MATH0487-2 – Éléments de statistique

Prof. L. Wehenkel

Partie 1 du projet personnel

Analyse des résultats de l'examen 2013 de probabilités

François Rigo

s123218

3ème Bachelier Ingénieur Civil

23 octobre 2014

Avant-propos : les résultats du tableau `Proba1ereSession20132014.xls` ont été arrondis par rapport aux vraies cotes, les étudiants n'ayant pas présenté l'examen ne sont pas comptabilisés (cela fausserait les résultats) et certains ont obtenu des cotes supérieures à 20 (bonus).
NB : En ne considérant pas les figures, annexe et tableaux, le rapport compte $4\frac{1}{2}$ pages.

1 Analyse descriptive

1.(a) Histogrammes de théorie (exécuter `Q1a.m`)

Les trois histogrammes se trouvent aux figures 1, 2 et 3. Tout d'abord, la question 1 semble celle qui a été le mieux réussie. En effet, la quasi-totalité des résultats sont supérieurs à 10. Quasi-totalité car il y a une valeur aberrante (0) dont nous reparlerons en 1(c). Il semble que les étudiants ont bien compris cette question. La question 2 présente des résultats acceptables avec un étalement autour de 10-15 et plusieurs gros échecs. Notons qu'à cette question, un étudiant à obtenu une cote de 22/20, due au bonus. Il y a un grand intervalle de valeurs [0,22] et la majorité des résultats se situe dans un intervalle de réussite. Les résultats de la question 3 sont plus "singuliers", il y a très peu d'étudiants l'ayant très bien réussie (mais un étudiant a obtenu 21), le reste des résultats est dispersé dans les échecs et un très grand nombre ont obtenus certaines valeurs particulières (8 et 13). Nous pouvons faire 2 hypothèses concernant la forte différence de résultats avec la théorie 3 : soit la matière concernant la théorie 3 n'a pas été comprise, soit les étudiants n'ont pas eu assez de temps pour cette dernière. La statistique ne nous permet pas de répondre à ces hypothèses, il faudrait avoir accès à d'autres données (ex : temps passé sur chaque questions,...), assez difficile en pratique. En conclusion, la distribution des 3 questions de théorie est assez différente et nous pouvons les classer selon leur difficulté (subjectivement) : 1, 2 et 3. Notons que nous avons utilisé la fonction `hist()` avec un pas de 1 (pour représenter toutes les cotes).

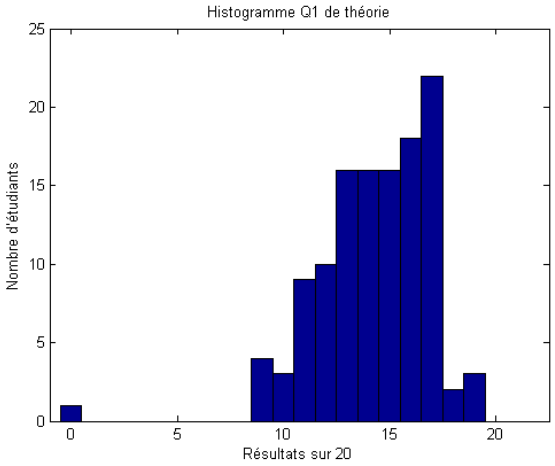


FIGURE 1 – Histogramme théorie 1

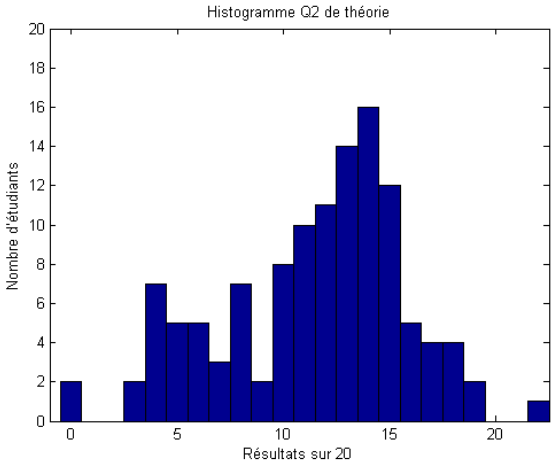


FIGURE 2 – Histogramme théorie 2

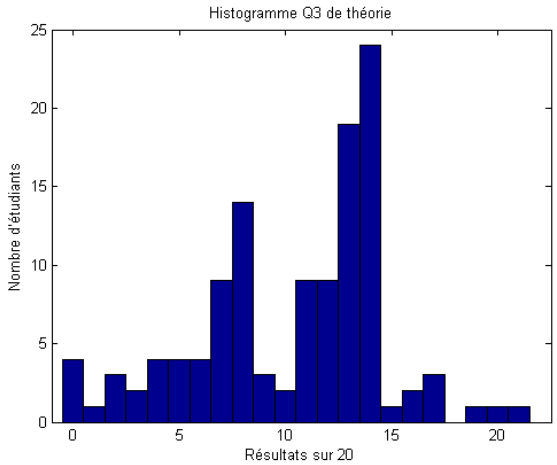


FIGURE 3 – Histogramme théorie 3

1.(b) Moyennes, médianes, modes, écart-types et lois normales des exercices (exécuter `Q1b.m`)

On a [1] :
- Moyenne $= m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ (utilisons la fonction `mean()`)
- Médiane= $F_x^{-1}(0.5)$ càd la valeur pour laquelle 50% des valeurs sont en-dessous (utilisons `median()`)
- Mode= $arg\ max_y\ f_x(y)$ càd la valeur "à la mode", celle qui arrive le plus souvent, avec $f_x(y)$ la fréquence relative (utilisons `mode()`)
- Ecart-type= $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2}$ càd la racine des écarts quadratiques à la moyenne m_x (utilisons `std()`)

Notons qu'ici nous utilisons `std(,1)` pour avoir s_x, l'écart-type corrigé $s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2}$ correspondant à `std(,0)`.
Nous obtenons :

	Moyenne	Médiane	Mode	Ecart-type
Exercice 1	10.8167	11	12	5.6480
Exercice 2	16.8083	18	20	3.7866
Exercice 3	7.7333	7.5	0	5.2816

TABLE 1 – Statistiques des exercices

Nous remarquons que globalement la question 1 a été moyennement réussie, avec une moyenne de presque 11 et un étalement de presque 6. La question 2 a par contre été particulièrement bien réussie. Il y a une moitié des étudiants ayant 18 ou plus, le résultat le plus récurrent étant 20. La question 3 a été particulièrement mal faite, la moitié des étudiants sont en cote d'exclusion et l'étalement autour de la moyenne de $\simeq 8$ est de 5.28. La cote la plus récurrente étant 0 !

