Rapport de projet
Elements de statistique
Partie 1

Plumacker Florence

3ème Bachelier Ingénieur Civil

Année académique 2014-2015

Analyse descriptive

(a) Après avoir chargé le fichier excel, on génère les différents histogrammes grâce à la fonction `hist` de MatLab. Les histogrammes trouvés sont représentés à la figure 1. Ces différents graphiques nous montrent que la première question de théorie semble avoir été la mieux faite et la troisième, la moins bien réalisée. Cela peut s'expliquer par le fait que les étudiants commencent le plus souvent par la première question, y répondent au maximum et ont donc moins de temps pour les deux suivantes.

(b) Les moyennes, médianes, modes et écart-types peuvent être générés respectivement à l'aide des fonctions `mean, median, mode` et `std`. Les valeurs trouvées sont répertoriées dans la figure 2.

Il semble ici évident que la question 2 est celle qui a été la mieux réalisée. La question 3, par contre, possède un mode de 0. Cela prouve que de nombreuses personnes n'y ont pas répondu. On peut relier ce résultat à la difficulté des questions.
Des résultats sont considérés comme normaux s'ils sont compris entre $\mu + \sigma$ et $\mu - \sigma$ où μ est la moyenne et σ l'écart-type. Les pourcentages de normalité pour les exercices 1, 2 et 3 sont respectivement 65%, 83.33% et 65.83%. Ces valeurs s'approchent toutes les trois de la valeur théorique de 68%.

(c) Les différentes boîtes à moustaches générées sont reprises à la figure 3. Des valeurs considérées comme aberrantes sont représentées par de petites croix rouges. Dans le cas du projet 2, par exemple, on se rend compte que la seule valeur aberrante est un zéro, venant probablement d'un étudiant qui n'a pas rendu son travail. Cette donnée tirerait clairement la moyenne vers le bas alors qu'il ne s'agit pas d'une donnée représentative du travail des étudiants. Cependant, on peut remarquer que la boîte à moustaches représentant les résultats à la question sur les projets ne possède aucune donnée aberrante. Cela est dû au grand nombre de 0 et de 20. Les quartiles sont repris dans la figure 4. Ceux-ci ont été calculés à l'aide de la fonction `quantile` de MatLab.

(d) Afin de calculer ce polygone des fréquences cumulées, on moyenne d'abord les résultats aux trois questions de théorie ainsi qu'aux trois questions d'exercice pour chaque étudiant. Les différents polygones sont affichés à l'aide de la fonction `cdfplot`. Ceux-ci sont représentés à la figure 5.
À l'aide du curseur, on peut déterminer les valeurs des fréquences cumulées pour une cote égale à 12 (43.33% pour la théorie et 5.67% pour les exercices) et une cote égale à 15 (88.33% pour la théorie et 80.83% pour les exercices). En soustrayant les deux fréquences cumulées, on trouve la proportion d'étudiants ayant un cote comprise entre 12 et 15 :

$88.33\% - 43.33\% = 45\%$ pour la théorie

$80.83\% - 51.67\% = 29.16\%$ pour les exercices

Ces valeurs sont évidemment des approximations étant donné que les polygones ne sont pas des fonctions continues.

(e) On remarque que, bien que la moyenne du projet 2 était fort élevée, la question n'a pas été très bien faite. En effet, de nombreux étudiants ayant eu un note élevée au projet en ont eu une assez faible pour la question. Le coefficient de corrélation trouvé à l'aide de la fonction `corrcoef` est 0.1407. On peut donc dire que notre interprétation était plutôt

juste. En effet, le coefficient de corrélation étant faible, le lien entre les points pour le rapport de projet et pour la question concernant ce projet n'est pas significatif du tout. Ainsi, ce n'est pas parce qu'un élève a bien réussi son rapport que l'on peut en déduire qu'il aura surement bien réussi la question sur le projet.

Génération d'échantillons i.i.d.

