MATH0487-1 - Eléments de statistiques : Projet 1 - Partie 1

Laurent Portelange - s103355

24 octobre 2013

Questions

1. Analyse descriptive

(a) Le code des trois histogrammes est disponible dans le fichier Q1a.m ainsi qu'en Annexe.
Les trois histogrammes sont disponibles aux Figures 1, 2 et 3. Globalement, on remarque que la question 1 de théorie est celle qui a été la mieux réussie. En effet, peu d'élèves ont une note inférieure à 10/20.

Concernant la question 2 de théorie, on voit qu'elle a été un peu bien moins faite mais les résultats restent acceptables, avec un certains nombre qui ont réussi à atteindre une très bonne note(18/20 minimum). La note de la plupart des élèves se situe entre 8/20 et 15/20.

Par contre, pour la question 3 de théorie, l'histogramme nous montre que le résultat des élèves est plus uniforme que les 2 questions précédentes. Il y a environ le même nombre d'étudiants qui ont réussi ou qui ont raté. Mais les résultats élevés (note supérieure ou égale à 18/20) sont assez rares. La fréquence relative de chaque note est globalement la même. Il y a notamment quelques données aberrantes. Le nombre d'étudiant ayant une cote égale à 0 est presque deux fois plus importante que n'importe laquelle des autres cotes, ainsi qu'une forte diminution en 5, 18, 19 et 20/20.
On peut ainsi tirer comme conclusion que la question 1 est celle qui a été la mieux réussie, suivie de la question 2. La question 3 est, quant à elle, celle qui a été la moins bien faite, c'est-à-dire où il y a le plus grand nombre d'échecs.

(b) Les trois moyennes, médianes, modes et écart-types des résultats des exercices sont disponibles dans la Table 1. Le code pour les obtenir est disponible dans le fichier Q1b.m.

La moyenne nous montre directement que l'exercice 2 est l'exercice qui a le mieux été réussi. Les exercices 1 et 3 semblent avoir posé problème aux étudiants. La médiane est fort proche de la moyenne, cela signifie qu'il y a une répartition à peu près symétrique des résultats obtenus. De plus, l'écart-type est fort similaire pour les 3 exercices, cela montre que les résultats se concentrent plus ou moins de la même façon autour de la moyenne pour chaque exercice. Enfin, le mode est une information supplémentaire qui confirme ce qui a été dit plus haut et suit la tendance prévue, c'est-à-dire que l'exercice 2 a été très bien compris par la majorité des élèves tandis que les exercices 1 et 3 ont posé problème (surtout le 3 avec des résultats catastophiques).

Les résultats "normaux" sont ceux qui se trouvent l'intervalle [moyenne - écart-type ; moyenne + écart-type]. Compte tenu des valeurs obtenues précédemment, on peut calculer cet intervalle pour chaque exercice.

Intervalle de l'exercice 1 = [3.537 ; 14.1116]
Intervalle de l'exercice 2 = [10.384 ; 19.71]
Intervalle de l'exercice 3 = [1.4753 ; 9.6869]

