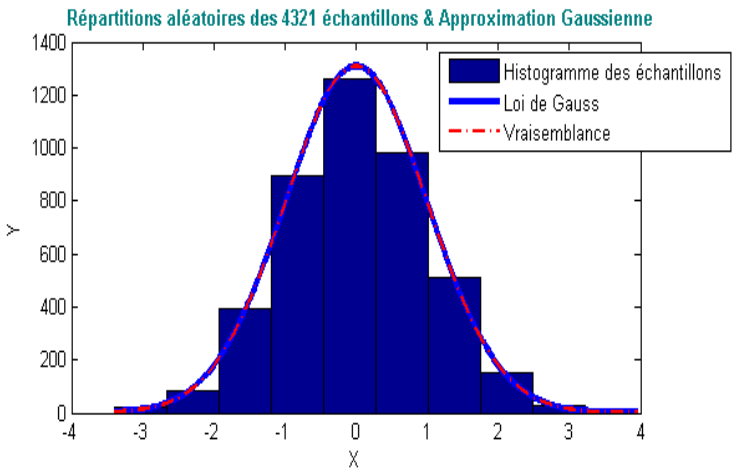
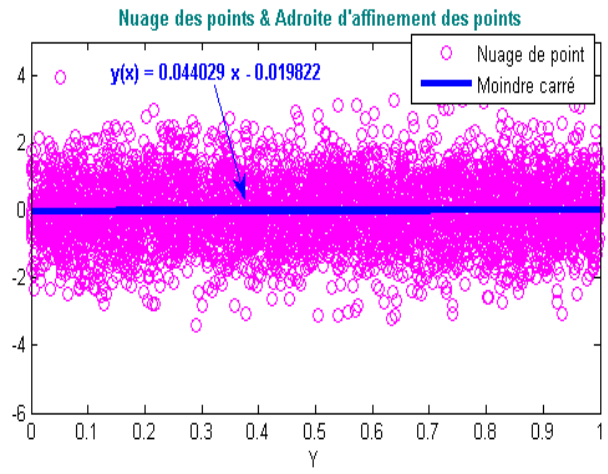
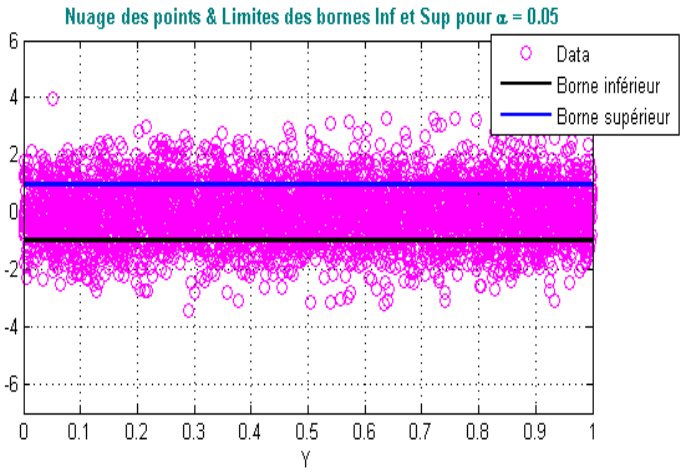
U.E : **Détection, Estimation & Information pour les Data Sciences**

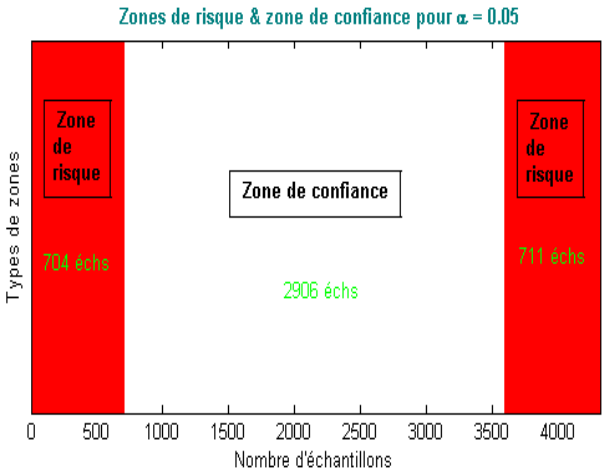
TP $n^0 1$

Maximum de Vraisemblance, Régression Linéaire, Zones de Confiance et Traitement de Data Massives Collectées