(a)

i. À l'aide de la fonction `randsample`, on génère un vecteur aléatoire de 20 nombres parmi 120, le nombre d'étudiants. On reprend ensuite les résultats des étudiants sélectionnés par le vecteur aléatoire. Comme à la question 1.(b), on calcule les moyennes, médianes et écarts-types à l'aide des fonctions appropriées. Ceux-ci sont répertoriés dans la figure 7. Quand on compare ces valeurs aux valeurs trouvées pour la population, on remarque que chacune des moyennes est un peu supérieure. Cela peut vouloir dire que notre échantillon contenait des élèves qui avaient relativement bien réussi les questions de théorie. Les écarts-types sont, quand à eux, généralement plus faibles.

ii. Les boîtes à moustache sont dessinées à l'aide de la fonction `boxplot`. Celles-ci sont reprises à la figure 8. Par rapport à la population, on remarque clairement un rétrécissement des boîtes comme le montre bien les axes. Les médianes sont, pour les projets, égales à celles de la population mais aussi égales au troisième quartiles contrairement à précédemment. Cela montre qu'une grande partie de l'échantillon a eu une note de 18 aux projets. Certains ont eu une note légèrement supérieure et quelques notes aberrantes font redescendre la médiane.

iii. Les polynômes sont obtenus de la même façon qu'à la question 1(c). Ceux-ci sont représentés à la figure 9. La distance de Kolmogorov-Smirnoff, pour cet échantillon, vaut 0.1526. Celle-ci est calculée à l'aide de la fonction `kstest2` appliquée aux fréquences cumulées obtenu avec la fonction `cdfcalc`. Les tendances générales des polynômes restent les mêmes. Cependant, les écarts sont parfois grands comme le montre bien la distance de Kolmogorov qui est quand même élevée.

(b)

i. Chaque échantillon est créé à l'aide de la fonction `randsample`. À l'aide d'une double application de la fonction `mean` aux résultats de l'exercice 1 des étudiants sélectionnés par les échantillons, on obtient une moyenne de 10.7540. Celle-ci est assez proche de la moyenne de la population. L'histogramme généré est repris par la figure 10. Sa forme fait penser à une distribution normale autour de la moyenne.

ii. De la même façon, on prend la moyenne des médianes obtenues pour chaque échantillon. Celle-ci vaut 10.4150. Ce qui est encore une fois assez proche de la valeur de la population qui vaut, elle, 11. L'histogramme est représenté à la figure 11. Cette fois-ci, la forme de l'histogramme est plus vague mais laisse toute fois entrevoir une distribution liée à une loi normale également.

iii. L'écart-type calculé vaut 5.4537 qui est fort proche de 5.6480, l'écart-type calculé pour la population. Celui-ci est cependant plus faible, ce qui se justifie facilement. En effet, chaque moyenne est faite par rapport à un échantillon de 20 étudiants. L'écart-type est donc, logiquement, plus faible car est plus représentatif de chaque donnée, celles-ci étant moins nombreuses. En faisant la moyenne de ceux-ci, il est donc normal de tomber sur une valeur inférieure à celle trouvée pour la population. L'histogramme représentant les écarts-types est repris à la figure 12. Une nouvelle fois, le diagramme fait penser à une loi normale.

iv. L'histogramme demandé est représenté à la figure 13.

v. Les histogrammes pour les exercices 2 et 3 sont représentés respectivement à la figure 14 et 15. Ceux-ci ont globalement la même allure bien que le graphique de l'exercice 2 soit centré sur une valeur de distance un peu plus élevée que les deux autres. En effet, pour l'exercice 2, il est centré sur 0.3 et pour les 1 et 3, ils sont plus centrés sur 0.2. La similitude des graphes montre bien que la distribution d'échantillonnage de la distance ne dépend pas de la fonction des fréquences cumulées, pourvu que celle-ci soit continue.

On peut également déduire des allures de nos graphiques que nos échantillons modélisent mieux les résultats aux exercices 1 et 3 qu'à l'exercice 2. En effet, les distances sont, en moyenne, supérieures pour l'exercice 2.

Codes

`Question_1a.m`

```
function Question_1a(NomDuFichier)
Resultats_Etudiants = xlsread(NomDuFichier);
subplot (221);
LOL1 = hist(Resultats_Etudiants(:,4), 21);%Résultats de la première ...
    question de théorie
bar(LOL1)
ylabel('Nombre d étudiants')
xlabel('Cote sur 20')
title('Histogramme de la première question de théorie')
subplot (222);
LOL = hist(Resultats_Etudiants(:,5), 21);%Résultats de la deuxième ...
    question de théorie
bar(LOL)
ylabel('Nombre d étudiants')
xlabel('Cote sur 20')
title('Histogramme de la deuxième question de théorie')
subplot(223);
LOL3  = hist(Resultats_Etudiants(:,6), 21);%Résultats de la troisième ...
    question de théorie
bar(LOL3)
ylabel('Nombre d étudiants')
xlabel('Cote sur 20')
title('Histogramme de la troisième question de théorie')
```