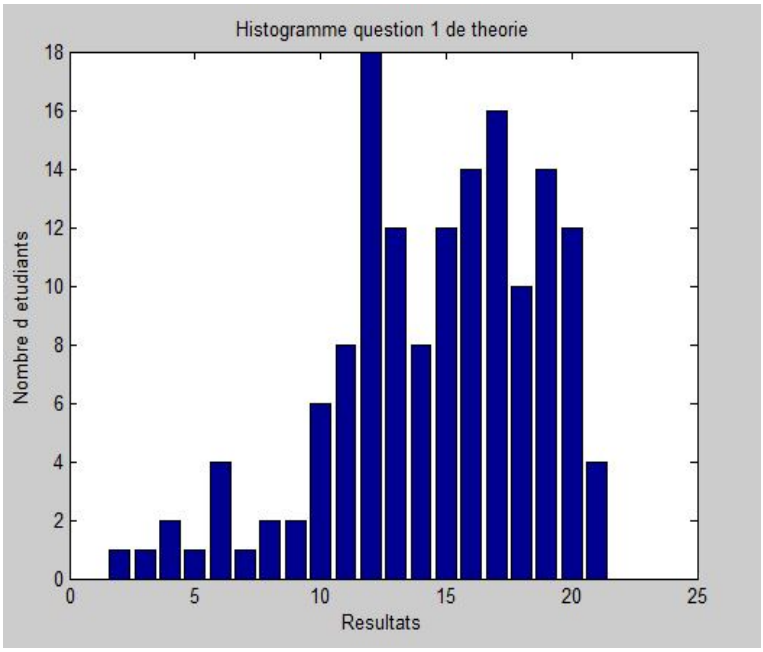

FIGURE 1

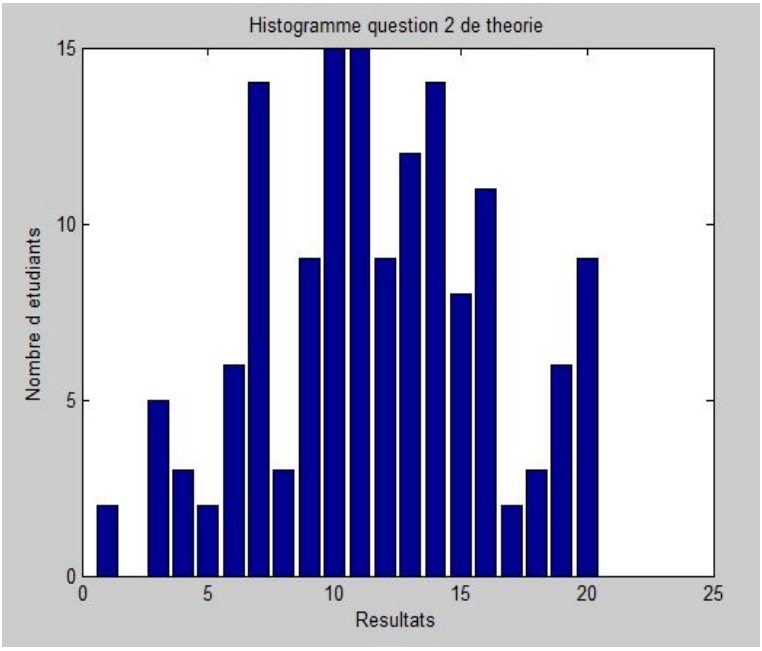

FIGURE 2

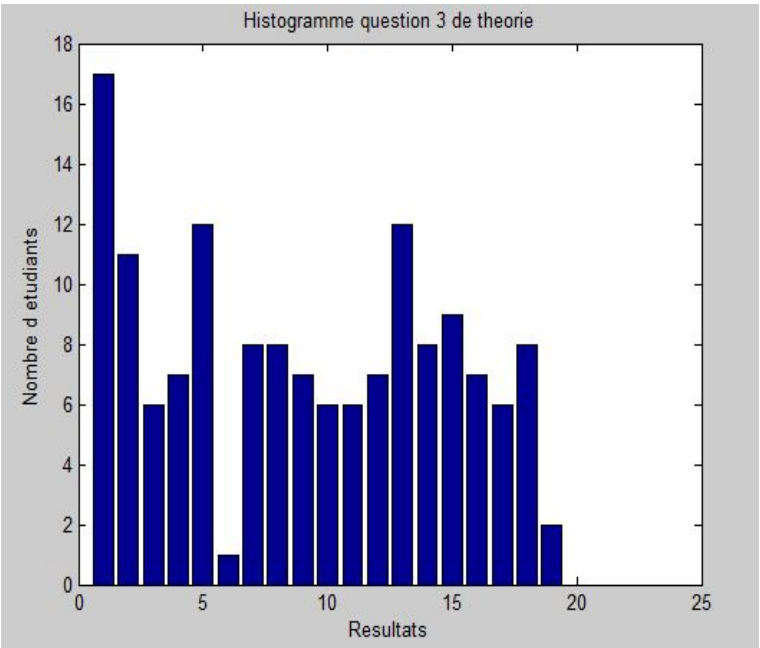

FIGURE 3

$Exercice$	$Moyenne$	$Médiane$	$Mode$	$Ecart-type$
1	8.8243	8	6	5.2873
2	15.0473	16	20	4.6633
3	5.5811	5	2	4.1058

TABLE 1

Pour connaitre la proportion d'étudiants ayant réalisé un travail "normal", il faut connaitre la proportion d'étudiants ayant une cote comprise dans les intervalles définis précédemment. Pour ce faire, je vais calculer les polygones des fréquences cumulées et faire la différence des ordonnées des points dont l'abscisse correspond aux bornes de l'intervalle. Le code pour obtenir ces polygones est disponible dans le fichier Q1b.m . En prenant les points (3.537,0.1757) et (14.1116, 0.8243) pour l'exercice 1, (10.384,0.1757) et (19.71,0.75) pour l'exercice 2 et (1.4753,0.1284) et (9.6869, 0.8649) pour l'exercice 3, on trouve une proportion de 64.86%, 57.43 % et 73.65 % respectivement pour l'exercice 1, 2 et 3.

(c) Les 4 boites à moustaches relatives aux résultats du projet sont disponibles aux figures 4, 5, 6 et 7. On peut voir qu'il y a effectivement des données aberrantes, représentées par des croix rouges. De plus, on remarque que le bas de la boite et la moustache inférieur se confondent. Cela est dû au grand nombre d'étudiants ayant eu une note nulle à cette question. Les valeurs des quartiles sont disponibles à la Table 2.

(d) Les polygones des fréquences cumulées des deux nouvelles variables sont disponibles aux figures 8 et 9. On peut estimer la proportion d'étudiants ayant une cote comprise entre 12 et 15 grâce à ces graphiques. Pour celle concernant la théorie, en prenant comme référence les points (12,0.6081) et (15, 0.8649) et en faisant la différence des ordonnées, on trouve une proportion d'étudiants de 25.68%. Pour celle concernant les exercices, en prenant comme référence les points (12,0.7432) et (15, 0.9257), et en faisant la différence des ordonnées, on trouve une proportion d'étudiants de 18.25%.

(e) Le scatterplot est disponible à la Figure 10. Avec la fonction corrcoef, on trouve un coefficient de corrélation de 0.2126. Ce coefficient exprime la causalité qu'il peut exister entre les deux variables. Ce coefficient étant faible, on peut donc dire que ces variables ne sont pas fortement liées. Cela est

	$Q1$	$Q2$	$Q3$
$Projet1$	16	18	19
$Projet2$	16	18	19
$Projet3$	17	18	19
$Projet4$	0	0	18