Objectifs du TP :

Le présent TP intitulé "`Maximum de Vraisemblance, Régression Linéaire, Zones de Confiance et Traitement de Data Massives Collectées`" est adossé à une partie du cours/ TDs de l'U.E. '***Détection, Estimation & Information pour les Data Sciences***'. Il est considéré comme un complément pour mieux assimiler et consolider les connaissances acquises. Il convoite à :

- comprendre comment générer aléatoirement des data massives, en particulier selon une distribution probabiliste (loi normale $\mathcal{N}(\mu, \sigma)$ dans le cas de ce TP) ;
- visualiser l'histogramme des data générées, puis le comparer par rapport à la loi normale en fonction du nombre de data étudiées ;
- appliquer la technique du maximum de vraisemblance pour représenter les données massives ;
- utiliser la technique du moindre carré pour approximer par une régression linéaire, les data générées ;
- représenter pour une probabilité astreinte, les bornes inférieur et supérieur relatives aux nuages des données massives générées en délimitant les types de zones concernées.
- obtenir les statistiques sur les data massives générées selon les techniques d'estimation les plus utilisées ;
- Manipuler, analyser et traiter un exemple de traitement de données massives collectées, stockées, provenant d'un sondage effectué auprès d'étudiants au sein d'un établissement universitaire.

À l'issue de ce TP, l'étudiant sera capable :

- d'estimer et d'interpréter les data massives générées aléatoirement ;
- de délimiter les zones de confiance ainsi que celles de risque selon une probabilité d'assurance prescrite ;
- de se familiariser et d'appréhender les techniques d'estimation les plus utilisées et de maîtriser leurs applications ;
- de se familiariser avec le traitement de méga données recueillies lors d'études, d'expériances, d'enquêtes, de sondages, ...

Concernant le compte-rendu du TP :

Ce TP doit faire l'objet d'un compte-rendu écrit qui devra être envoyé par e-mail à l'adresse électronique "*k.ghoumid@ump.ac.ma*". Il doit être mis sous la forme d'un seul fichier au format pdf, accompagné d'autres fichiers séparés au format Matlab (.m) des codes générés.
Les fichiers demandés doivent respecter les points prescrits énumérés ci-dessous.
Le compte-rendu doit être transmis au plus tard dans un délai d'une semaine à compter de cette séance.
Le fichier au format pdf du compte-rendu doit contenir :

- des réponses expressives et des explications circonstanciées à chacune des questions sollicitées.
- des figures qui doivent avoir des légendes et qui doivent être bien expliquées et commentées.
- en annexe les codes générés.

Thèmes rencontrés dans ce TP :

Données massives, Loi normale, Théorème central limite, Régression linéaire, Estimation ponctuelle, Estimation par intervalles, Moindre carré, Maximum de vraisemblance, Intervalles de confiance, Méga Data collectées, ...

TRAVAIL DEMANDÉ

1. Génération aléatoires de données massives.

1.1. Préambule

La naissance du **Big Data** est liée aux progrès des capacités des systèmes de stockage, de fouille et d'analyse de l'information numérique, c'est le reflet d'un changement plus profond qui concerne le passage d'une ère industrielle vers une ère numérique. Les Méga-Données sont un domaine qui traite des moyens d'analyser, d'extraire ou de traiter des ensembles de données trop volumineux ou trop complexes, pour être traités par des logiciels d'application de traitement de données traditionnels. Cette explosion quantitative, souvent redondante, des Big-Data, provenant d'une large variété de sources numériques, permet fréquemment l'utilisation des approches statistiques (parfois nouvelles), de l'analyse prédictive, ou de certaines autres méthodes avancées de données, pour les analyser. Ces méga-données étant très diverses, de par leur nature et/ou leur niveau de structuration, leur analyse sera donc différente. Il s'agit en fait de construire des modèles destinés à mieux comprendre des phénomènes et des comportements insaisissables jusqu'alors.
Dans ce TP, on va commencer par la génération d'un volume colossal de données numériques, puis par sa mise à disposition, dans une optique de faciliter sa compréhension, son traitement et son interprétation par la suite afin d'améliorer les prises de décisions. La discussion des problèmes d'estimation et des questions de test d'hypothèses viendra également après, et ceci en mettant l'accent sur les méthodes les plus appropriées dans ce type de traitement. Puis on donnera deux exemples de traitements de données massives collectées, relatives à deux applications différentes.