NB : il faut bien faire la différence entre la moyenne et la médiane. En effet, la moyenne considère toutes les valeurs et en fait une moyenne arithmétique. Elle n'est donc pas à l'abri de données aberrantes qui fausseraient la valeur réaliste de la *moyenne*. Par contre, la médiane est la valeur pour laquelle 50% des étudiants ont obtenu moins de cette valeur : elle ne fait pas intervenir toutes les valeurs mais regarde la proportion [2]. Elle n'est donc pas sensible aux valeurs aberrantes. Nous pouvons nous en assurer en remplaçant le premier résultat de l'exercice 1 par 10^4. Nous obtenons : $moyenne_{test} = 94$ et $mediane_{test} = 11$. La valeur de la médiane n'a pas été modifiée et reste réaliste. Par contre la moyenne devient totalement absurde. Nous savons donc que : la médiane est parfois un meilleur estimateur de la *moyenne* que la moyenne elle-même. La boîte à moustache correspondante est :

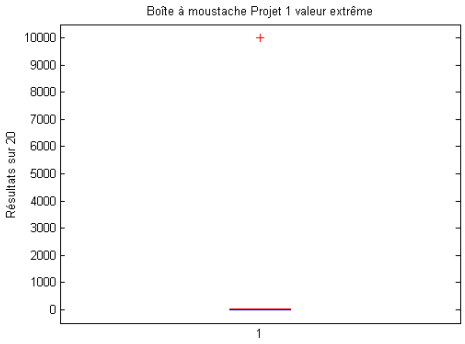


FIGURE 4 – Boxplot avec valeur aberrante extrême

Nous avons ensuite comparé (figures 5 à 7) la distribution en histogramme des exercices à celle de la loi normale qui correspondrait (avec les valeurs des moyennes et d'écart-type du tableau 1) avec la fonction `normpdf()`. L'écart en hauteur entre l'histogramme et la gaussienne vient du fait que la première est discrète et la seconde continue (il faudrait des pas très petits pour passer de discret à continu). Nous remarquons que l'allure correspond mais les exercices ne suivent pas à proprement parlé une loi normale, des valeurs extrêmes apparaissant.

1. D'après le formulaire d'examen du cours de statistique
2. La médiane sera entière ou demi-entière car il y a un nombre pair de valeurs et il faut faire la moyenne des 2 résultats situé à 50% de l'effectif : entière si 2 mêmes valeurs et demi-entière si 2 valeurs adjacentes

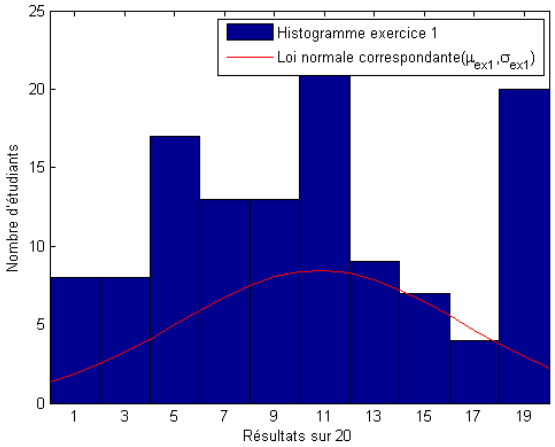


FIGURE 5 – Histogramme exercice 1

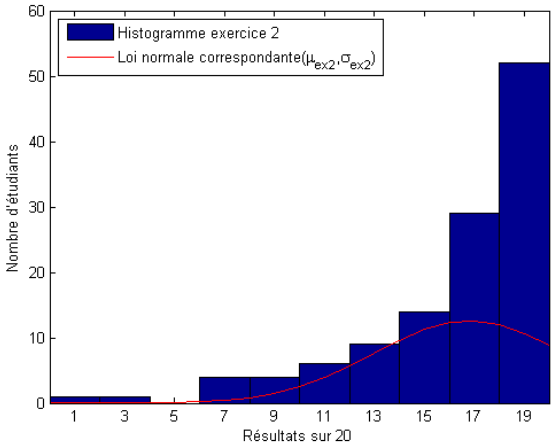


FIGURE 6 – Histogramme théorie 2

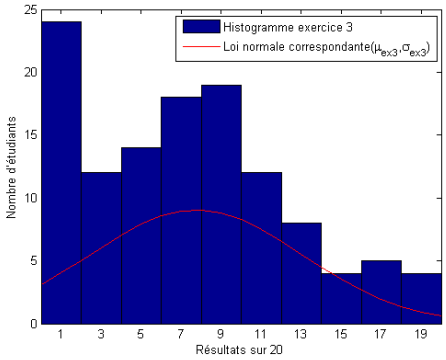


FIGURE 7 – Histogramme théorie 3

Nous avons aussi calculer la proportion d'étudiants "normaux" c'est-à-dire ayant une cote comprise dans $[m_x - s_x, m_x + s_x]$. Nous obtenons :

	Intervalle résultats normaux	Intervalle arrondi	Proportion d'étudiants
Exercice 1	[5.1687 ; 16.4647]	[6 ; 16]	65%
Exercice 2	[13.0217 ; 20.5949]	[14 ; 20]	83.33%
Exercice 3	[2.4517 ; 13.015]	[3 ; 13]	65.83%

TABLE 2 – Caractéristiques "normales" des exercices

Les résultats de l'exercice 1 sont bien répartis et ont l'air de suivre une loi normale. L'exercice 2 a été tellement bien réussi qu'il est n'est pas possible de couvrir tous les résultats normaux (l'intervalle normal excède la borne supérieur de résultats, 20). Encore une fois c'est la question 3 qui a été le plus mal faite : manque de temps ou manque de préparation des étudiants.

1.(c) Boîtes à moustache et quartiles des projets

Nous avons tracé les boîtes à moustache des projets à la figure 8 avec la fonction `boxplot()`, elles ne nous apprennent rien sur la moyenne : la boîte bleue est délimitée par les quartiles Q1 ($F_x^{-1}(0.25)$ et Q3 ($F_x^{-1}(0.75)$) et la médiane Q2 est la ligne rouge ($F_x^{-1}(0.5)$). C'est la question 3 qui apparaît la "plus symétrique" car la médiane est +- au milieu de la boîte bleue et les moustaches sont +- de même longueur. Nous avons alors la table 3. Nous remarquons que Q2=Q3 pour le projet 1. Cela signifie qu'il y a 50% et 75% des étudiants ayant moins de 18. Il y a donc eu un grand nombre d'étudiant ayant obtenu exactement 18 (25%). La moitié des étudiants ont un échec pour la Q projet. Les moustaches sont les valeurs min et max (hors valeurs aberrantes représentées par les croix rouges). Ces valeurs aberrantes sont telles que :

$$x \notin [Q3 - 1.5 * (Q3 - Q1); Q1 + 1.5 * (Q3 - Q1)]$$

Et nous obtenons comme valeurs aberrantes [0,10,12] pour le projet 1 et 0 pour le projet 2. La Q projet est mieux répartie (dans tout [0 ;20])et il n'y a pas de valeurs aberrantes, car sinon elle seraient hors de [0 ;20] ce qui est impossible car il n'y a avait pas de bonus/malus.

