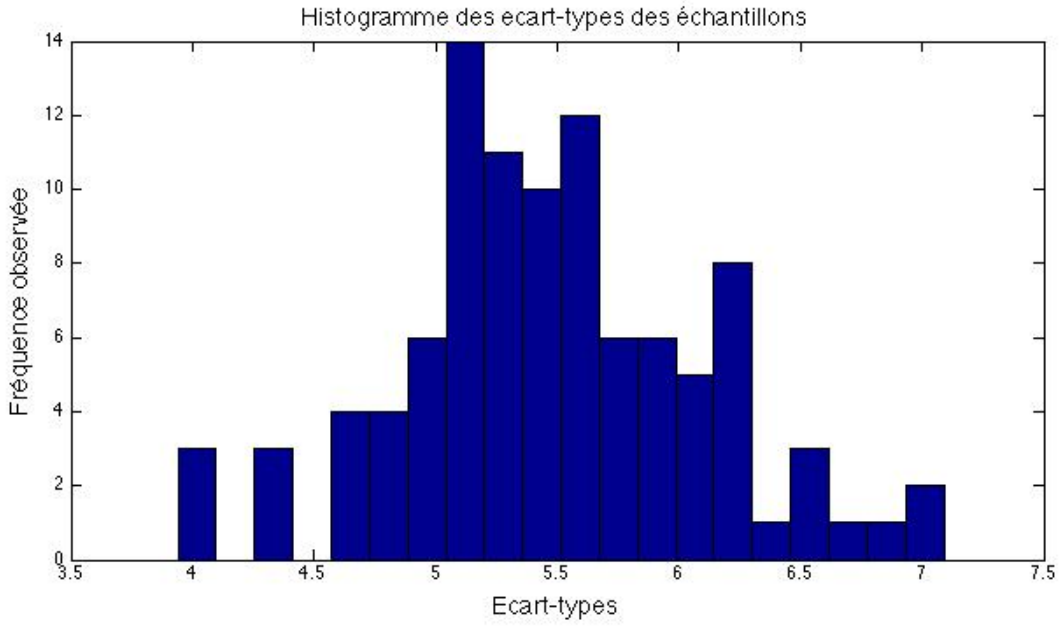


Figure 8.Histogramme des écart-types.

A nouveau, l'allure de l'histogramme fait penser que la variable contenant les écart-types suit une loi normale. La moyenne de ces 100 écart-types est de 5.477. cette valeur est proche mais inférieure à la valeur de l'écart-type calculé pour la population.
La moyenne étant faite sur un échantillon de 20 étudiants, l'écart-type sera logiquement plus faible puisque le nombre de données est moindre que celui de la population. La moyenne des écarts-types sera donc plus faible que celui de la population.
Théoriquement, on sait que la statistique s^2 sous-estime en moyenne σ^2 ; l'estimateur sans-biais de σ^2 étant s_{n-1}^2.

iv. L'histogramme des distances de Kolgomorov-Smirnov pour l'exercice 1 est illustré à la Figure 9 ci-dessous.