1.2. Génération des données

À cet effet, on dispose dans ce présent TP de données massives que l'on désire y tirer des conclusions. Pour parvenir à tel résultat, on fait l'hypothèse d'existence d'une loi de probabilité sous-jacente, qui considère ces data comme des variables aléatoires indépendantes ayant cette loi. Ensuite on cherche à l'aide des approches généralement probabilistes, à évaluer, approximativement des paramètres et des quantités, a priori inconnus, et qui apparaît préalablement difficile de les évaluer directement.
On commence dans un premier temps par la génération des données massives à traiter. Pour cela, on va utiliser la méthode de Box-Muller. Cette dernière consiste à générer des paires de nombres aléatoires à distribution normale centrée réduite, à partir d'une source de nombre aléatoires de loi uniforme.

✓ Introduire le nombre n d'échantillons aléatoires à transmettre (de préférence grand) ;

Pour la suite de ce TP, on définit dès à présent :

- Le pourcentage qui minore la probabilité d'inclure la valeur à estimer. Il s'agit en

fait d'encadrer la valeur estimée par une marge d'erreur astreinte selon les exigences des intervalles de confiance.

- La variance des données étudiées.

✓ Générer à l'aide de l'instruction "$rand(n, 1)$" un vecteur X selon la distribution uniforme, puis à l'aide de la fonction "$Box\_Muller(n)$" le vecteur Y des paires de nombres aléatoires à distribution normale centrée réduite, liés à X, en imitant la syntaxe suivante :

```
clear all;
close all;
clc;
n = input(Éntrer le nombre d'échantillons à tester  =  ');
alpha = input(Éntrer la valeur de alpha (comprise entre 0 et 1)  =  ');
var = input(Éntrer la valeur attribuée à la variance (entre 0 et 1)  =  ');
```

Question 1 :

1-1. Écrire les lignes de commandes liées aux vecteurs X et Y.

1-2. Tracer à l'aide de l'instruction "***hist(Y)***", les histogrammes des data massives générées correspondant aux nombres $n = 20, 50, 300, 1000$ et 7000. Conclure.

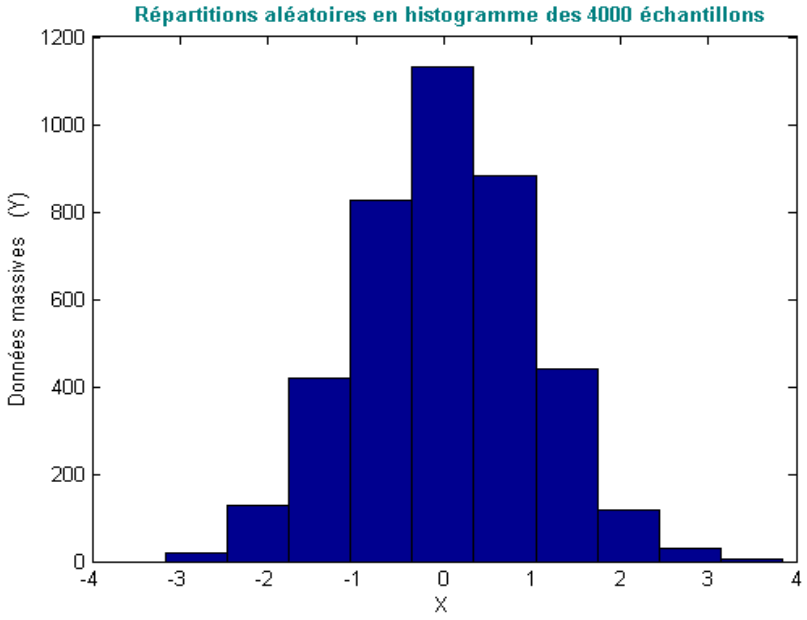


FIGURE 1 – Exemple d'un histogramme des data pour $n = 4000$.