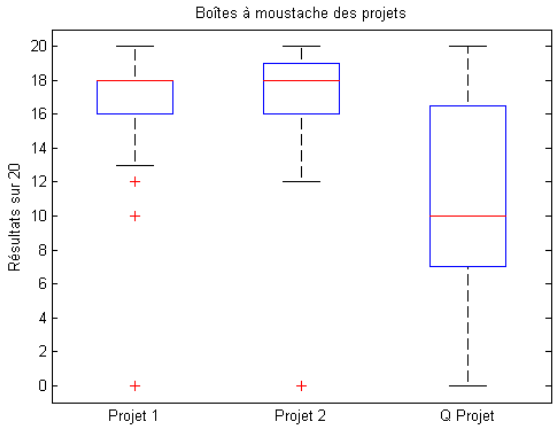


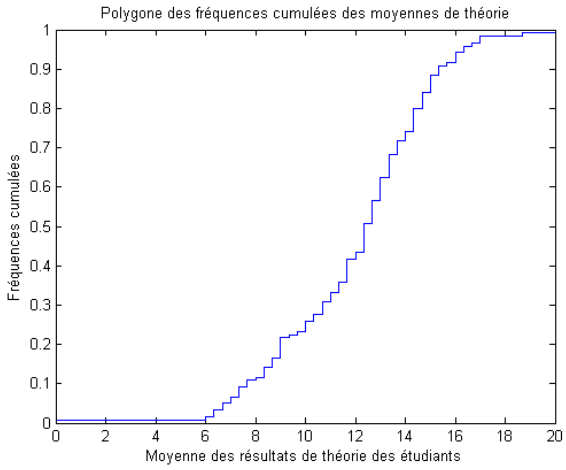
FIGURE 8 – Boxplot des projets

	Q1(0.25)	Q2(médiane)	Q3(0.75)
Projet 1	16	18	18
Projet 2	16	18	19
Q Projet	7	10	16.5

TABLE 3 – Les 3 quartiles des projets

1.(d) Fréquences cumulées de la moyenne de théorie et de la moyenne des exercices pour chaque étudiant

Nous avons réalisé la moyenne des 3 questions de théorie et des 3 exercices pour chaque étudiant (figures 9 et 10). Le polygone des fréquences cumulées en escalier a été généré avec la fonction `cdfplot()`.



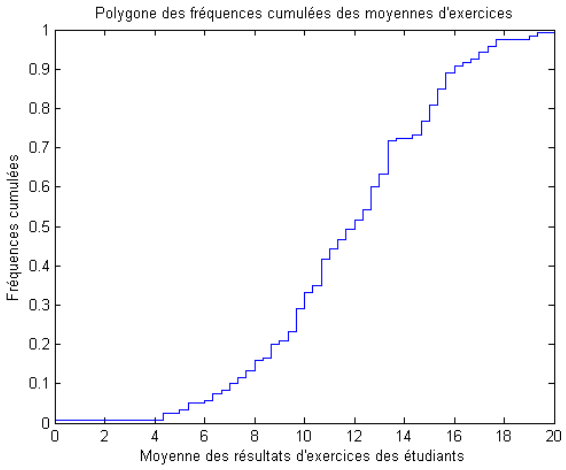


FIGURE 9 – Fréquences cumulées moyenne théorie

FIGURE 10 – Fréquences cumulées moyenne exercice

Nous avons ensuite calculé la proportion d'étudiants ayant une cote entre 12 et 15 (compris). Cela revient à calculer :

$$\hat{F}_x(15) - \hat{F}_x(12)$$

$\hat{F}_x(15)$ correspond à la proportion d'étudiant ayant une cote au plus égale 15 et $\hat{F}_x(12)$ une cote au plus égale à 12. Nous obtenons alors :

$$\%\text{ étudiants théorie } [12\,;15] = 0.45 \;\Rightarrow \text{Nombre} = 120 * 0.45 = 54$$

$$\% \text{ étudiants exercices } [12\,;15] = 0.2917 \;\Rightarrow \text{Nombre} = 120 * 0.2917 = 35$$

Notons que les valeurs de $\hat{F}_x(y)$ ont été obtenues à l'aide de `ecdf()`. Nous pouvons aussi calculer les proportions en utilisant le curseur sur la figure de MATLAB, les valeurs sont en effet identiques. Il faut bien prendre les valeurs du dessus des escaliers et pas du creux des escaliers (générés par `stairs()`.

1.(e) Nuage de points projet 2 - Q projet

Nous avons réalisé le nuage de points (figure 11) des projets 2 et Q avec la fonction `scatterplot()`. Les points représentent des étudiants qui ont obtenu le couple de cotes (Q projet, projet 2).

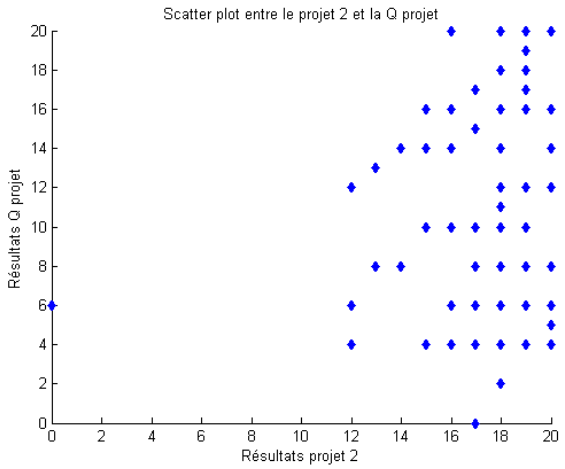


FIGURE 11 – Scatterplot des projets 2 et Q

Une répartition équitable entre les 2 projets aurait donné un scatterplot carré symétrique (mêmes résultats entre 2 et Q). Cependant, on remarque que le projet 2 a été beaucoup mieux fait que la Q projet. Tous les points sont en effet localisés à droite (mise à part un étudiant qui n'a pas rendu son projet 2 et a obtenu 0). Nous remarquons ce déséquilibre des résultats entre 2 et Q en calculant le coefficient de corrélation linéaire (x correspondant à 2 et y à Q) : [3]

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 \sum i = 1^n (y_i - m_y}^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - m_x)}{s_x}\frac{(y_i - m_y)}{s_y}$$

En utilisant la fonction `corrcoef()`, nous trouvons : $r_{2,Q} = \begin{pmatrix} 1 & 0.1407 \\ 0.1407 & 1 \end{pmatrix}$

La diagonale contient des 1 car une valeur est parfaitement corrélée avec elle même. La corrélation entre 2 et Q est donnée par 0.1407 ce qui est assez faible comme nous en avions fait l'hypothèse au scatterplot. L'intensité de liaison entre les projets 2 et Q est donc bien faible. Il semble donc que certains étudiants n'aient pas très bien compris ce qu'ils ont fait au projet 2. Cependant, le projet a été réalisé à domicile, avec tous les outils à disposition et hors conditions d'examen (stress,...). Il n'est donc pas évident de tirer trop de conclusions quand aux raisons de la faible corrélation vu les conditions différentes.