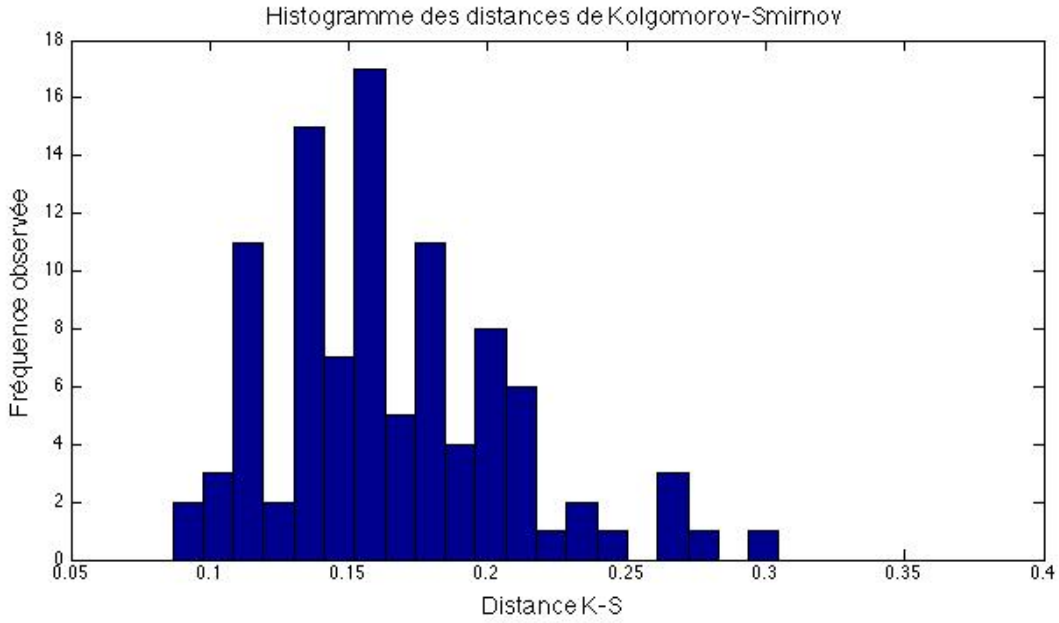


Figure 9. Distances de Kolgomorov-Smirnov de l'exercice 1

v. Les histogrammes des distances K-S pour l'exercice 2 et 3 sont présentés respectivement aux Figure 10 et Figure 11 ci-dessous.