TABLE 2

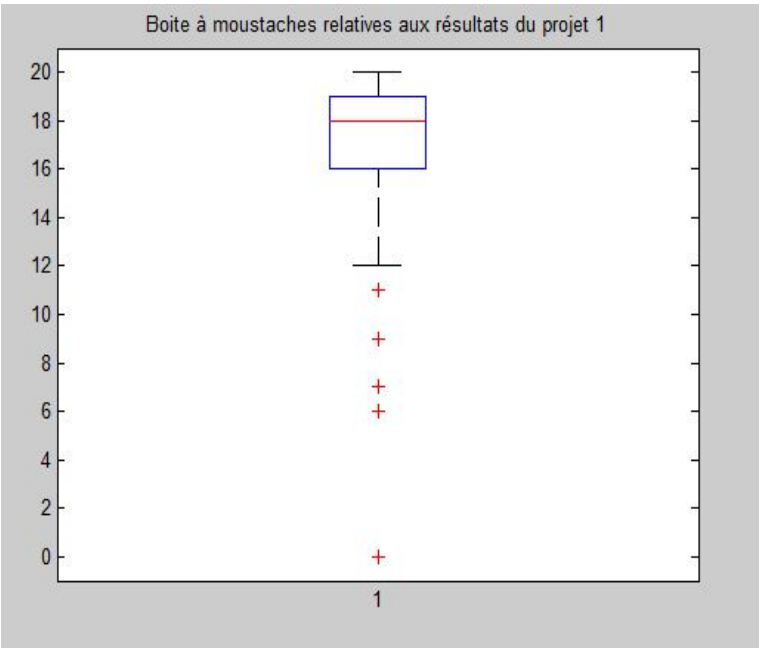

FIGURE 4

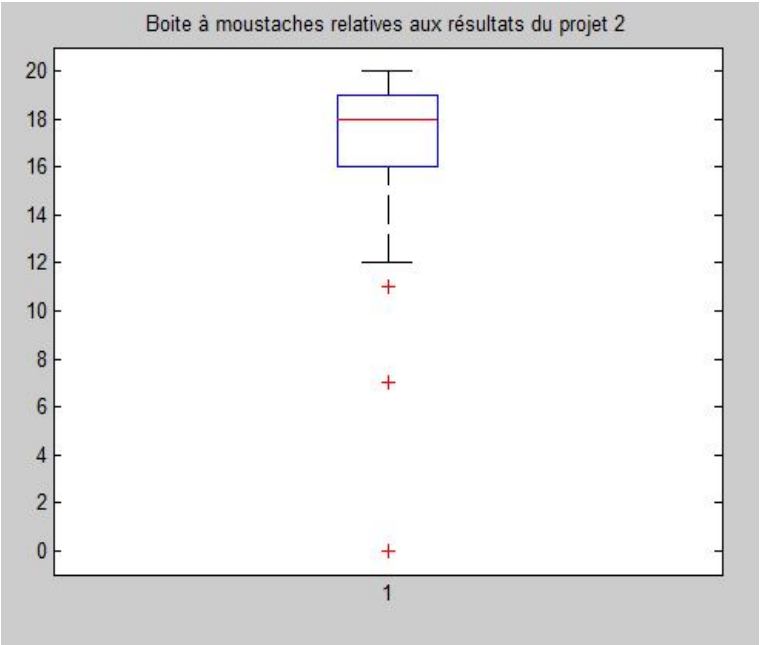

FIGURE 5

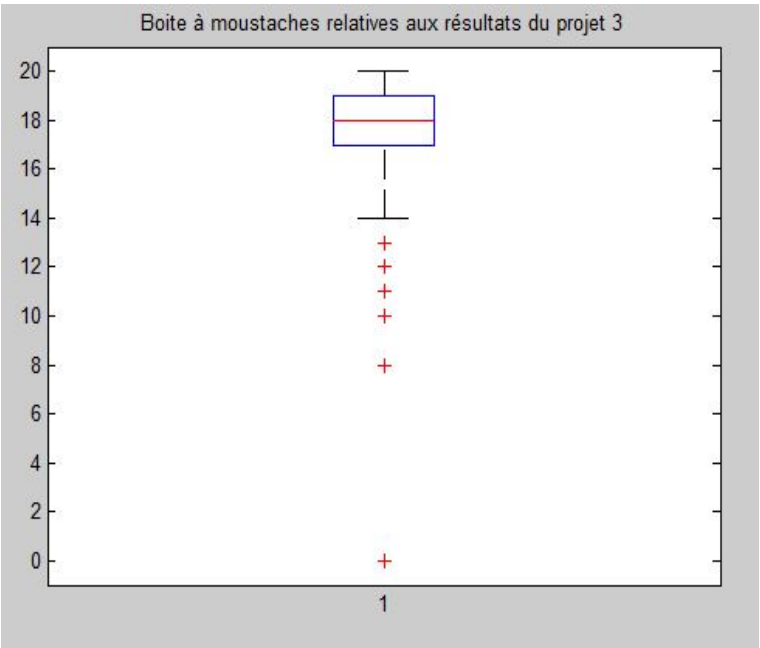

FIGURE 6

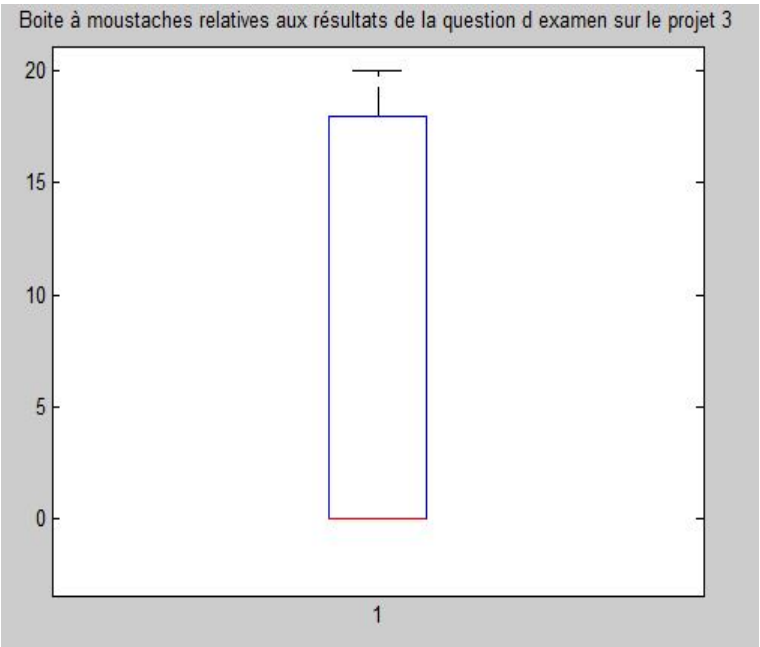

FIGURE 7

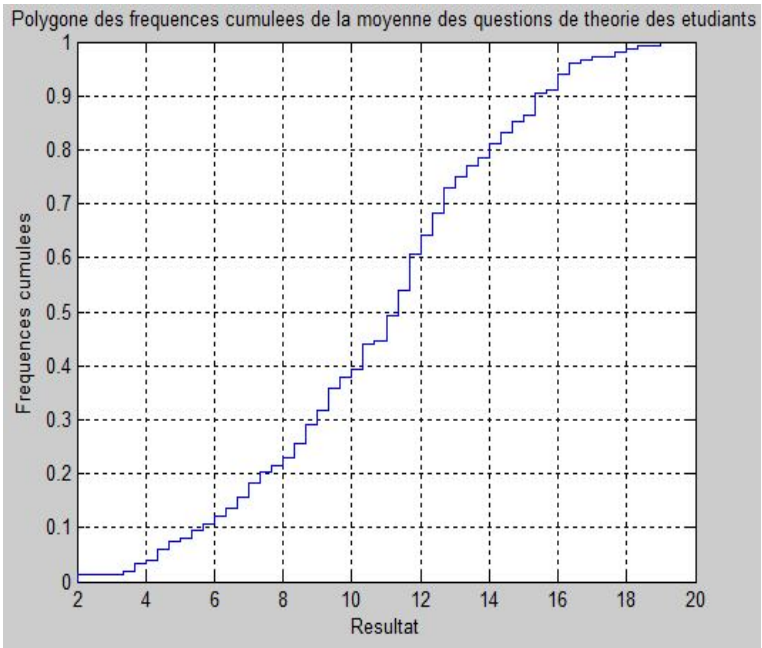

FIGURE 8

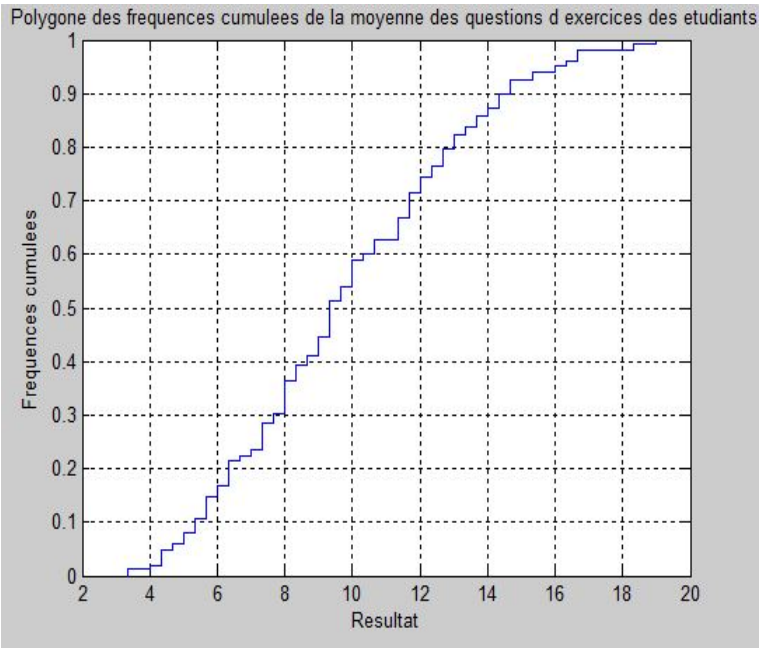

FIGURE 9

Exercice	*Moyenne*	*Médiane*	*Mode*	*Ecart − type*
1	8	6.5	3	6.2070
2	15.8	16.5	20	3.6216
3	6.1	5	5	4.6442

TABLE 3

dû par le nombre important de données aberrantes. En effet, sur le scatterplot, on peut voir qu'un certain nombre non négligeable d'étudiants ont une cote supérieure ou égale à 10/20 pour le projet 3, mais ont une cote nulle pour la question d'examen portant sur le projet. Cela fait donc fortement baisser le coefficient de corrélation.