2 Génération d'échantillons i.i.d.

Dans cette partie, nous n'allons plus étudier l'ensemble de la population (120 étudiants) mais tirer des échantillons i.i.d. de 20 étudiants. Pour se faire, nous utilisons la fonction `randsample(120,20)` qui tire 20 nombres au hasard parmi 120. Ces échantillons seront bien indépendants, identiquement distribués (tirage au hasard avec remise).

2.(a) Tirage d'un échantillon i.i.d. de 20 étudiants

Nous tirons un seul échantillon de 20 et utilisons ce même échantillon pour l'ensemble de la section 2.(a). L'échantillon contient les indices suivant d'élèves : [51, 12, 72, 57, 84, 84, 77, 5, 9, 39, 64, 79, 49, 99, 87, 117, 64, 40, 13, 74]

3. $r_{x,y}$ est compris entre -1 et 1 et fait intervenir la covariance des 2 données

2.(a).i Moyennes, médianes et écart-types des exercices

Nous calculons les 3 moyennes, médianes et écart-types des exercices et obtenons (à comparer avec les valeurs de la population)(table 4) :

	Echantillon (n=20)			Population (n=120)		
	Moyenne	Médiane	Ecart-type	Moyenne	Médiane	Ecart-type
Exercice 1	11.2	8.5	6.2738	10.8167	11	5.6480
Exercice 2	15.95	18	4.7167	16.8083	18	3.7866
Exercice 3	9.15	10	4.2223	7.7333	7.5	5.2816

TABLE 4 – Statistiques comparées des exercices

Les valeurs des moyennes, des médianes et des écart-types de l'échantillon ne sont pas très éloignées de celles de la population(en général 1-2 points de différence). Les valeurs ne sont pas identiques (sauf pour la médiane de l'exercice 2, par "chance") mais nous voyons apparaître les mêmes tendances. Notons que l'écart-type varie assez fort d'une exécution à l'autre.

2.(a).ii Boîtes à moustache des projets

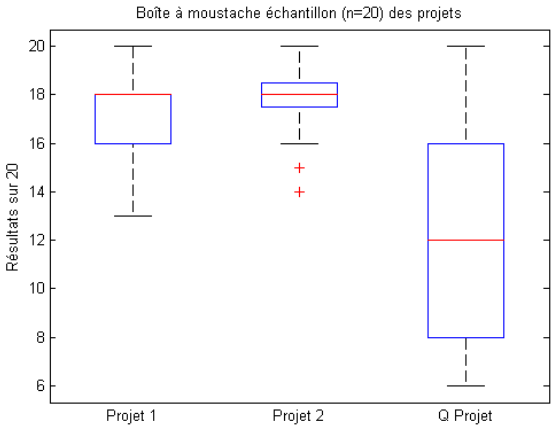


FIGURE 12 – Boxplot projets échantillon

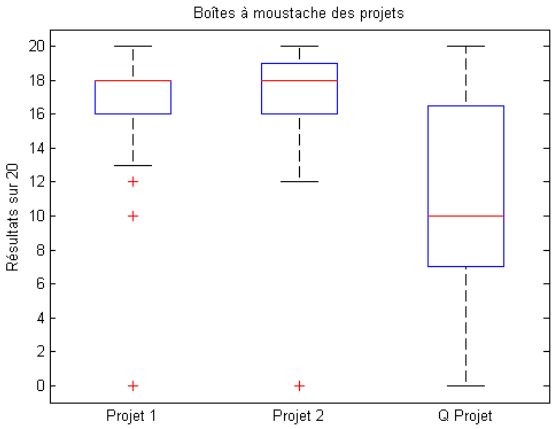


FIGURE 13 – Boxplot projets population

Nous remarquons des tendances très semblables même s'il n'y a pas égalité parfaite. Vu que la taille d'échantillon est bien plus faible que la population (n=20 contre n=120), il y a moins de chance pour tomber sur des valeurs aberrantes (qui sont peu nombreuses dans la population). Nous avons en effet mois de valeurs aberrantes pour l'échantillon que pour la population. Les projet 2 et Q projet de l'échantillon présentent une boîte symétrique (médiane au milieu de Q1 et Q3).

2.(a).iii Fréquences cumulées de la moyenne de théorie et distance de Kolmogorov-Smirnov

Nous avons superposé les polygones de fréquences cumulées de la population et de l'échantillon avec la fonction de répartition normale qui correspondrait à la population ($m_{pop} = 12.0194$,$s_{pop} = 3.0521$). [4]
Vu la similitude entre le polygone de fréquences cumulées des moyennes de théorie de la population et la fonction de répartition de la loi normale correspondante, nous remarquons que les moyennes de théorie de la population suivent une loi normale. Le polygone de fréquences cumulées de l'échantillon a la même allure que celui de la population mais avec certains sauts et décalages(dûs au faible nombre de valeurs, 20, on voit d'ailleurs une vingtaine d'escaliers). Certains étudiants pourraient être représentés plusieurs fois dans un même échantillon de 20 mais avec une assez faible probabilité. Les résultats gardent la même tendance que ceux de la population mise à part les décalages et erreurs intrinsèquement liées au contingentement de l'échantillon. Nous pouvons alors calculer la distance de Kolmogorov-Smirnov entre le polygone de l'échantillon et celui de la population, définie comme la distance maximale entre les 2 courbes. Nous obtenons : [5]

$$D^{KS} = 0.2333$$

4. m_{pop}=moyenne(moyennes de théorie de chaque étudiant) et s_{pop}=écart-type(moyennes de théorie de chaque étudiant)
5. Nous avons réalisé une fonction `dist_kol.m` qui calcule D^{KS} mais il existe une fonction `kstest2()` de MATLAB qui donne la même valeur

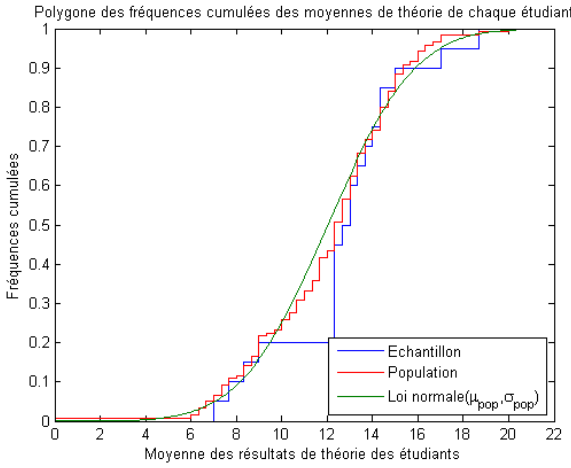


FIGURE 14 – Polygones des fréquences cumulées

Notons que cette valeur est toujours comprise dans [0,1] (normal vu $\hat{F}_x(y) \in [0,1]$) et vu la valeur obtenue, la fonction empirique (échantillon) n'est pas très éloignée de la fonction de référence (population).