`Question_1b.m`

```
function C =  Question_1b(NomDuFichier)


Resultats_Etudiants = xlsread(NomDuFichier);
C(:,1) = mean(Resultats_Etudiants(:, 7:9));
C(:,2) = median(Resultats_Etudiants(:, 7:9));
C(:,3) = mode(Resultats_Etudiants(:, 7:9));
C(:,4) = std(Resultats_Etudiants(:, 7:9),1));

%Calcul de l'intervalle normal

Interv(:,2) = C(:,1)+C(:,4)
Interv(:,1) = C(:,1) - C(:,4)

%Calcul du pourcentage compris dans cet intervalle
S = size(Resultats_Etudiants);
for j = 7:9
    Count(j-6) = 0;

for i=1:S(1)
    if Interv(j-6,1)≤ Resultats_Etudiants(i,j) &&  ...
        Resultats_Etudiants(i,j)≤Interv(j-6,2)
        Count (j-6) = Count(j-6) + 1;
    end
end
Norm(j-6) = Count(j-6)/120;
end
Norm
```

Question_1c.m

```

function  Question_1c(NomDuFichier)


Resultats_Etudiants = xlsread(NomDuFichier);

%Affichage des boîtes à moustache
subplot(131) ;
boxplot(Resultats_Etudiants(:,1));
title('Projet 1');
subplot(132);
boxplot(Resultats_Etudiants(:,2));
title('Projet 2');
subplot(133);
boxplot(Resultats_Etudiants(:,3));
title('Question sur le projet');

%Calcul des quartiles
Quartile2(1) = quantile(Resultats_Etudiants(:,1),0.5);
Quartile2(2) = quantile(Resultats_Etudiants(:,2),0.5);
Quartile2(3) = quantile(Resultats_Etudiants(:,3),0.5);
Quartile1(1) = quantile(Resultats_Etudiants(:,1),0.25);
Quartile1(2) = quantile(Resultats_Etudiants(:,2),0.25);
Quartile1(3) = quantile(Resultats_Etudiants(:,3),0.25);
Quartile3(1) = quantile(Resultats_Etudiants(:,1),0.75);
Quartile3(2) = quantile(Resultats_Etudiants(:,2),0.75);
Quartile3(3) = quantile(Resultats_Etudiants(:,3),0.75);
Quartile1
```

```
29 Quartile2
30 Quartile3
```

Question_1d.m

```
1  function FreqCum =  Question_1d(NomDuFichier)
2  Resultats_Etudiants = xlsread(NomDuFichier);
3
4  %Calcul des moyennes
5  for i = 1:120
6  MoyenTheo(i) = mean(Resultats_Etudiants(i,4:6));
7  MoyenExo(i) = mean(Resultats_Etudiants(i,7:9));
8  end
9
10
11 %Affichage des polygones
12 subplot(121)
13 cdfplot(MoyenTheo)
14 title('Polygone des fréquences cumulées pour la théorie')
15 xlabel('Cote sur 20')
16 ylabel('Fréquences cumulées')
17 subplot(122)
18 cdfplot(MoyenExo)
19 title('Polygone des fréquences cumulées pour les excercices')
20 xlabel('Cote sur 20')
21 ylabel('Fréquences cumulées')
22 FreqCum = cdfcalc(MoyenTheo);
23 end
```

Question_1e.m

```
1
2  function Question_1e(NomDuFichier)
3
4
5  Resultats_Etudiants = xlsread(NomDuFichier);
6  scatter(Resultats_Etudiants(:,2), Resultats_Etudiants(:,3));
7  corrcoef(Resultats_Etudiants(:,2), Resultats_Etudiants(:,3));
8  end
```