2. Génération d'échantillons i.i.d.

(a) i. Les résultats de l'échantillon i.i.d. des 20 étudiants est disponible à la Table 3.
On remarque que les résultats de l'échantillon sont fort similaires aux résultats de la population. On peut donc considérer cet échantillon comme étant une bonne approximation du comportement générale des étudiants.

ii. On remarque directement la diminution du nombre de données aberrantes, due à la petite taille de l'échantillon i.d.d. Les boites à moustaches, disponibles aux Figures 11, 12, 13 et 14, sont relativement similaires, sauf celle qui concerne la question du projet de l'examen où le troisième quartile est nettement inférieur à celle de la population. Cela est probablement dû à un manque de données aberrantes.

iii. On voit qu'en général, le polygone de l'échantillon, disponible à la Figure 15, suit celui de la population. Sauf aux limites mais cela est dû au fait qu'il n'y a pas de résultat considéré comme "extrême" dans l'échantillon. Le code pour calculer la distance de Kolomogorov est disponible dans le fichier Q2a.m ainsi qu'en Annexe. On trouve une distance de 0.1797.

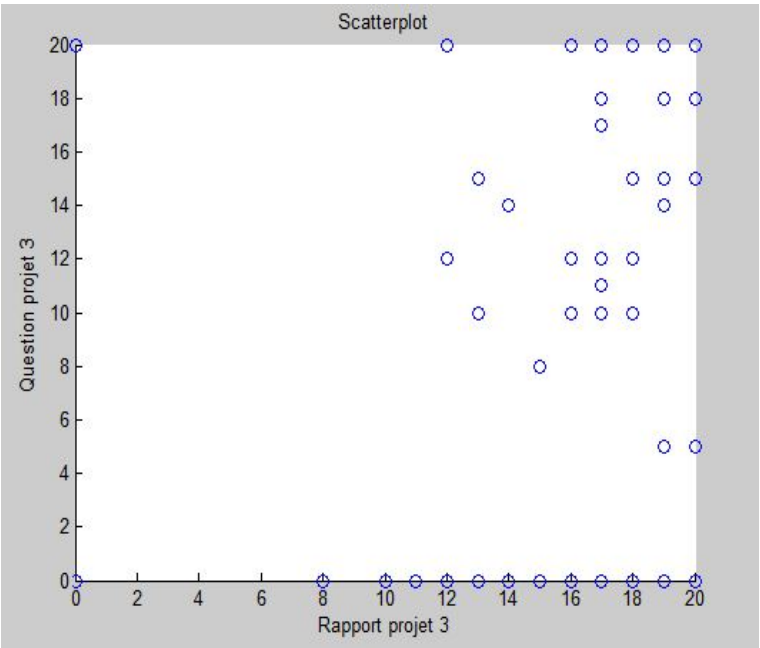

FIGURE 10

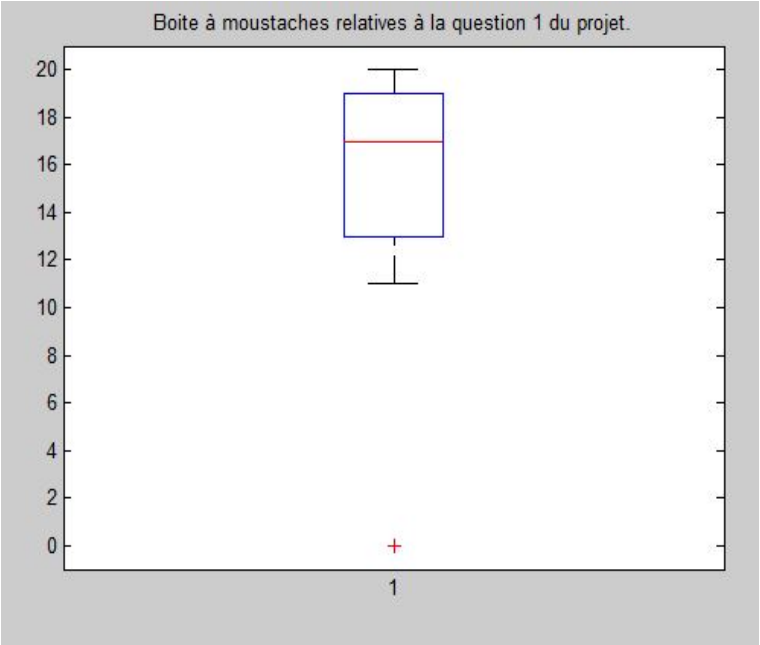

FIGURE 11

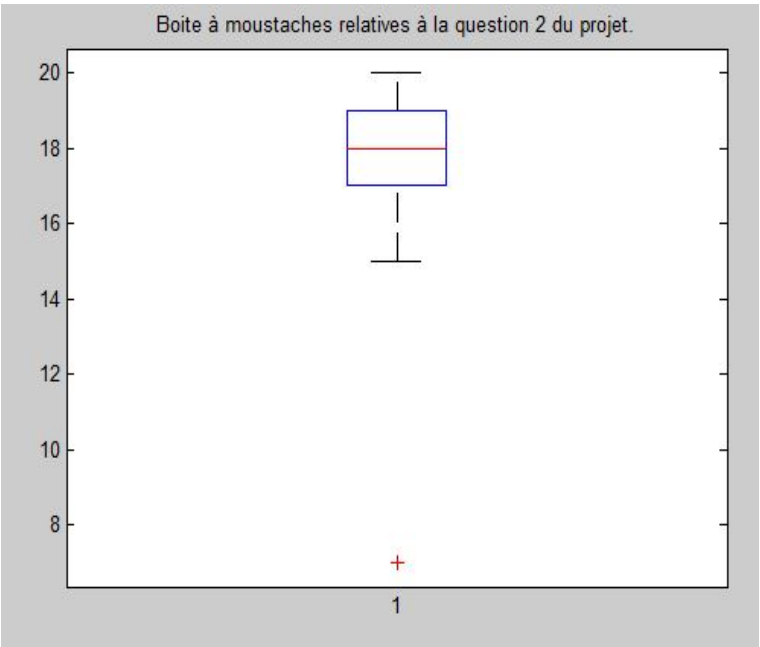

FIGURE 12

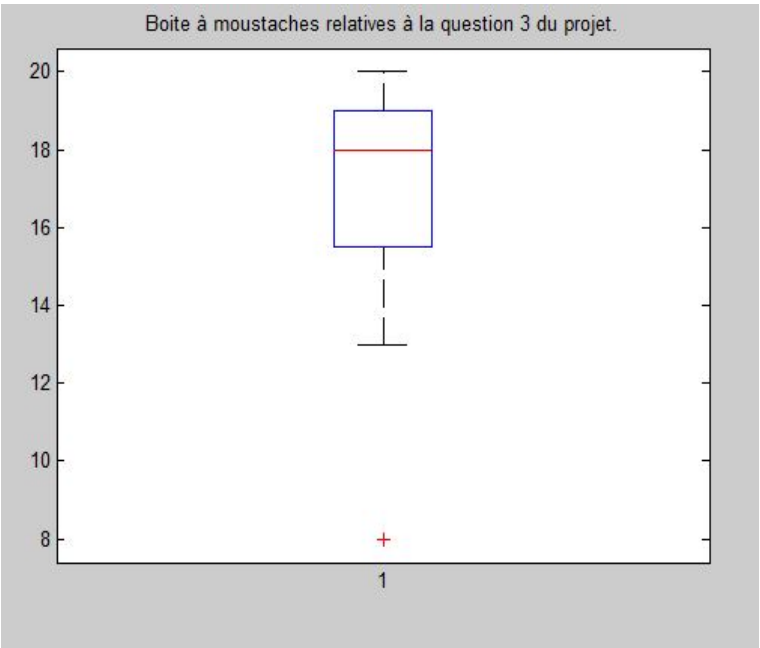

FIGURE 13

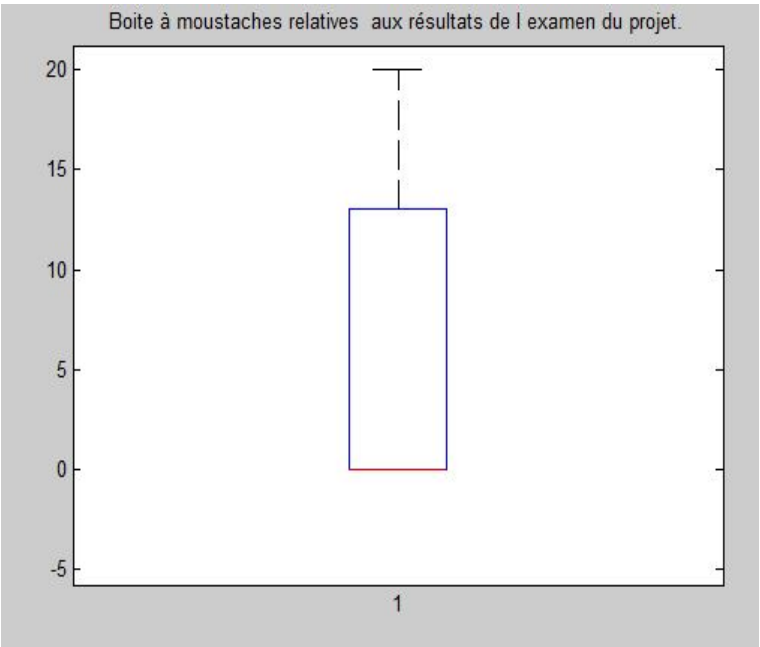

FIGURE 14

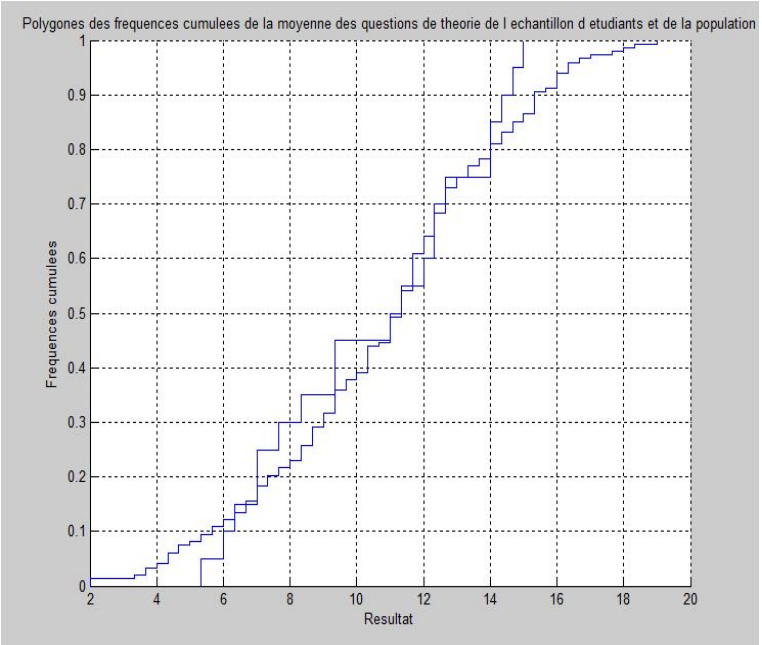

FIGURE 15

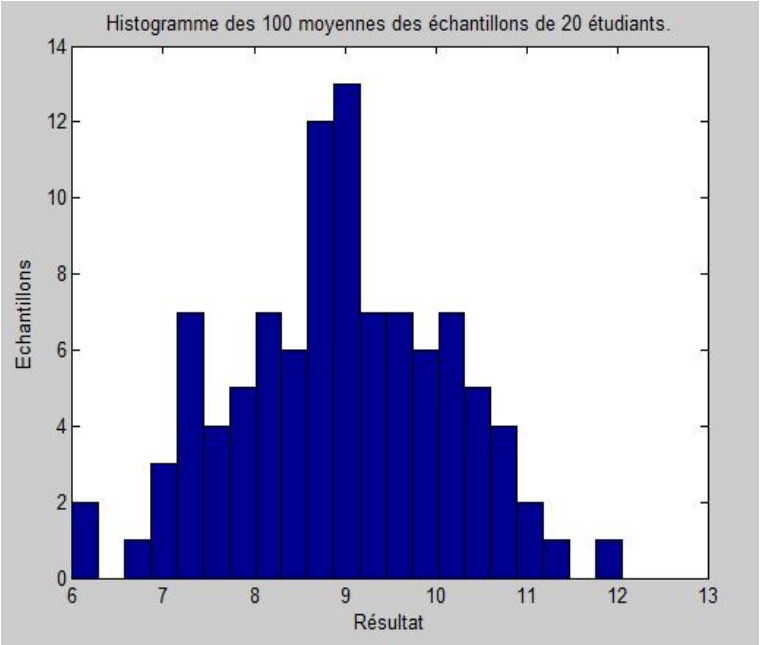

FIGURE 16

(b) i. La moyenne de la nouvelle variable vaut 8.8574 et celle de la population est de 8.8243. On peut donc en conclure qu'elles sont fort proches. L'allure de l'histogramme à la Figure 16 ressemble à une distribution de la loi normale, ce qui montre bien que la moyenne obtenue à cet histogramme tend bien vers la moyenne de la population.

ii. L'allure de l'histogramme, disponible à la figure 17 est parsemée de trous et ne fait penser à aucune loi connue. La moyenne de la nouvelle variable vaut 8.2144, on voit donc qu'elle est plus proche de la moyenne obtenue par la population à lŠexercice 1 que la valeur calculée à la fin du point précédent.