2. Comparaison des data générées avec la distribution Normale $\mathcal{N}(\mu, \sigma)$.

À partir des n (on lui assignera par la suite une grande valeur) échantillons aléatoires générés (data), on cherche à tracer l'allure de la courbe obtenue, puis de la comparer par rapport à l'histogramme de la courbe ci-haut.

✓ En utilisant les commandes "***mean()***" et "***std()***", calculer pour le vecteur Y défini ci-dessus, la moyenne et l'écart-type en mimant les lignes suivantes :

```
mu = mean(Y);
sigma = std(Y);
```

✓ Définir le point G représentant le barycentre des data générées avec la syntaxe suivante :

```
Point_G = [mean(X),mean(Y)];
```

Question 2 :

2-1. Utiliser l'instruction "**disp()**" ou "**fprintf()**" pour afficher les valeurs, de la moyenne μ et de l'écart-type σ, calculées ci-dessus.

2-2. Recalculer les valeurs de μ_X, μ_Y, σ_X et σ_Y avec l'instruction "**fitdist(x,'Normal')**".

2-3. Avec les valeurs numériques de μ et σ et en se servant de la fonction "***GaussLoi()***" (une macro) ou bien de la fonction Matlab "***normpdf()***", superposer sur une même figure l'allure de l'histogramme des données et celle de la loi normale, lacées aux n données massives. Interpréter.

3. Estimation par Maximum de Vraisemblance (Maximum Likelihood Estimation).

Une fois les données massives générées par la technique de Box-Muller sont comparées par rapport à la distribution normale $\mathcal{N}(\mu, \sigma)$, elles seront estimées par la suite en utilisant le principe du Maximum de Vraisemblance (MV, en anglais "Maximum Likelihood method").
Cette technique statistique relative au maximum de vraisemblance permettant l'estimation des paramètres d'un modèle de régression, est utilisée pour inférer les paramètres de la loi de probabilité des données générées (vecteur Y dans notre cas) en recherchant les valeurs des paramètres maximisant la fonction de vraisemblance. En d'autres termes, la méthode consiste à trouver la valeur la plus vraisemblable du paramètre global partant d'un échantillon donné.

Question 3 :

3-1. Au moyen de l'instruction "***mle()***" qui renvoie les paramètres d'estimateur recherchés par la méthode de maximum de vraisemblance et en utilisant la commande "***fprintf()***", afficher :

- La valeur de la moyenne μ_{MV} et son intervalle de confiance à 95%.
- La valeur d'écart-type σ_{MV} et son intervalle de confiance à 95%.

Imiter la syntaxe suivante pour rechercher les estimateurs $\widehat{\mu}_{MV}$ et $\widehat{\sigma}_{MV}$:

```
MVS1 = mle(Y,'distribution','normal');
```

3-2. Comparer les valeurs obtenues par l'estimation MV ($\widehat{\mu}_{MV}$ et $\widehat{\sigma}_{MV}$) à celles calculées par les formules mathématiques.

3-3. Superposer sur la même figure l'allure de l'histogramme des data, celle de la loi normale de paramètres issus des formules statistiques et celle de la loi normale de paramètres issus de l'estimation MV, liées aux données massives générées aléatoirement. Interpréter.