Question_2a.m

```
1  function C = Question_2a(NomDuFichier)
2  Resultats_Etudiants = xlsread(NomDuFichier);
3  A = size(Resultats_Etudiants);
4
5  %%Génération du vecteur aléatoire.
6  Vecteur_Aleatoire = randsample (A(1), 20, true);
7  Echantillon = Resultats_Etudiants(Vecteur_Aleatoire,:);
8
9  %%Calcul des moyennes, médiannes, écart-types
10 C(:,1) = mean(Echantillon(:, 7:9));
11 C(:,2) = median(Echantillon(:, 7:9));
```

```
C(:,3) = std(Echantillon(:, 7:9),1);

%%Affichage des boites à Moustache
subplot(131) ;
boxplot(Echantillon(:,1));
title('Projet 1');
subplot(132);
boxplot(Echantillon(:,2));
title('Projet 2');
subplot(133);
boxplot(Echantillon(:,3));
title('Question sur le projet');

%%Polygones des fréquences cumulées
for i = 1:20
MoyenEchTheo(i) = mean(Echantillon(i,4:6));
end


figure
FreqcumEchT = cdfcalc(MoyenEchTheo);

cdfplot(MoyenEchTheo)
title('Polygone des fréquences cumulées des résultats de la théorie ...
    pour un échantillon')

xlabel('Cote sur 20')
ylabel('Fréquences cumulées')



%Distance de Kolmogorov–Smirnov

FreqCumPop =  Question_1d(NomDuFichier);
[¬,¬, Kolmo] = kstest2(FreqcumEchT, FreqCumPop);
DistKolmo = max(Kolmo)
end
```

Question_2b.m

```

function FreqCum =  Question_1d(NomDuFichier)
Resultats_Etudiants = xlsread(NomDuFichier);

%Calcul des moyennes
for i = 1:120
MoyenTheo(i) = mean(Resultats_Etudiants(i,4:6));
MoyenExo(i) = mean(Resultats_Etudiants(i,7:9));
end


%Affichage des polygones
subplot(121)
cdfplot(MoyenTheo)
title('Polygone des fréquences cumulées pour la théorie')
xlabel('Cote sur 20')
ylabel('Fréquences cumulées')
subplot(122)
cdfplot(MoyenExo)
```

```
20 title('Polygone des fréquences cumulées pour les excercices')
21 xlabel('Cote sur 20')
22 ylabel('Fréquences cumulées')
23 FreqCum = cdfcalc(MoyenTheo);
24 end
```

Figures

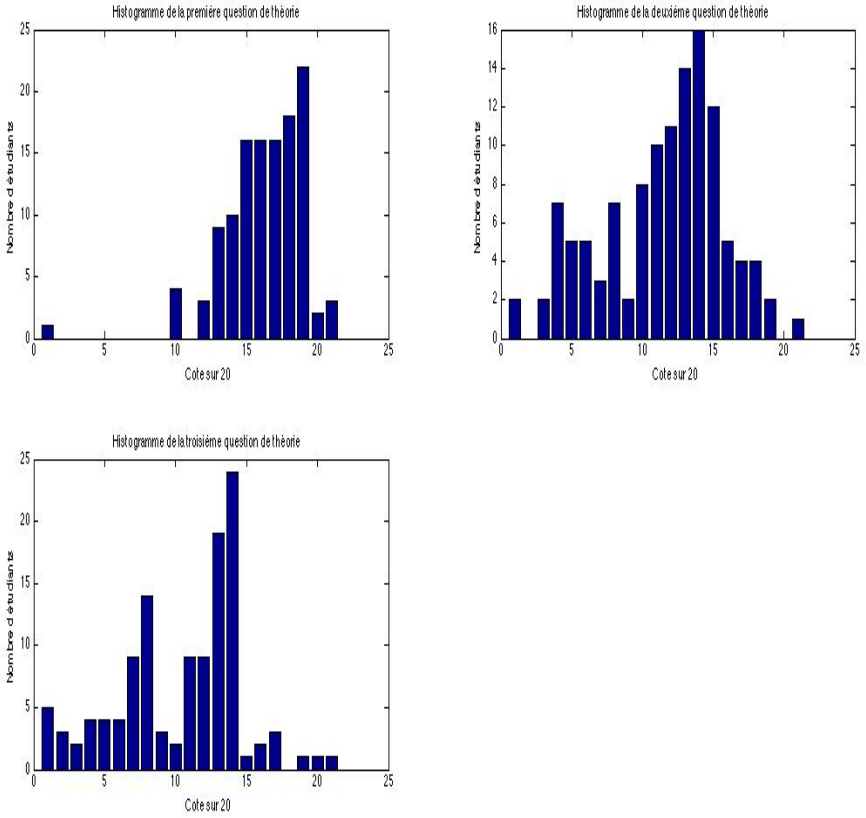


FIGURE 1 – Histogramme des questions de théorie

	Moyenne	Médiane	Mode	Ecart-type	Plage de normalité
1ère question	10.8167	11.0000	12.0000	5.648	[5.1687 ;16.4647]
2ème question	16.8083	18.0000	20.0000	3.7866	[13.0217 ;20.5949]
3ème question	7.7333	7.5000	0	5.2816	[2.4517 ;13.0150]