iii. L'allure de l'histogramme, disponible à la figure 18 ressemble vaguement à une distribution de la loi normale malgré quelques "sauts" dans la courbe et quelques données aberrantes. La moyenne de la nouvelle variable vaut 5.2374 et celle de la population est de 5.2873. Elles sont donc fort proches. Puisqu'on a affaire à une distribution de la loi normale, il est normal que la moyenne obtenue tende bien vers celle de la population.

iv. L'histogramme de la variable vector_kolmo1 qui contient les distances de Kolmogorov Smirnov obtenues concernant l'exercice 1 est disponible à la figure 19.

v. L'histogramme de la variable vector_kolmo2 qui contient les distances de Kolmogorov Smirnov obtenues concernant l'exercice 2 est disponible à la figure 20. L'histogramme de la variable vector_kolmo3 qui contient les distances de Kolmogorov Smirnov obtenues concernant l'exercice 3 est disponible à la figure 21. On constate que les histogrammes des distances obtenues pour l'exercice 1 et 2 sont continus (il n'y a pas de trous) et ont une allure un peu similaire, on peut donc en conclure qu'ils suivent une même loi de répartition et que les échantillons reflètent bien le comportement de la population. Par contre, le troisième histogramme n'est pas continu et est parsemé de trous. Il ne ressemble pas aux 2 histogrammes précédents. Il suit donc une loi différente des deux autres. De plus, le "pic" du troisième histogramme est plus élevé que celui des 2 autres. Cela nous montre que les échantillons utilisés pour réaliser le troisième histogramme ne sont pas très représentatif de la population.