2.(b) Tirage de 100 échantillons i.i.d. de 20 étudiants

En procédant de la même manière qu'en 2.(a) nous avons cette fois tiré 100 échantillons i.i.d. de longueur 20 que nous sauvegardons dans une matrice.

2.(b).i Moyenne de l'exercice 1

Nous avons calculé la moyenne de l'exercice 1 de chaque échantillon et sauvegardé les 100 moyennes dans une nouvelle variable. L'histogramme correspondant (par pas de 1) est :

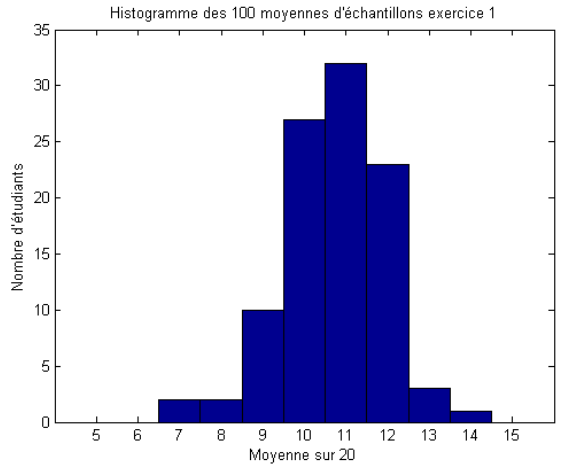


FIGURE 15 – Histogramme des moyennes des 100 échantillons exercice 1

Il apparaît clairement que cette distribution ressemble à une distribution normale, centrée sur une valeur très proche de la moyenne de la population, avec une faible variance donc l'estimation avec 100 échantillons semble acceptable. Nous obtenons une moyenne totale (sur les 100) de **10.7840**, assez proche de la moyenne de référence 10.8167 (1(b))(moyenne inférieure à celle de la population, mais pas à chaque exécution, il arrive que la moyenne des moyennes soit plus élevée que 10.8167). En effet, sur 100 échantillons de 20 parmi une population de 120, la moyenne semble être un bon estimateur de la moyenne car la majorité des étudiants apparaîtront sûrement une fois dans ces 100 échantillons, l'écart avec la vraie valeur vient du fait que certains étudiants seront plus souvent représentés que d'autres. Si à chaque tirage on avait des étudiants différents, il ne suffirait que de 6 échantillons de 20 pour couvrir les 120. Attention dans ce projet nous avons généré des vecteurs ALÉATOIRES, certains étudiants apparaissent donc plusieurs fois. Rigoureusement, il y a $C_{120}^{20} = 2.94 * 10^{22}$ façon différentes de prendre 20 étudiants parmi 120 (chaque étudiant apparaîtra exactement 120 fois chacun). Vu cette valeur extrêmement

élevée on peut supposer que nous aurons une moyenne très proche de celle de la population en utilisant 10^5 échantillons de 20 aléatoirement (les décimales en-dessous de 10^{-2} sont peu intéressantes en pratique).

2.(b).ii Médiane de l'exercice 1

Les médianes des 100 échantillons sont représentés à la figure 16. Ici la distribution normale est moins évidente à percevoir. Il y a un creux là où on s'attendait à un pic (valeur de la moyenne de la population) et l'étalement est moins symétrique. Cependant, d'une exécution à l'autre nous avons quand même observé globalement des distributions plus "normales" (gaussiennes). Nous obtenons **10.3950** comme moyenne des 100 médianes, valeur assez proche de la moyenne 10.8167 mais plus éloignée de la médiane 11 (normal car la médiane et une valeur entière ou demi-entière). Comme expliqué en 1(b), la médiane n'est pas sensible aux valeurs aberrantes contrairement à la moyenne. La médiane est donc considérée comme un bon estimateur de la moyenne. Remarquons que la moyenne des médianes est dans notre cas inférieure à la moyenne des moyennes. C'est assez logique vu que la médiane est entière et a tendance à minimiser la moyenne (en ne tenant pas compte des valeurs aberrantes). Nous pouvons aussi le voir par l'écart-type plus élevé des médianes : $\sigma_{100\ med} = 1.7060$ et $\sigma_{100\ moy} = 1.2367$. La moyenne semble donc être un meilleur estimateur que la médiane.

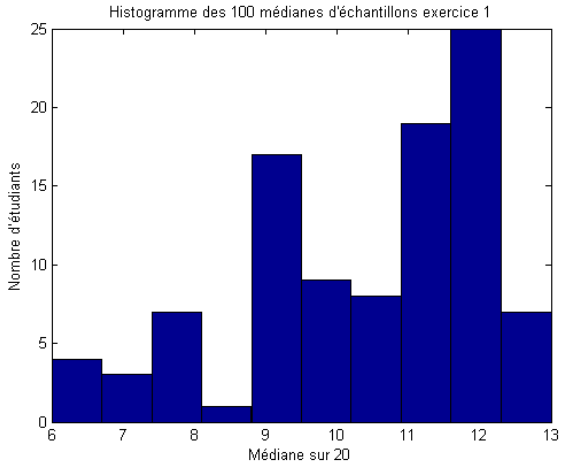


FIGURE 16 – Histogramme des médianes des 100 échantillons exercice 1

2.(b).iii Écart-type de l'exercice 1

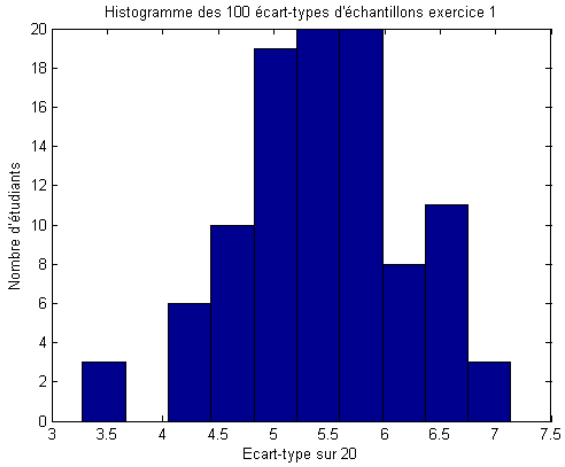


FIGURE 17 – Histogramme des écart-types des 100 échantillons exercice 1

Les écart-types de l'exercice 1 des 100 échantillons se trouvent à la figure 17. La ressemblance avec une loi normale est plus évidente qu'au point précédent, mise à part un trou à gauche. Notons que la présence de trous dans les histogrammes dépend aussi de l'échelle utilisée (vu que nous sommes dans le cas discret, des intervalles trop petits engendreront un grand nombre de trous). Cette gaussienne est centrée sur une valeur proche de