FIGURE 2 – Moyenne, médiane, mode et écart-type des résultats aux exercices

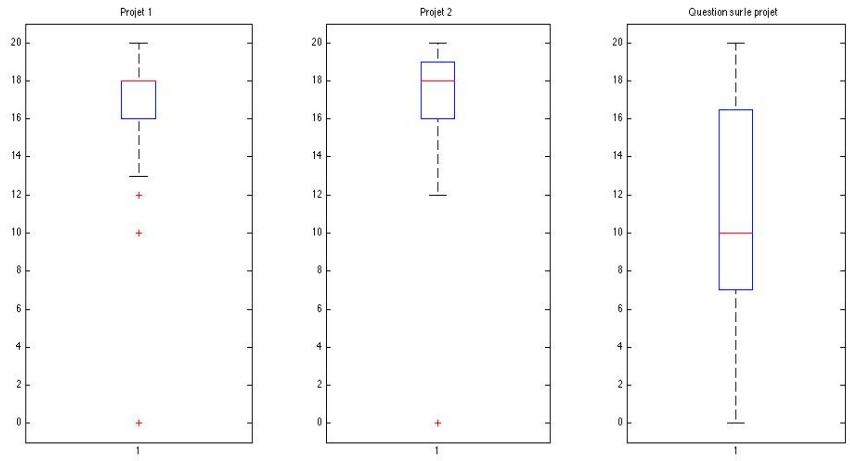


FIGURE 3 – boîtes à moustaches relatives aux résultats des projets.

	Quartile 1	Quartile 2(Médiane)	Quartile 3
Projet 1	16	18	18
Projet 2	16	18	19
Question sur les projets	7	10	16.5

FIGURE 4 – Différents quartiles des résultats de projet

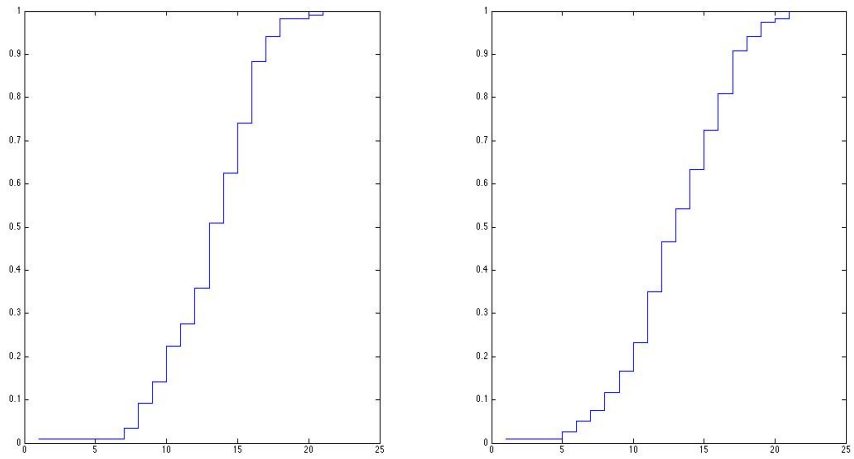
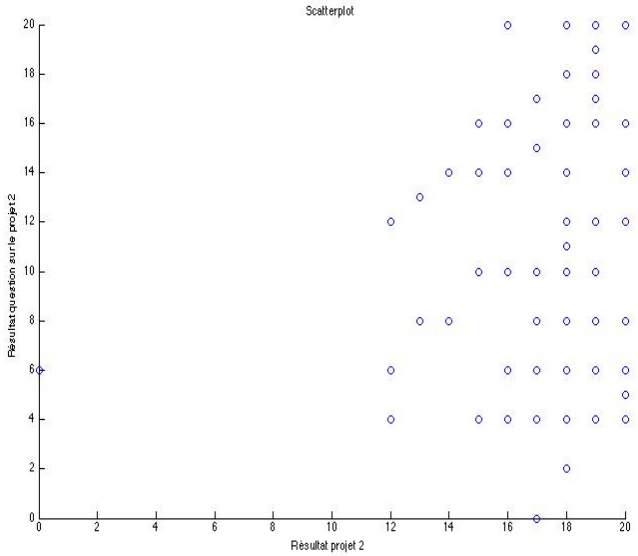


FIGURE 5 – Polygones des fréquences cumulées



,

FIGURE 6 – Scatterplot comparant les résultats obtenus au rapport du projet 2 et les résultats obtenus lors de la question sur le projet 2.