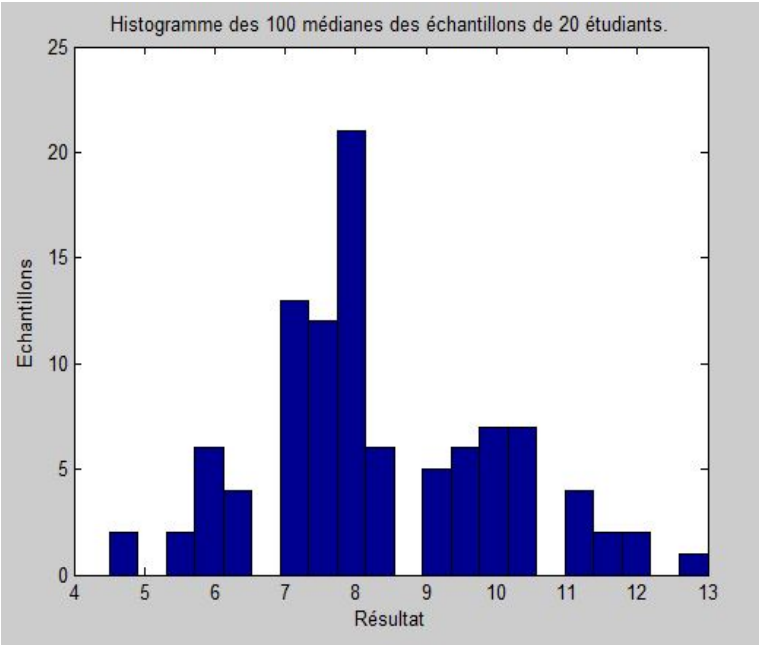

FIGURE 17

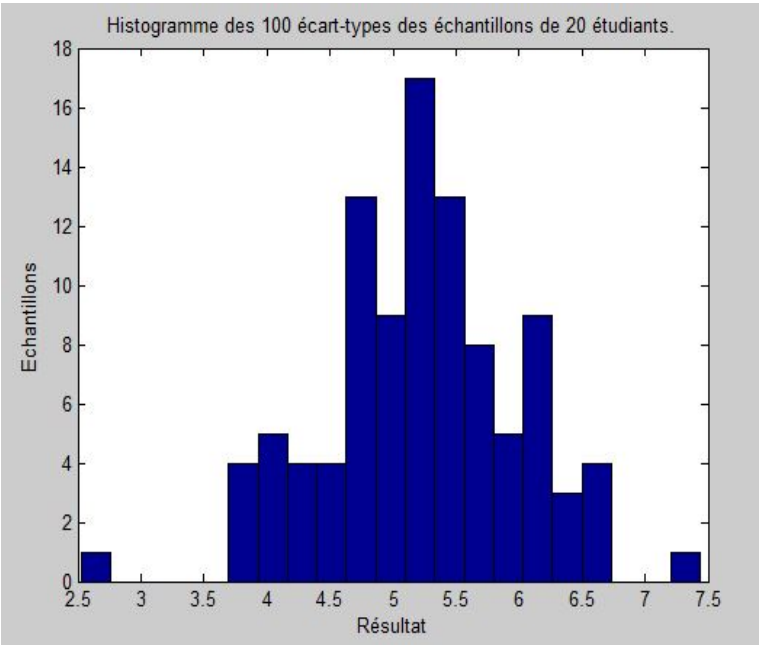

FIGURE 18

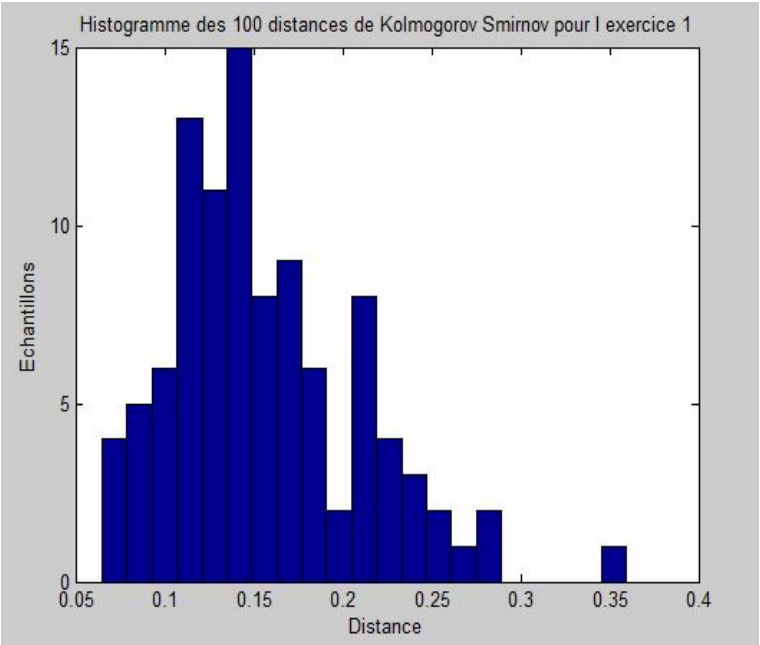

FIGURE 19

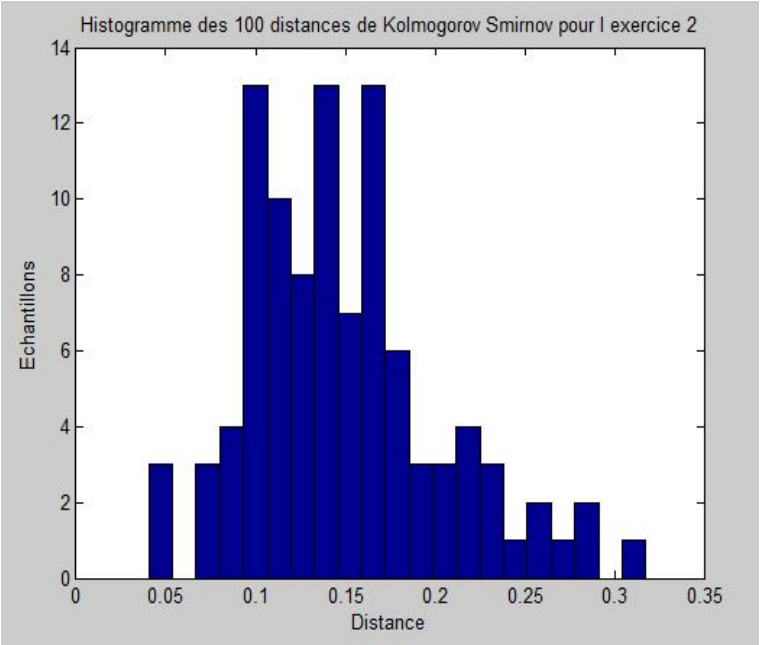

FIGURE 20

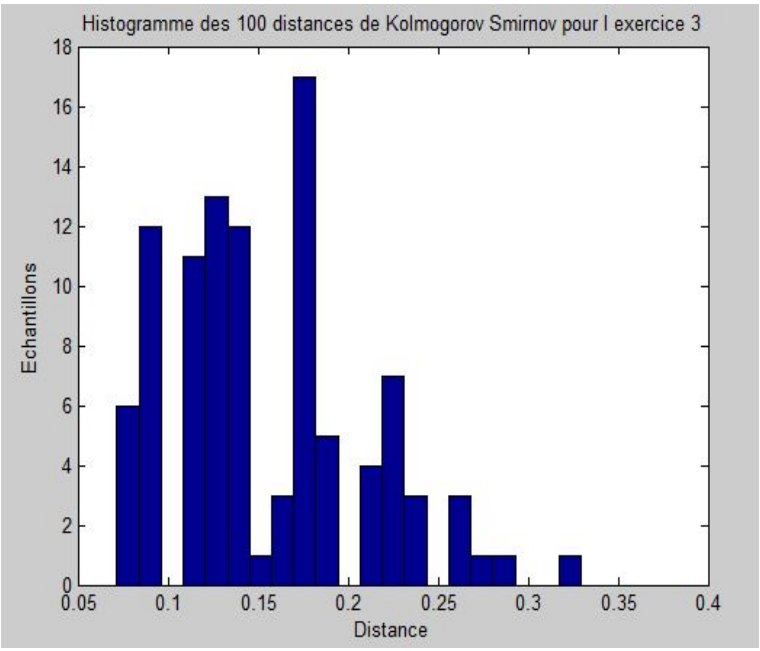

FIGURE 21

A Annexes

A.1 Question 1. (a)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);

x = map(:,5);
binranges = 0:1:20;
bincounts = histc(x,binranges);
bar(bincounts);
xlabel('Resultats');
ylabel('Nombre d etudiants');
title('Histogramme question 1 de theorie');
figure

y = map(:,6);
binranges = 0:1:20;
bincounts = histc(y,binranges);
bar(bincounts);
xlabel('Resultats');
ylabel('Nombre d etudiants');
title('Histogramme question 2 de theorie');
figure

z = map(:,7);
binranges = 0:1:20;
bincounts = histc(z,binranges);
bar(bincounts);
xlabel('Resultats');
ylabel('Nombre d etudiants');
```

```
title('Histogramme question 3 de theorie');
```

A.2 Question 1. (b)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
clear mean;
X = mean(map,1);
mean = [X(8), X(9), X(10)]
clear median;
median = median(map);
median = [median(8), median(9), median(10)]
M = double(map);
clear mode;
M = mode(M);
mode = [M(8), M(9), M(10)]
mat_var = double(map);
clear var;
var = var(mat_var);
st_deviation = [sqrt(var(8)), sqrt(var(9)), sqrt(var(10))]

cdfplot(map(:,8));
title('Polygone des fréquences cumulées des résultats de l exercice 1');
xlabel('Résultat');
ylabel('Fréquences cumulées');
axis([0 22 0 1]);
figure;
cdfplot(map(:,9));
title('Polygone des fréquences cumulées des résultats de l exercice 2');
xlabel('Résultat');
ylabel('Fréquences cumulées');
axis([0 22 0 1]);
figure;
cdfplot(map(:,10));
title('Polygone des fréquences cumulées des résultats de l exercice 3');
xlabel('Résultat');
ylabel('Fréquences cumulées');
axis([0 22 0 1]);
```

A.3 Question 1. (c)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
map = double(map);

boxplot(map(:,1));
title('Boite à moustaches relatives aux résultats du projet 1')
figure
boxplot(map(:,2))
title('Boite à moustaches relatives aux résultats du projet 2')
figure
boxplot(map(:,3))
title('Boite à moustaches relatives aux résultats du projet 3')
figure
boxplot(map(:,4))
```

```
title('Boite à moustaches relatives aux résultats de la question d examen sur le projet 3')

quartileProj1 = quantile(map(:,1),[.25 .50 .75])
quartileProj2 = quantile(map(:,2),[.25 .50 .75])
quartileProj3 = quantile(map(:,3),[.25 .50 .75])
quartileProj3bis = quantile(map(:,4),[.25 .50 .75])
```

A.4 Question 1. (d)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
mat_theory = [map(:,5) map(:,6) map(:,7)];
mat_exercise = [map(:,8) map(:,9) map(:,10)];

clear mean;
mean_theory = mean(mat_theory,2);
mean_exercise = mean(mat_exercise,2);

cdfplot(mean_theory);
title('Polygone des frequences cumulees de la moyenne des questions de theorie des etudiants');
ylabel('Frequences cumulees');
xlabel('Resultat');
figure

cdfplot(mean_exercise);
title('Polygone des frequences cumulees de la moyenne des questions d exercices des etudiants');
ylabel('Frequences cumulees');
xlabel('Resultat');
```