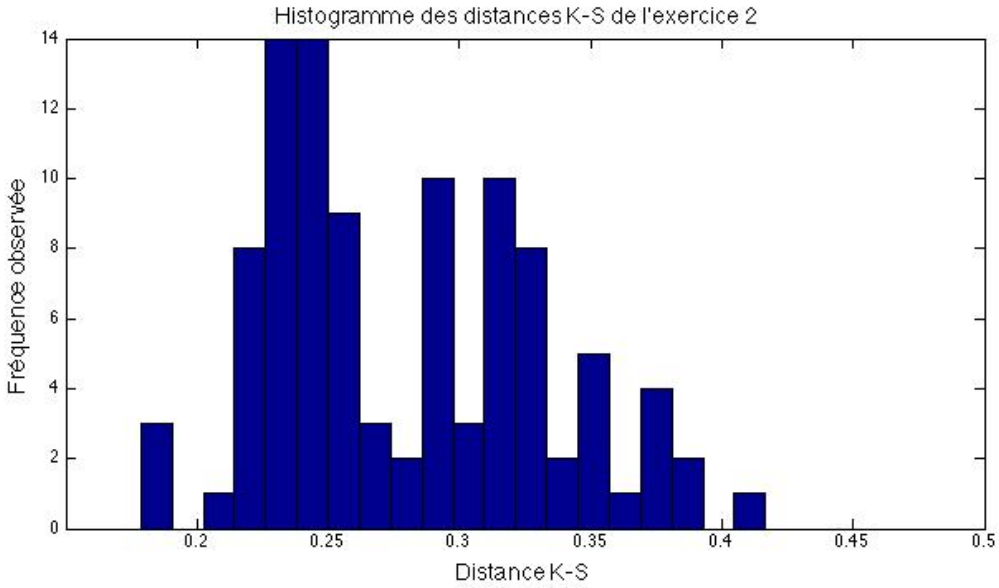


Figure 10. Distances K-S de l'exercice 2

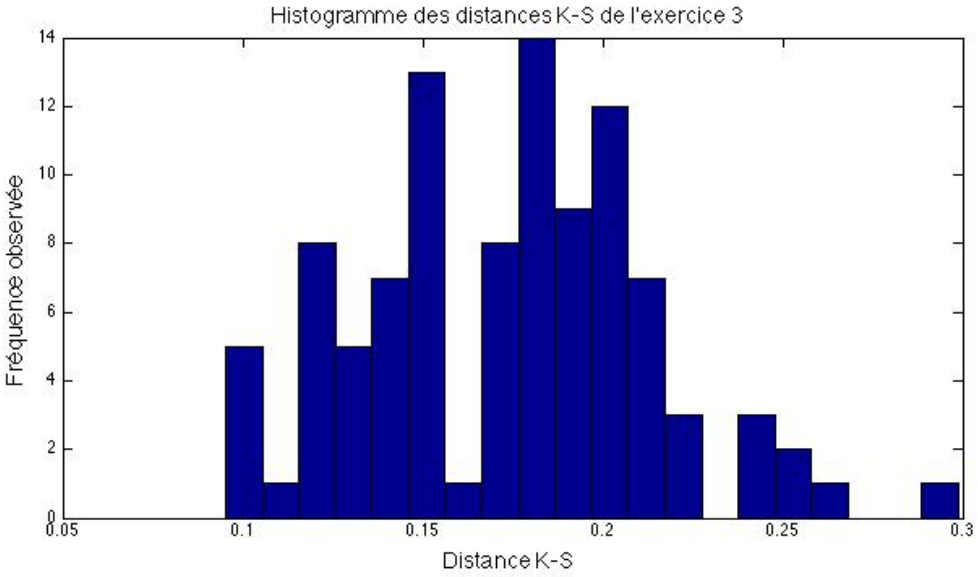


Figure 11.Distances K-S de l'exercice 3

Les trois histogrammes des distances K-S présentent la même allure qui fait encore une fois penser à l'allure du graphe d'une loi normale.

Les trois nouvelles variables calculées au point iv. et v. (distances K-S entre les fonctions de répartition empirique des échantillons et de la population) sont des variables représentant les statistiques de Kolgomorov-Smirnov, notées $D_{n,\chi}^{K\check{S}}$. Or, la distribution d'échantillonnage de $D_{n,\chi}^{K\check{S}}$ ne dépend pas de l'allure de F_χ lorsque celle-ci est continue. Les résultats obtenus ayant tous le même ordre de grandeur, cette affirmation est vérifiée.

La distance de Kolgomorov-Smirnov nous donne une idée de la précision de la modélisation de la population par l'échantillon. Par le Théorême de Glivenko-Cantelli, pour un échantillon de taille infinie, la distance de Kolgomorov-Smirnov doit tendre vers 0. Ce sera donc l'exercice pour lequel la moyenne des distances K-S sera la plus petite (la convergence uniforme est la meilleure) qui sera le mieux modélisé par les échantillons. Les moyennes des distances K-S sont de 0.1639 , 0.2829 et 0. 1782 . L'exercice 1 est donc celui qui est le mieux modélisé.

Annexe

Q1, point a à c :

```
function Q1

X = xlsread('Proba1ereSession20132014.xls');

[m n] = size(X);

% Q1a : génération des histogrammes
T1 = X(:,4);
T2 = X(:,5);
T3 = X(:,6);
subplot(1,3,1)
hist(T1,20);
subplot(1,3,2)
hist(T2,20);
axis([0 22 0 18]);
subplot(1,3,3)
hist(T3,20);
axis([0 22 0 25]);
figure

%Q1b : calcul des statistiques uni-variées.
M1 = mean(X(:,7))
M2 = mean(X(:,8))
M3 = mean(X(:,9))

med1 = median(X(:,7))
med2 = median(X(:,8))
med3 = median(X(:,9))

mod1 = mode(X(:,7))
mod2 = mode(X(:,8))
mod3 = mode(X(:,9))

s1 = sqrt(var(X(:,7),1))
s2 = sqrt(var(X(:,8),1))
s3 = sqrt(var(X(:,9),1))

%Q1b : calcul de la proportion de résultats normaux
norm1 = [M1-s1, M1+s1]
norm2 = [M2-s2, M2+s2]
norm3 = [M3-s3, M3+s3]

n1 = 0;
n2 = 0;
n3 = 0;
for i=1:1:m

    if X(i,7)>= norm1(1) && X(i,7)<= norm1(2)
        n1 = n1+1;
    end

    if X(i,8)>= norm2(1) && X(i,8)<= norm2(2)
        n2 = n2+1;
    end

    if X(i,9)>= norm3(1) && X(i,9)<= norm3(2)
```

```
        n3 = n3+1;
    end

end

prop1 = n1/m
prop2 = n2/m
prop3 = n3/m

%Q1c : génération des boxplot et calcul des quartiles

subplot(1,3,1)
moustache1 = boxplot(X(:,1))
subplot(1,3,2)
moustache2 = boxplot(X(:,2))
subplot(1,3,3)
moustache3 = boxplot(X(:,3))

quart1 = quantile(X(:,1),[.25 .50 .75])
quart2 = quantile(X(:,2),[.25 .50 .75])
quart3 = quantile(X(:,3),[.25 .50 .75])

end
```