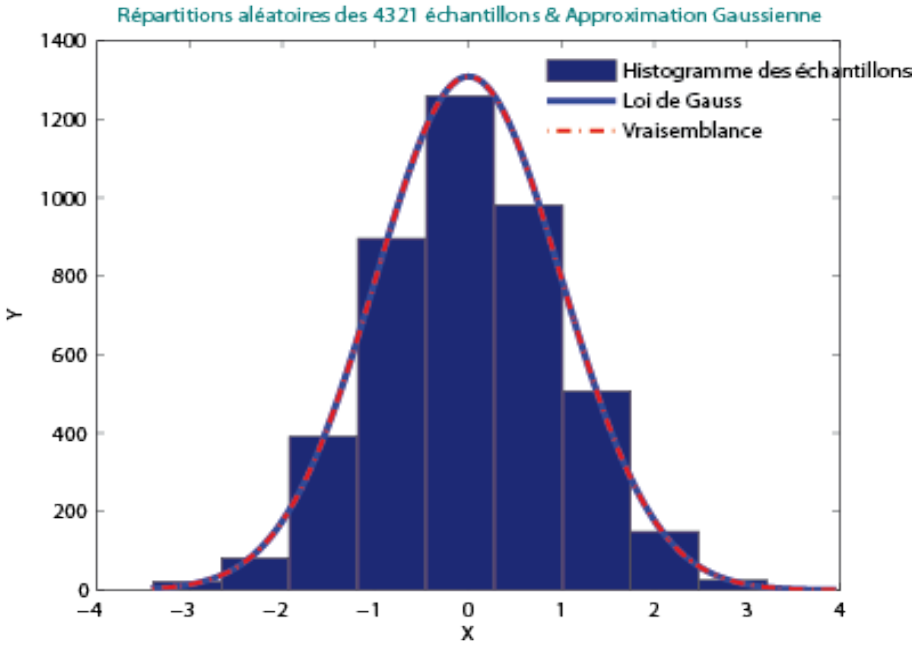


FIGURE 2 – Histogramme des $n = 4321$ data générées, leur approximation selon la loi Normale et leurs estimation par la méthode du Maximum de Vraisemblance.

4. Estimation par Moindre Carré (Least Squares) et Régression linéaire.

Dans ce paragraphe, on cherche à utiliser l'approche de la régression linéaire pour établir une fonction prédictive linéaire dont les paramètres de modèle inconnus sont estimés à partir des data générées.
En d'autres termes, on cherche l'expression de la fonction affine la plus proche conduisant à un ajustement linéaire, permettant d'expliquer le comportement de la variable statistique Y en fonction X.
Il est à rappeler que cette régression linéaire est souvent estimée par la méthode des moindres carrés.

✓ Plagier la ligne suivante pour obtenir l'ajustement linéaire le plus proche associé aux données massives traitées.

```
P = polyfit(X,Y',1) ;
```

Question 4 :

4-1. Donner l'équation de la droite $y(x)$ obtenue par le modèle de la régression linéaire.

4-2. Tracer sur la même figure les nuages des données massives ainsi que l'adroite de l'ajustement linéaire associé.

4-3. Retrouver les estimateurs ponctuels $\widehat{a}_{LS}$ et $\widehat{b}_{LS}$ représentant les coefficients de la droite $y(x)$ en utilisant la fonction "***MoindreCarres()***".

4-4. Vérifier que le point G ($Point\_G(= [mean(X), mean(Y)]$) qui représente le barycentre des data générées, appartient à l'adroite de la régression linéaire.

4-5. Définir puis calculer le coefficient de détermination lié à cet ajustement linéaire. Comment peut-on juger la qualité de cette régression linéaire simple ?