A.5 Question 1. (e)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
map = double(map);
scatter(map(:,3), map(:,4))
xlabel('Rapport projet 3');
ylabel('Question projet 3');
title('Scatterplot');
coeff_corr = corrcoef(map(:,3), map(:,4))
```

A.6 Question 2. (a)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
map = double(map);

vector_random = randsample(148,20,true);

proj_one = [];
proj_two = [];
proj_three = [];
proj_four = [];
th_one = [];
th_two = [];
```

```
th_three = [];
ex_one = [];
ex_two = [];
ex_three = [];

for i = 1:20
    proj_one = [proj_one map(vector_random(i), 1)];
    proj_two = [proj_two map(vector_random(i), 2)];
    proj_three = [proj_three map(vector_random(i),3)];
    proj_four = [proj_four map(vector_random(i),4)];
    th_one = [th_one map(vector_random(i), 5)];
    th_two = [th_two map(vector_random(i), 6)];
    th_three = [th_three map(vector_random(i),7)];
    ex_one = [ex_one map(vector_random(i),8)];
    ex_two = [ex_two map(vector_random(i),9)];
    ex_three = [ex_three map(vector_random(i), 10)];

end
proj_one = proj_one';
proj_two = proj_two';
proj_three = proj_three';
proj_four = proj_four';
th_one = th_one';
th_two = th_two';
th_three = th_three';
ex_one = ex_one';
ex_two = ex_two';
ex_three = ex_three';

mat = [proj_one proj_two proj_three proj_four th_one th_two th_three ex_one ex_two ex_three];
%%%% i %%%%
clear mean;
X = mean(mat,1);
mean = [X(8), X(9), X(10)]
clear median;
median = median(mat);
median = [median(8), median(9), median(10)]
clear mode;
M = mode(mat);
mode = [M(8), M(9), M(10)]
clear var;
var = var(mat);
st_deviation = [sqrt(var(8)), sqrt(var(9)), sqrt(var(10))]

%%%% ii %%%%
boxplot(mat(:,1))
title('Boite à moustaches relatives à la question 1 du projet')
figure
boxplot(mat(:,2))
title('Boite à moustaches relatives à la question 2 du projet')
figure
boxplot(mat(:,3))
title('Boite à moustaches relatives à la question 3 du projet')
figure
```

```
boxplot(mat(:,4))
title('Boite à moustaches relatives  aux résultats de l examen du projet')
%%%% iii %%%%

mat_theory = [mat(:,5) mat(:,6) mat(:,7)];
mat_theory_tot = [map(:,5) map(:,6) map(:,7)];
clear mean;
mean_theory = mean(mat_theory,2);
clear mean;
mean_theory_tot = mean(mat_theory_tot, 2);
binranges = 0:1:20;
bincounts = histc(mean_theory,binranges);

for i = 1:21
    freq(i) = bincounts(i) / sum(bincounts);
end

for i = 2:21
    freq(i) = freq(i-1)+freq(i);
end
figure;
hold on;
cdfplot(mean_theory);
cdfplot(mean_theory_tot);


title('Polygones des frequences cumulees de la moyenne des questions de theorie de l echantillon d e
ylabel('Frequences cumulees');
xlabel('Resultat');


mat_theory = [map(:,5) map(:,6) map(:,7)];
clear mean;
mean_theory = mean(mat_theory,2);
binranges = 0:1:20;
bincounts = histc(mean_theory,binranges);

for i = 1:21
    freq2(i) = bincounts(i) / sum(bincounts);
end

for i = 2:21
    freq2(i) = freq2(i-1) + freq2(i);
end

dist_kolmog = max(abs(freq-freq2))
```

A.7 Question 2. (b)

```
map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);
map = double(map);

vector_mean = [];
```

```
vector_median = [];
vector_st_deviation = [];
vector_kolmo1 = [];
vector_kolmo2 = [];
vector_kolmo3 = [];
binranges = 0:1:20;

for i = 1:100
    ex_one = [];
    ex_two = [];
    ex_three = [];
    vector_random = randsample(148,20,true);
    for j = 1:20
        ex_one = [ex_one map(vector_random(j),8)];
        ex_two = [ex_two map(vector_random(j),9)];
        ex_three = [ex_three map(vector_random(j),10)];

    end

    bincounts = histc(ex_one,binranges);
    bincounts2 = histc(ex_two,binranges);
    bincounts3 = histc(ex_three,binranges);
    ex_one = ex_one';
    ex_two = ex_two';
    ex_three = ex_three';


     clear mean;
     X = mean(ex_one);
     vector_mean = [vector_mean X];

     clear median;
     Y = median(ex_one);
     vector_median = [vector_median Y];

     clear var;
     Z = sqrt(var(ex_one));
     vector_st_deviation = [vector_st_deviation Z];

    for k = 1:21
    freq11(k) = bincounts(k) / sum(bincounts);
    freq21(k) = bincounts2(k) / sum(bincounts);
    freq31(k) = bincounts3(k) / sum(bincounts);
    end

    for k  = 2:21
    freq11(k) = freq11(k-1) + freq11(k);
    freq21(k) = freq21(k-1) + freq21(k);
    freq31(k) = freq31(k-1) + freq31(k);
    end

    bincounts = histc(map(:,8),binranges);
    bincounts2 = histc(map(:,9),binranges);
    bincounts3 = histc(map(:,10),binranges);
```

```
    for k = 1:21
    freq12(k) = bincounts(k) / sum(bincounts);
    freq22(k) = bincounts2(k) / sum(bincounts);
    freq32(k) = bincounts3(k) / sum(bincounts);
    end

    for k = 2:21
        freq12(k) = freq12(k-1) + freq12(k);
        freq22(k) = freq22(k-1) + freq22(k);
        freq32(k) = freq32(k-1) + freq32(k);

    end
    dist_kolmo1 = max(abs(freq11-freq12));
    dist_kolmo2 = max(abs(freq21-freq22));
    dist_kolmo3 = max(abs(freq31-freq32));

    vector_kolmo1 = [vector_kolmo1 dist_kolmo1];
    vector_kolmo2 = [vector_kolmo2 dist_kolmo2];
    vector_kolmo3 = [vector_kolmo3 dist_kolmo3];

end
vector_mean = vector_mean';
vector_median = vector_median';
vector_st_deviation = vector_st_deviation';

vector_kolmo1 =  vector_kolmo1';
vector_kolmo2 =  vector_kolmo2';
vector_kolmo3 =  vector_kolmo3';

hist(vector_mean, 21);
title('Histogramme des 100 moyennes des échantillons de 20 étudiants.')
xlabel('Résultat')
ylabel('Echantillons')

clear mean;
mean_i = mean(vector_mean)
clear mean;
mean_ii = mean(vector_median)
clear mean;
maen_iii = mean(vector_st_deviation)

figure;
hist(vector_median,21);
title('Histogramme des 100 médianes des échantillons de 20 étudiants.')
xlabel('Résultat')
ylabel('Echantillons')
figure;

hist(vector_st_deviation, 21);
title('Histogramme des 100 écart-types des échantillons de 20 étudiants.')
xlabel('Résultat')
ylabel('Echantillons')
```

```
figure;
hist(vector_kolmo1,21);
title('Histogramme des 100 distances de Kolmogorov Smirnov pour l exercice 1')
xlabel('Distance')
ylabel('Echantillons')
figure;
hist(vector_kolmo2,21);
title('Histogramme des 100 distances de Kolmogorov Smirnov pour l exercice 2')
xlabel('Distance')
ylabel('Echantillons')
figure;
hist(vector_kolmo3,21);
title('Histogramme des 100 distances de Kolmogorov Smirnov pour l exercice 3')
xlabel('Distance')
ylabel('Echantillons')
```