	Moyenne	Médiane	Ecart-type
1ère question	10.8500	10.5000	4.3682
2ème question	17.0500	18.0000	3.7902
3ème question	8.2000	8.0000	5.7363

FIGURE 7 – Moyenne, médiane et écart-type des résultats aux questions de théorie pour l'échantillon

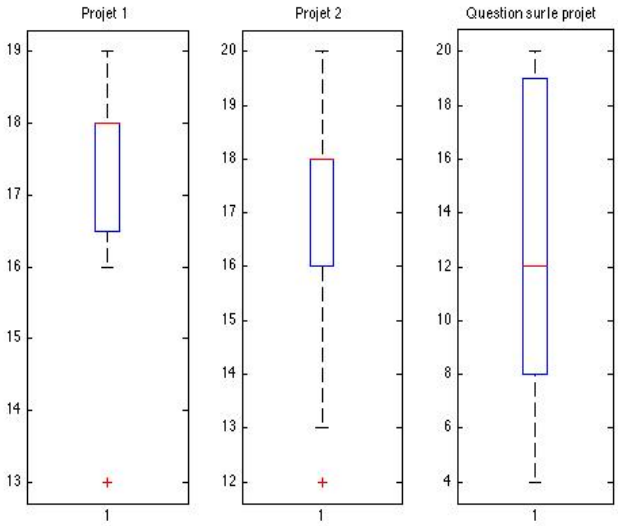


FIGURE 8 – Boites à moustaches relatives aux résultats des projets.