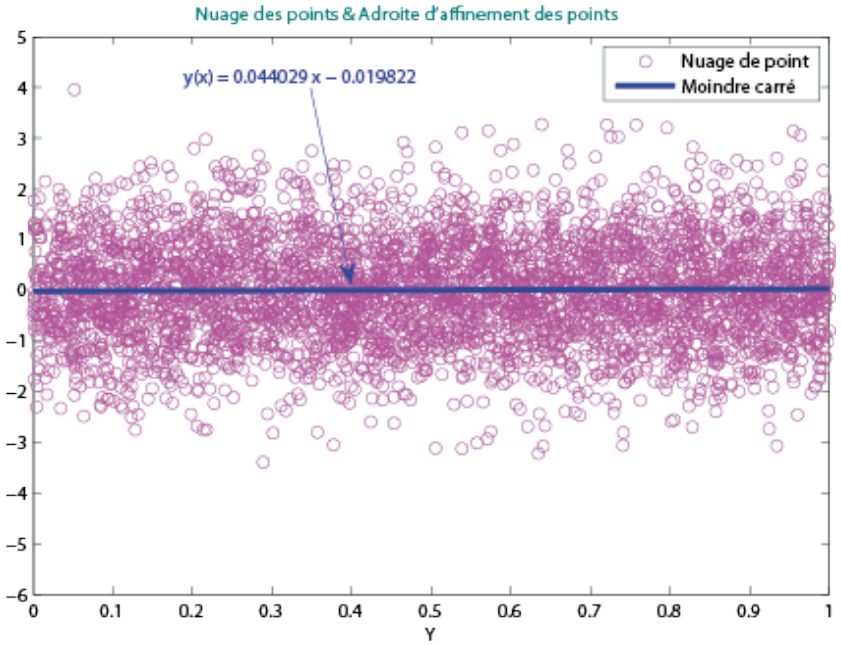


FIGURE 3 – Nuage des données massives et ajustement linéaire par la droite d'équation $y(x)$ obtenue par la méthode des moindres carrés.

5. Estimation par intervalles de confiance à une tolérance astreinte.

Dans ce paragraphe, on cherche à encadrer l'estimateur (non biaisé) de la variance des data générées aléatoirement et ayant la loi de probabilité définie ci-dessus.
Pour évaluer le degré de la confiance relative à cet estimateur du paramètre de la variance, il est judicieux de déterminer un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur du paramètre à estimer. Une telle estimation par intervalle de confiance, permet la précision de l'incertitude sur les estimations, encadre la valeur réelle que l'on cherche à estimer à l'aide de mesures aléatoires et permet aussi de définir une marge d'erreur entre les résultats.
Cet encadrement de la valeur réelle par une estimation qui y précise des incertitudes et y engendre des marges d'erreurs, consiste à identifier deux bornes, inférieure et supérieure, qui dépendent des données massives générées.

✓ Imiter la syntaxe disposée ci-dessous pour encadrer l'estimateur de la variance par un intervalle de confiance, susceptible de contenir la valeur du paramètre estimé avec une probabilité $(1-\alpha)$ fixée a priori.

Question 5 :

5-1. Tracer sur la même figure le nuage des points relatifs aux données massives de la zone de confiance et celui des zones de risque (coefficient α de Student).

5-2. Donner la figure et les statistiques relatives aux data de la zone de confiance ainsi que celles liées aux zones délimitant les bornes inférieur et supérieure.

5-3. Donner les commandes qui permettent les régressions linéaires (ajustements par des droites approximées selon les moindres carrés) des nuages des bornes des zones de risque.

5-4. Superposer sur la même figure le nuage des points relatifs aux données massives et les droites d'ajustement cadrant les bornes inférieur et supérieur.

5-5. Récapituler en affichant à l'aide des instructions "***fprintf()***" ou "***disp()***", les statistiques relatives aux nombres de data de la zone de confiance et des zones de risque.

```
Xe = 0;
Ye = 0;
for   i = 1 : n
      Xe = Xe + X(i);
      Ye = Ye + Y(i);
end
Xe = Xe/n;
Ye = Ye/n;

stats = 0;
for   i = 1 : n
      stats = stats + err(i)*err(i);
end
stats = stats/(n-2);

En1 = abs(Y - Yx);
En2 = 100*En1;

p = 1-alpha/2;
z = ltqnorm(p);

for   i = 2 : n
      somme = 0;
      for   j= 1 : n
            somme = somme + (X(i)-Xe) ^ 2;
      end
      borne_inf(i) = Yx(i) - (z*sqrt(stats)*sqrt(1+(1/(i-1)) + (((X(i)-Xe) ^ 2)/somme)));
      borne_sup(i) = Yx(i) + (z*sqrt(stats)*sqrt(1+(1/(i-1)) + (((X(i)-Xe) ^ 2)/somme)));
end
```

——————- Ci-dessous un exemple de simulations et de courbes obtenues ——————-

» Pour le nuage des points :

* Selon les formules statistiques :
  - La valeur moyenne = 0.0024
  - L'écart-type = 1.000
* Selon la méthode du maximum de vraisemblance :
  - La valeur moyenne = 0.0024, l'intervalle de confiance = [0 , 0.0323]
  - L'écart-type = 1.0003, l'intervalle de confiance = [0.9798 , 1.0220]