Q1, point d et e :

```
% Q1d

function Q1d

X = xlsread('Proba1ereSession20132014.xls');

[u v ] = size(X)
n1 = 0;
n2 = 0;

%génération des nouvelles variables et calcul de la proportion entre 12 et 15
for i=1:1:u

moythe(i) = mean(X(i,4:6));

if moythe(i)>=12 && moythe(i)<=15
    n1 = n1+1;
end

moyexo(i) = mean(X(i,7:9));

if moyexo(i)>=12 && moyexo(i)<=15
    n2 = n2+1;
end


end
prop1 = n1/u
prop2 = n2/u

% génération des polygones de fréquences cumulées
subplot(1,2,1)
cdfplot(moythe);
subplot(1,2,2)
```

```
cdfplot(moyexo);

%deuxième méthode de calcul de la proportion entre 12 et 15
[prop_th x_th] = cdfcalc(moythe);
[prop_ex x_ex] = cdfcalc(moyexo);

borneinfth = prop_th(min(find(x_th>12)))
bornesupth = prop_th(min(find(x_th>15)))

borneinfex = prop_ex(min(find(x_ex>12)))
bornesupex = prop_ex(min(find(x_ex>15)))

propestimeeth = bornesupth-borneinfth
propestimeeex = bornesupex-borneinfex

%Q1e
figure;

scatter(X(:,2),X(:,3))

corrcoef(X(:,2),X(:,3))

end
```

Q2 :

```
function Q2

X = xlsread('Proba1ereSession20132014.xls');

[m n] = size(X);

%génération de l'échantillon

echantillon = randsample(m,20,true)

%génération des nouvelles variables

for i=1:1:20

    val_echantillon(i,1:9) = X(echantillon(i),:);
    moytheech(i) = mean(val_echantillon(i,4:6));

end

for k=1:1:m
    moythe(k) = mean(X(k,4:6));
end
%Q2ai
m_ex1_ech = mean(val_echantillon(:,7))
m_ex2_ech = mean(val_echantillon(:,8))
m_ex3_ech = mean(val_echantillon(:,9))

med_ex1_ech = median(val_echantillon(:,7))
med_ex2_ech = median(val_echantillon(:,8))
med_ex3_ech = median(val_echantillon(:,9))

s1 = std(val_echantillon(:,7),1)
s2 = std(val_echantillon(:,8),1)
```

```
s3 = std(val_echantillon(:,9),1)

%Q2aii
subplot(1,3,1);
beard1 = boxplot(val_echantillon(:,1));
ylim([0 20]);
subplot(1,3,2);
beard2 = boxplot(val_echantillon(:,2));
ylim([0 20]);
subplot(1,3,3);
beard3 = boxplot(val_echantillon(:,3));
ylim([0 20]);


quart1 = quantile(val_echantillon(:,1),[.25 .50 .75])
quart2 = quantile(val_echantillon(:,2),[.25 .50 .75])
quart3 = quantile(val_echantillon(:,3),[.25 .50 .75])

%Q2aiii
figure

cdfplot(moytheech)
hold on
cdfplot(moythe)
hold off

ksdist1 = cdfcalc(moytheech)
ksdist2 = cdfcalc(moythe)

[h p dist] = kstest2(ksdist1,ksdist2)

%partie b

pop1 = cdfcalc(X(:,7));
pop2 = cdfcalc(X(:,8));
pop3 = cdfcalc(X(:,9));

%génération des 100 échantillons
for i=1:1:100
    echantillon = randsample(m,20,true);
    for k=1:1:20
        valeur_echantillon(k,1:9) = X(echantillon(k),:);
    end

    ex1 = cdfcalc(valeur_echantillon(:,7));
    ex2 = cdfcalc(valeur_echantillon(:,8));
    ex3 = cdfcalc(valeur_echantillon(:,9));

    [a(i) b(i) dist1(i)] = kstest2(ex1,pop1);
    [c(i) d(i) dist2(i)] = kstest2(ex2,pop2);
    [e(i) f(i) dist3(i)] = kstest2(ex3,pop3);

 moyex1iid(i)= mean(valeur_echantillon(:,7));

 medianeiid(i) = median(valeur_echantillon(:,7));

 siid(i) = std(valeur_echantillon(:,7),1);

end

%Q2bi-ii-iii
figure
```

```
hist(moyex1iid,20);
figure
hist(medianeiid,20);
figure
hist(siid,20);

moyenne = mean(moyex1iid)
moymed = mean(medianeiid)
moyectype=mean(siid)

%Q2biv-v

figure
hist(dist1,20)
mean(dist1)
figure
hist(dist2,20)
mean(dist2)
figure
hist(dist3,20)
mean(dist3)

end
```