l'écart-type de la population : nous obtenons **5.4353** comme moyenne des écart-types contre 5.6480 pour la population. Notons que nous avons utilisé la forme "s_x" de l'écart-type (c'est-à-dire biaisé). En prenant s_{n-1} nous obtenons[6] **5.5765**, plus proche de 5.6480. C'est assez normal, on préfère toujours utiliser un estimateur non-biaisé car :

$$E\{s_x^2\} = \frac{n-1}{n}\sigma_X^2$$
$$E\{s_{n-1}^2\} = \sigma_X^2$$

Notons que ce sont des variances ici, et vu l'inégalité de Jenssen $E\{\sqrt{s^2}\} \leq \sqrt{E\{s^2\}}$, la statistique s_{n-1} sous-estime σ_X (population) ![7] On peut relier $sn-1$ et s_x par :

$$s_{n-1} = \sqrt{\frac{n}{n-1}}s_x \Longrightarrow 5.5765 = \sqrt{\frac{20}{19}}5.4353$$

2.(b).iv Distance de Kolmogorov-Smirnov de l'exercice 1

2.(b).v Distance de Kolmogorov-Smirnov des exercices 2 et 3

Les histogrammes des DKS des 100 échantillons des exercices 1 à 3 se trouvent aux figures 18 à 20.

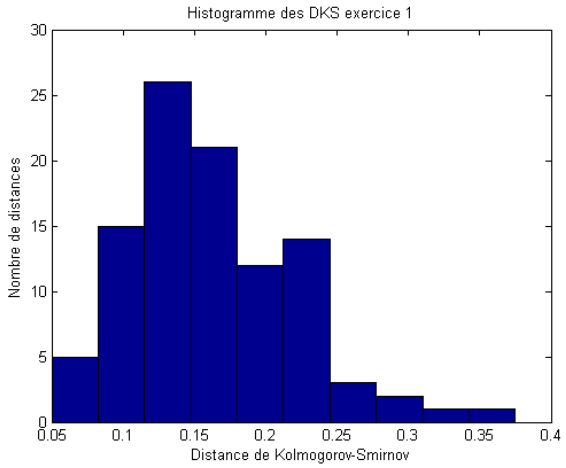


FIGURE 18 – Histogramme des DKS des 100 échantillons exercice 1

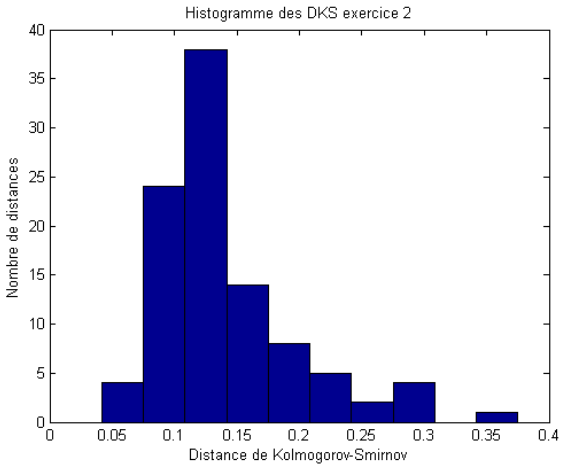


FIGURE 19 – Histogramme des DKS des 100 échantillons exercice 2

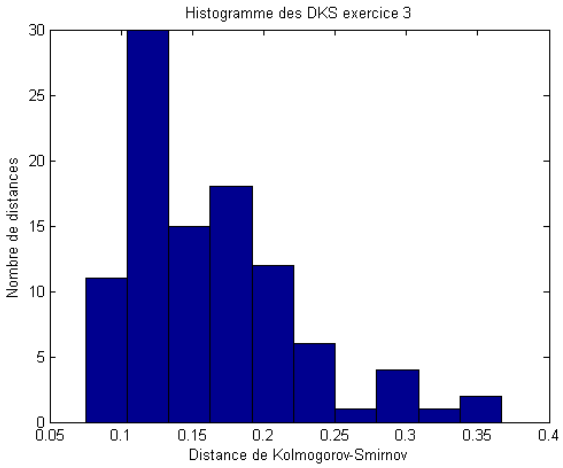


FIGURE 20 – Histogramme des DKS des 100 échantillons exercice 3

Ces distributions sont assez similaires, ce qui montre bien que les D^{KS} (distribution de la distance) sont indépendantes de la forme des F_x (polygones de fréquences cumulées différents pour les 3 exercices). En observant

6. calculée avec `std(,0)`
7. *Éléments de statistique*, L. Wehenkel

la moyenne des 3 D^{KS}, on observe qu'elle est la moins élevée pour l'ex 2 (0.1388) par rapport aux autres (0.1605 pour l'ex 1 et 0.1636 pour l'ex 3). Nous constatons donc que l'exercice 2 a été mieux approché que les 2 autres (dans le cas de nos 100 échantillons), en effet ces moyennes varient à chaque exécution mais restent dans [0.1 ;0.2]). Le théorème de Glivenko-Cantelli nous dit que cette distance tend vers 0 pour un échantillon infini. [8]