» Statistiques concernant les data testées :

* Le nombre total d'échantillons = 4321.0000
* Le nombre d'échantillons appartenant à la borne inférieur = 704.0000

* Le nombre d'échantillons appartenant à la borne supérieur = 711.0000
* Le nombre d'échantillons appartenant à la zone de confiance = 2906.0000

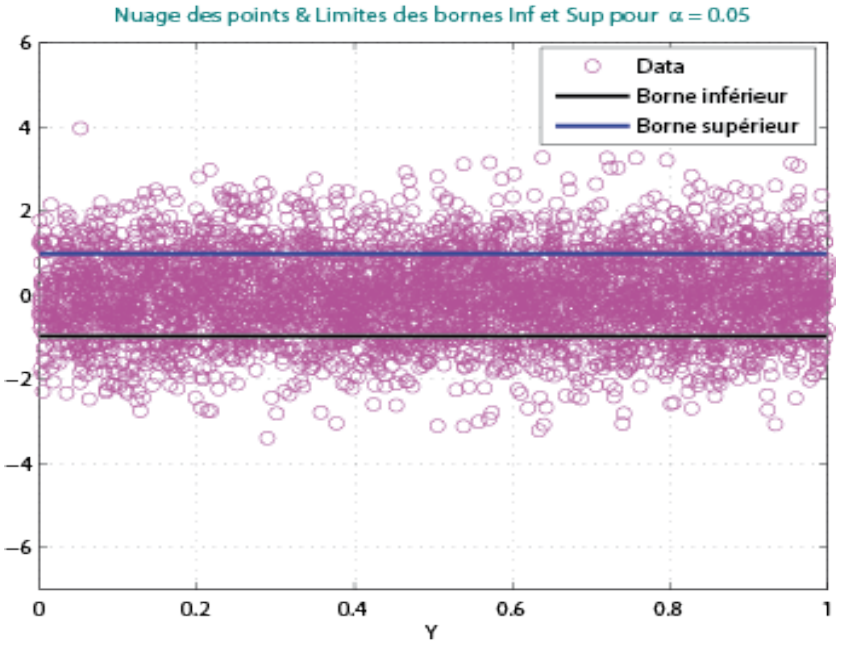


FIGURE 4 – Le nuage et les limites des bornes inférieur et supérieur des données massives générées, pour un niveau de confiance de l'ordre de 95%.

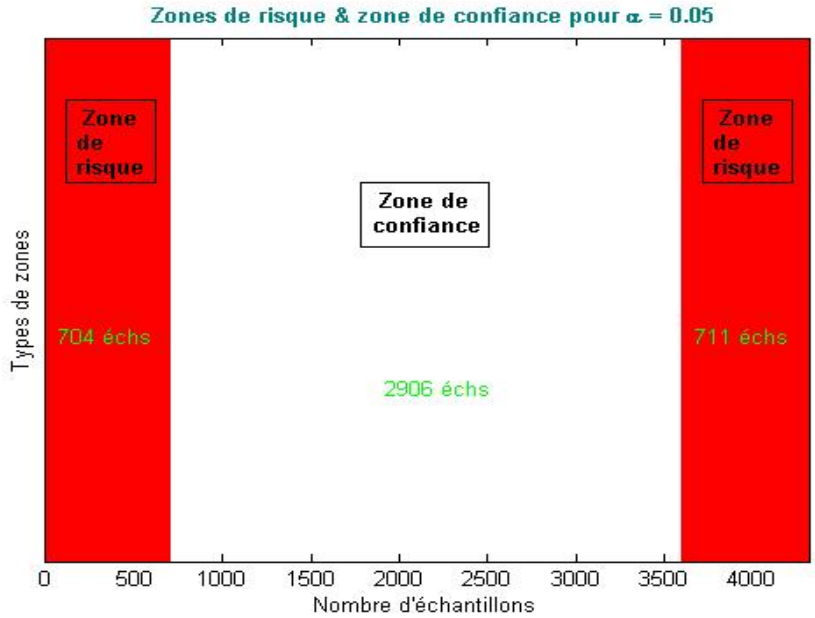


FIGURE 5 – Zones de confiance (2906 échantillons) et zones de risque (704 et 711 échantillons) pour un niveau de tolérance prescrit de l'ordre de 95%.