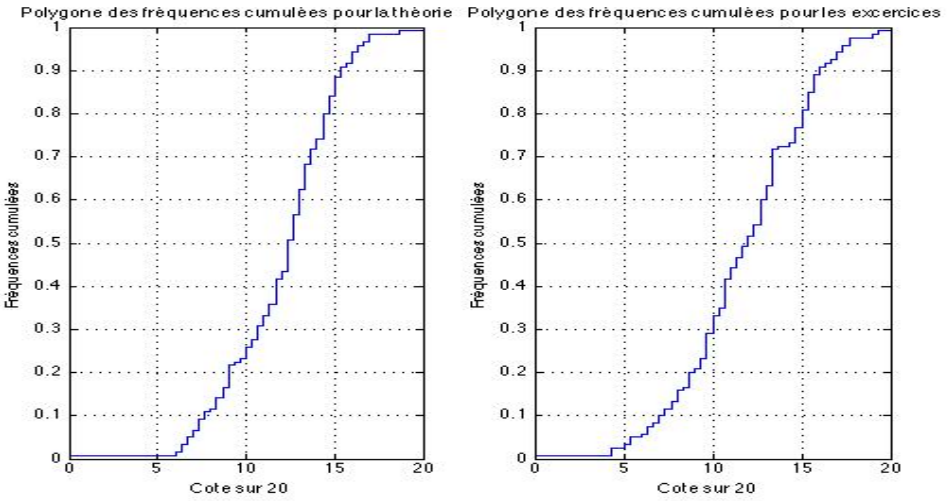


FIGURE 9 – Polygones des fréquences cumulées de l'échantillon

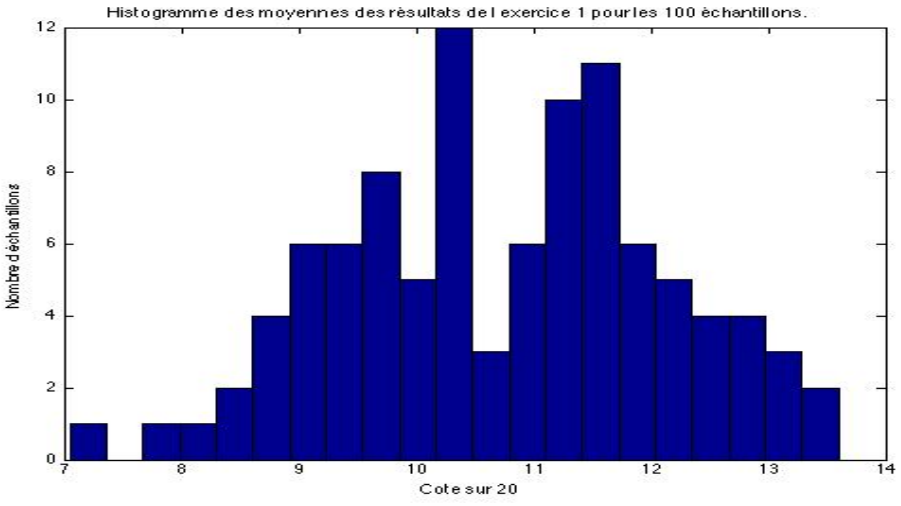


FIGURE 10 – Histogramme des résultats de l'exercice 1 pour les 100 échantillons.

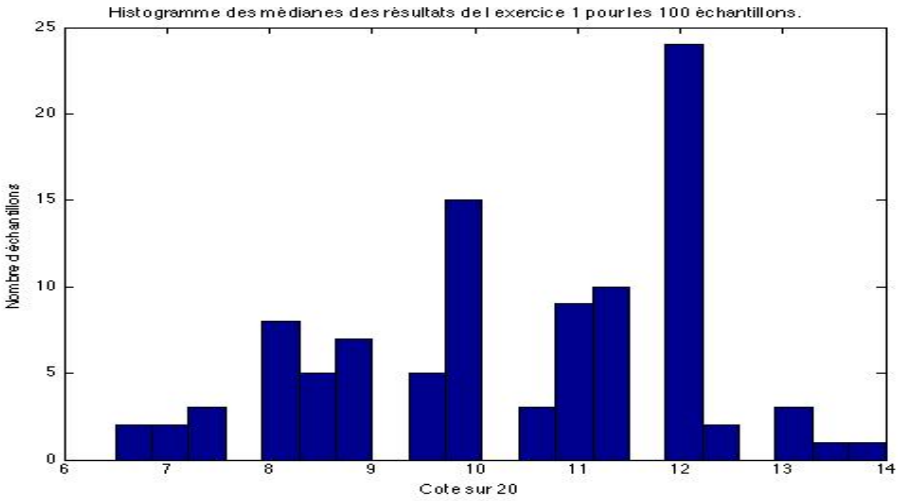


FIGURE 11 – Histogramme des médianes des résultats de l'exercice 1 pour les 100 échantillons.

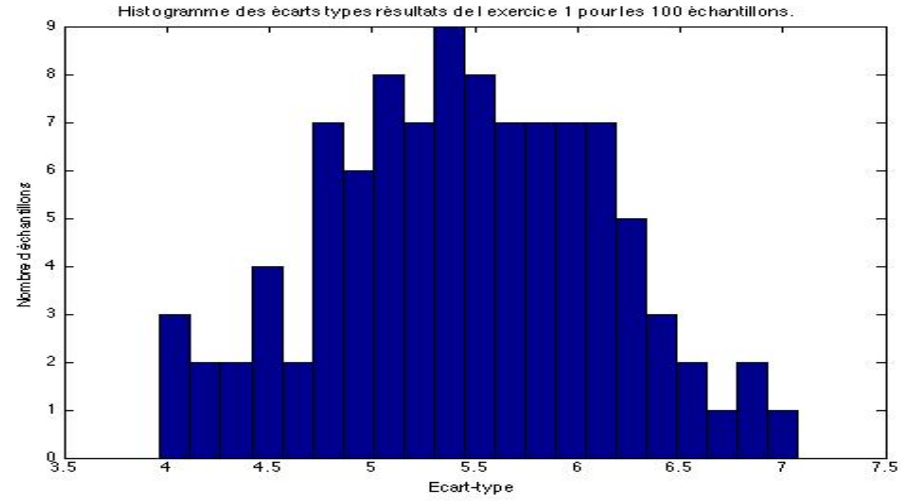


FIGURE 12 – Histogramme des écarts-types des résultats de l'exercice 1 pour les 100 échantillons.