Les 3 distributions semblent décrire des lois normales décalées vers la gauche (avec un pic pour des valeurs de D^{KS} de 0.1-0.2 et quelques sauts irrégularités dues aux échantillons). **Cependant, la distribution normale n'est pas la plus adéquate pour décrire nos 3 histogrammes !** En effet, ,il semble plus judicieux de parler de loi Bêta (Be(n,p)) dont la distribution est de la forme : [9]

$$f_x(x) = \frac{(n+p-1)!}{(n-1)!(p-1)!} x^{n-1}(1-x)^{p-1}$$

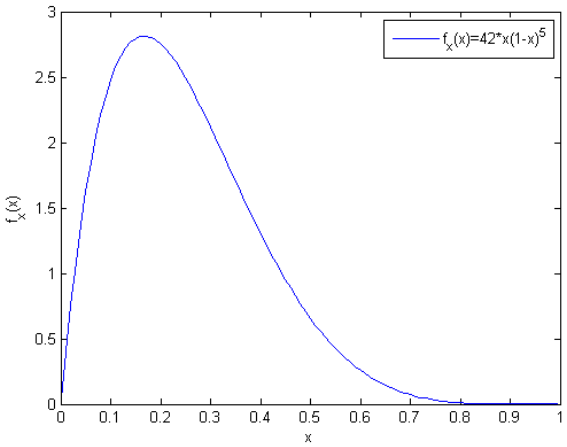


FIGURE 21 – Loi Be(2,6)

Cette loi est continue, ce qui explique la hauteur inférieure aux histogrammes (en effet la hauteur de chaque pics des histogrammes diminuent quand les intervalles diminuent). Cette loi est toujours comprise dans [0 ;1], comme D^{KS} d'ailleurs. Nous l'avons abordée au cours théorique dans le cadre de la démarche bayesienne destinée à estimer la probabilité d'avoir une variable aléatoire θ sachant un certain échantillon D_n. Les 3 histogrammes décrivent en fait le nombre d'échantillons ayant une certaine distance D^{KS} avec la population, sachant un certain échantillon. La démarche réalisée au point 2(b) correspond bien à la volonté d'estimer le mieux possible les statistiques de l'ensemble de la population (n=120) sur base d'un ensemble de 100 estimateurs (n=20) de ses statistiques.

3 Annexe

projet1.m

```
% éRsolution de l'ensemble du projet 1
clear all;
format long;
Q1a;
Q1b;
Q1c;
Q1d;
Q1e;
Q2a;
Q2b;
```

Q1a.m

8. *Éléments de statistique*, chap 3, L. Wehenkel
9. *Éléments de statistique*, chap 4, L. Wehenkel

```
pts=xlsread('Proba1ereSession20132014.xls');
x=0:max(pts(:));
figure;
hist(pts(:,4),x)
axis([-1 max(pts(:))+0.5 0 25]);
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme Q1 de éthorie');
figure;
hist(pts(:,5),x)
axis([-1 max(pts(:))+0.5 0 20]);
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme Q2 de éthorie');
figure;
hist(pts(:,6),x)
axis([-1 max(pts(:))+0.5 0 25]);
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme Q3 de éthorie');
```

Q1b.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q1b---------');
%Moyenne, émdiane, mode et écart-type des Ex1,2,3
moy_Ex=mean(pts(:,7:9))
med_Ex=median(pts(:,7:9))
mod_Ex=mode(pts(:,7:9))
std_Ex=std(pts(:,7:9),1)
% Remarque moyenne!=émdiane
pts_test=pts;
pts_test(1,7)=10000;
figure;
boxplot(pts_test(:,7));
ylabel('éRsultats sur 20')
title('îBote à moustache Projet 1 valeur êextrme')
moy_test=mean(pts_test(:,7))
med_test=median(pts_test(:,7))
%Superposition des histogrammes avec les lois normales correspondantes
x=0:20;
Gauss_E1=120*normpdf(x,moy_Ex(1),std_Ex(1));
figure;
hist(pts(:,7));
hold on;
plot(x,Gauss_E1,'r');
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
legend('Histogramme exercice 1','Loi normale correspondante(\mu_{ex1},\sigma_{ex1})')
Gauss_E2=120*normpdf(x,moy_Ex(2),std_Ex(2));
figure;
hist(pts(:,8));
hold on;
plot(0:20,Gauss_E2,'r');
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
legend('Histogramme exercice 2','Loi normale correspondante(\mu_{ex2},\sigma_{ex2})'
    ,2)
Gauss_E3=120*normpdf(x,moy_Ex(3),std_Ex(3));
figure;
hist(pts(:,9));
hold on;
plot(0:20,Gauss_E3,'r');
xlabel('éRsultats sur 20');
ylabel('Nombre d''étudiants');
legend('Histogramme exercice 3','Loi normale correspondante(\mu_{ex3},\sigma_{ex3})')
```

```
%éRsultats "normaux" et proportion dé'udiants
interv_norm_1=[moy_Ex(1)-std_Ex(1) moy_Ex(1)+std_Ex(1)];
k=1;
for i=0:20
    if i>=interv_norm_1(1) && i<=interv_norm_1(2)
        res_norm_1(k)=i;
        k=k+1;
    end
end
cnt=0;
for i=1:120
    if pts(i,7)>=interv_norm_1(1) && pts(i,7)<=interv_norm_1(2)
        cnt=cnt+1;
    end
end
prop_et_norm_1=cnt/120
interv_norm_2=[moy_Ex(2)-std_Ex(2) moy_Ex(2)+std_Ex(2)];
k=1;
for i=0:20
    if i>=interv_norm_2(1) && i<=interv_norm_2(2)
        res_norm_2(k)=i;
        k=k+1;
    end
end
cnt=0;
for i=1:120
    if pts(i,8)>=interv_norm_2(1) && pts(i,8)<=interv_norm_2(2)
        cnt=cnt+1;
    end
end
prop_et_norm_2=cnt/120
interv_norm_3=[moy_Ex(3)-std_Ex(3) moy_Ex(3)+std_Ex(3)];
k=1;
for i=0:20
    if i>=interv_norm_3(1) && i<=interv_norm_3(2)
        res_norm_3(k)=i;
        k=k+1;
    end
end
cnt=0;
for i=1:120
    if pts(i,9)>=interv_norm_3(1) && pts(i,9)<=interv_norm_3(2)
        cnt=cnt+1;
    end
end
prop_et_norm_3=cnt/120
```

Q1c.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q1c---------');
boxplot([pts(:,1),pts(:,2),pts(:,3)],{'Projet 1','Projet 2','Q Projet'});
title('îBotes à moustache des projets');
ylabel('éRsultats sur 20')
quartiles_proj12Q=quantile(pts(:,1:3),[0.25 0.5 0.75])
Q=quartiles_proj12Q;
%Remarque valeurs aberrantes
k=1;
for i=1:length(pts)
    for j=1:3
if pts(i,j)<Q(1,j)-1.5*(Q(3,j)-Q(1,j)) || pts(i,j)>Q(3,j)+1.5*(Q(3,j)-Q(1,j))
    val_aberrant(k,1)=i;
    val_aberrant(k,2)=j;
    val_aberrant(k,3)=pts(i,j);
    k=k+1;
end
    end
```

```
end
```