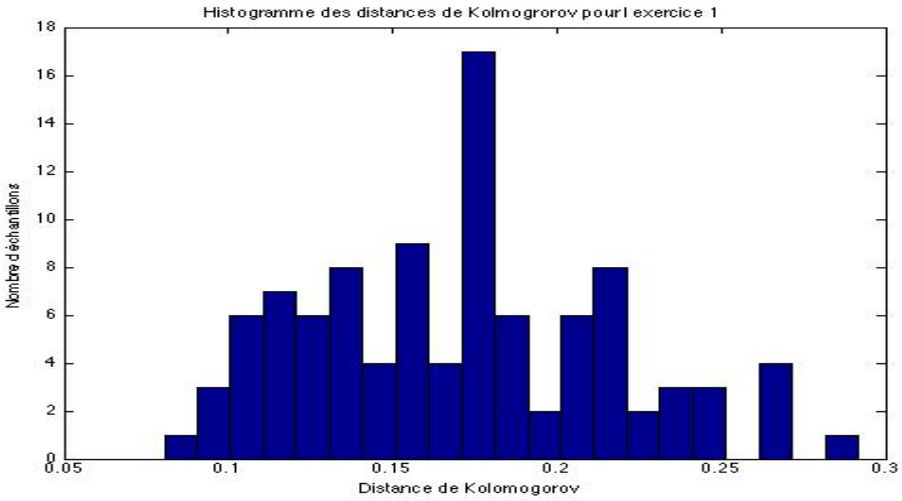


FIGURE 13 – Histogramme des distances de Kolmogrorov pour l'exercice 1.

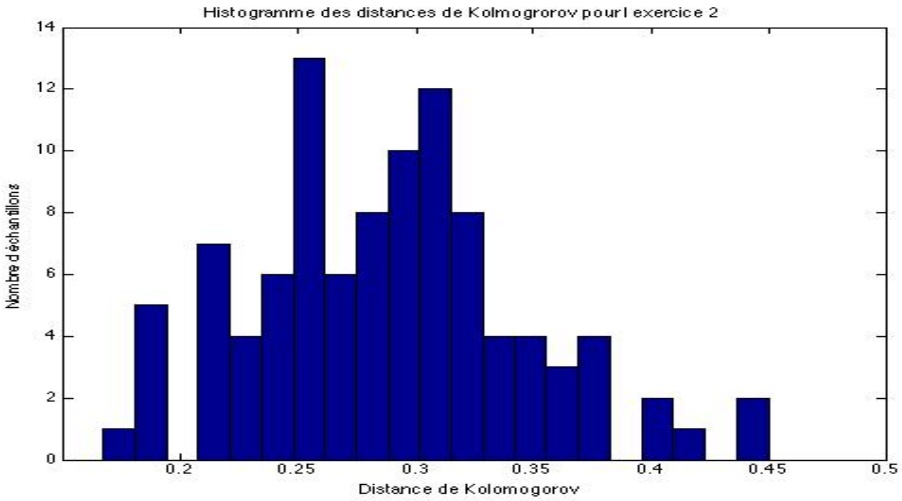


FIGURE 14 – Histogramme des distances de Kolmogrorov pour l'exercice 2.

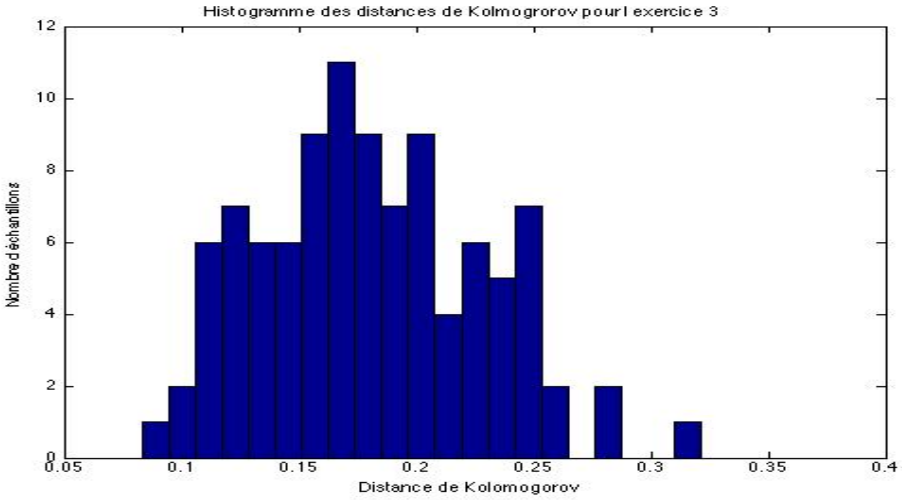


FIGURE 15 – Histogramme des distances de Kolmogrorov pour l'exercice 3.