Q1d.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q1d---------');
moy_th=mean(pts(:,4:6)');
moy_ex=mean(pts(:,7:9)');
[f_th,x_th]=ecdf(moy_th);
cdfplot(moy_th);
grid off;
title('Polygone des éfrquences écumules des moyennes de éthorie')
xlabel('Moyenne des érsultats de éthorie des étudiants')
ylabel('éFrquences écumules')
figure;
cdfplot(moy_ex);
grid off;
title('Polygone des éfrquences écumules des moyennes d''exercices')
xlabel('Moyenne des érsultats d''exercices des étudiants')
ylabel('éFrquences écumules')
prop_th_12_15=f_th(find(x_th==15))-f_th(find(x_th==12))
[f_ex,x_ex]=ecdf(moy_ex);
prop_ex_12_15=f_ex(find(x_ex==15))-f_ex(find(x_ex==12))
```

Q1e.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q1e---------');
scatter(pts(:,2),pts(:,3),'filled','d');
xlabel('éRsultats projet 2')
ylabel('éRsultats Q projet')
title('Scatter plot entre le projet 2 et la Q projet')
coef_corr=corrcoef(pts(:,2),pts(:,3))
```

Q2a.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q2a---------');
ech1=randsample(120,20,true);
%(i) moyenne, émdiane et écart-type des exercices, sur lé'chantillon
moy_Ex_ech1=mean(pts(ech1,7:9))
med_Ex_ech1=median(pts(ech1,7:9))
std_Ex_ech1=std(pts(ech1,7:9),1)
%(ii) îbotes à moustache des projets, sur lé'chantillon
figure;
boxplot([pts(ech1,1),pts(ech1,2),pts(ech1,3)],{'Projet 1','Projet 2','Q Projet'});
title('îBote à moustache échantillon (n=20) des projets');
ylabel('éRsultats sur 20')
quartiles_ech_proj12Q=quantile(pts(ech1,1:3),[0.25 0.5 0.75]);
%(iii) distance de Kolmogorov-Smirnov entre la fonction de érpartition de
%la population et celle de lé'chantillon
moy_th_ech=mean(pts(ech1,4:6)');
moy_th_pop=mean(pts(:,4:6)');
[dist_kolmogorov1_moy_th,x,f_th_tot,f_th_ech]=dist_kol(moy_th_pop,moy_th_ech);
dist_kolmogorov1_moy_th
% éVrification de la distance de kolmogorov avec la fonction kstest2
[h,p,dist_kolmogorov1_moy_th_kstest2]=kstest2(moy_th_pop,moy_th_ech);
dist_kolmogorov1_moy_th_kstest2
figure;
stairs(x,f_th_ech);
hold all;
stairs(x,f_th_tot,'r');
axis([0 22 0 1]);
plot(0:0.1:22,normcdf(0:0.1:22,mean(moy_th_pop), (std(moy_th_pop,1))))
title('Polygone des éfrquences écumules des moyennes de éthorie de chaque étudiant')
xlabel('Moyenne des érsultats de éthorie des étudiants')
ylabel('éFrquences écumules')
```

```
legend('Echantillon','Population','Loi normale(\mu_{pop},\sigma_{pop})',4);
```

Q2b.m

```
pts=xlsread('Proba1ereSession20132014.xls');
disp('---------Q2b---------');
% Pour l'ex 1: (i) moyennes 100 éch, (ii) émdianes 100 éch,
% (iii) écart-types 100 éch
for i=1:100
    ech100(:,i)=randsample(120,20,true);
    moy_Ex1_ech100(i)=mean(pts(ech100(:,i),7));
    med_Ex1_ech100(i)=median(pts(ech100(:,i),7));
    std_Ex1_ech100(i)=std(pts(ech100(:,i),7),1);
    std_Ex1_ech100_corr(i)=std(pts(ech100(:,i),7),0);
end
figure;
hist(moy_Ex1_ech100,5:15)
xlabel('Moyenne sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme des 100 moyennes d''échantillons exercice 1');
moy_moy_Ex1_ech100=mean(moy_Ex1_ech100)
moy_Ex1=mean(pts(:,7))
figure;
hist(med_Ex1_ech100)
xlabel('éMdiane sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme des 100 émdianes d''échantillons exercice 1');
moy_med_Ex1_ech100=mean(med_Ex1_ech100)
med_Ex1=median(pts(:,7))
std_med_Ex1_ech100=std(med_Ex1_ech100);
std_moy_Ex1_ech100=std(moy_Ex1_ech100);
figure;
hist(std_Ex1_ech100)
xlabel('Ecart-type sur 20');
ylabel('Nombre d''étudiants');
title('Histogramme des 100 écart-types d''échantillons exercice 1');
moy_std_Ex1_ech100=mean(std_Ex1_ech100)
moy_std_Ex1_ech100_corr=mean(std_Ex1_ech100_corr)
std_Ex1=std(pts(:,7),1)
%Pour les exercices 1 (iv) à 3(v): DKS pour les 100 échantillons
for i=1:100
    dist_kolmogorov100_Ex1(i)=dist_kol(pts(:,7),pts(ech100(:,i),7));
    dist_kolmogorov100_Ex2(i)=dist_kol(pts(:,8),pts(ech100(:,i),8));
    dist_kolmogorov100_Ex3(i)=dist_kol(pts(:,9),pts(ech100(:,i),9));
end
figure;
hist(dist_kolmogorov100_Ex1)
xlabel('Distance de Kolmogorov-Smirnov');
ylabel('Nombre de distances');
title('Histogramme des DKS exercice 1');
figure;
hist(dist_kolmogorov100_Ex2)
xlabel('Distance de Kolmogorov-Smirnov');
ylabel('Nombre de distances');
title('Histogramme des DKS exercice 2');
figure;
hist(dist_kolmogorov100_Ex3)
xlabel('Distance de Kolmogorov-Smirnov');
ylabel('Nombre de distances');
title('Histogramme des DKS exercice 3');
moy_DKS_Ex=[mean(dist_kolmogorov100_Ex1) mean(dist_kolmogorov100_Ex2) mean(
    dist_kolmogorov100_Ex3)];
%Loi êBta
xx =0:0.01:1;
plot(xx,42*xx.*(1-xx).^5)
legend('f_x=42*x(1-x)^5')
```

```
ylabel('f_x(x)')
legend('f_x(x)=42*x(1-x)^5')
```

dist_kol.m

```
function [dist_kolmogorov,x,f_pop,f_ech]=dist_kol(pop,ech)
[f_p,x_pop]=ecdf(pop);
[f_e,x_ech]=ecdf(ech);
clear x;
x=sort([x_pop' x_ech']);
f_ech=cumsum(histc(ech,x)/sum(histc(ech,x)));
f_pop=cumsum(histc(pop,x)/sum(histc(pop,x)));
dist_kolmogorov=max(abs(f_pop-f_ech));
end
```