6. Exemple 1 de data collectées : Âges des étudiants d'une université.

Le tableau étalé ci-dessous représente un exemple de data amassées issues d'une statistique relative à l'âge d'une population étudiante d'un établissement universitaire. Cette enquête statistique assurée par un organisme spécifique, a pour but d'étudier le temps nécessaire aux étudiants universitaires pour terminer leurs études.
Il s'agit en fait d'un sondage effectué dans un contexte d'accorder la mis en place des politiques abouties (pays X) visant à inciter davantage les étudiants à entrer plus tôt (avec un âge moins jeune) sur le marché du travail. Ainsi l'analyse de cette pléthore numérique d'informations peut apporter un éclairage au suivi de l'évolution et peut aussi améliorer la prise de décisions en temps réel. Par ailleurs, cette investigation sur le devenir des étudiants a pour objectif d'informer le public, d'éclairer la réflexion politique et sociale et d'aider à l'amélioration des formations pour l'insertion future dans le monde professionnel.
Les données recueillies, exhibées ci-dessous, concernent la distribution d'âge des étudiants d'une formation tertiaire, inscrits entre la 1ère et la 5ème année universitaire. Ces data sont enregistrées aussi dans un fichier Excel baptisé "AgesEtudiants.xls".

FIGURE 6 – Photo d'étudiants universitaires dans un amphi.

Question 6 :

6-1. Charger à l'aide de l'instruction "***load()***" le fichier Excel "AgesEtudiants.xls", afin de récupérer les data y contenues et de les rendre disponibles dans le workspace.
Pour cela, imiter la syntaxe suivante :

$$\text{Data\_Ages} = \text{load('AgesEtudiants.xls')}\,;$$

6-2. Vérifier la longueur et les dimensions du fichier en question, puis donner l'âge du plus jeune et celui du plus âgé parmi ces étudiants.

6-3. Représenter à l'aide de l'instruction "***plot()***" le nuage correspondant à ces data glanées issues de cette enquête.

6-4. Calculer par les formules mathématiques, puis en utilisant l'instruction "***med()***" liée à la technique d'estimation par maximum de vraisemblance, la moyenne et l'écart-type, relatives aux données des âges des étudiants.
Remarque : Pour la méthode d'estimation par maximum de vraisemblance, afficher les résultats relatifs à un intervalle de confiance de l'ordre de 95%.

6-5. Superposer sur la même figure l'allure de l'histogramme des data collectées et celle de la loi normale relative à la moyenne et à l'écart-type trouvés ci-dessus. Interpréter.

23.61,21.84,21.41,21.37,20.46,22.05,24.59,20.90,21.98,21.20,20.18,20.93,21.97,22.24,20.09,21.59,21.21,20.96,22.71,22.25,22.25,21.23,22.94,20.67,22.00,22.08,20.39,21.34,22.32,17.02,
21.13,20.68,24.41,24.66,22.25,22.23,21.77,25.94,21.84,22.20,21.42,22.07,19.84,22.88,20.61,22.78,21.83,23.59,19.4751,20.6783,21.9028,21.189,22.0225,24.3884,20.2533,21.9838,17.38
55,23.4847,22.8176,24.2479,23.1894,20.0105,21.324,25.1516,24.2354,20.4426,21.5356,23.9258,22.0556,21.1614,19.5814,23.9314,22.3567,20.5084,21.3977,23.8482,21.3288,23.4556,2
3.0146,24.0439,24.3452,23.7383,23.4509,23.5211,21.1749,20.6369,20.1915,23.2342,20.722,22.487,21.7913,22.8422,23.1657,21.8039,24.1067,23.5942,24.4566,21.1762,20.6732,24.446
8,22.4504,23.0571,21.7554,23.2389,20.7707,21.5503,23.3176,22.3837,23.067,25.3339,22.6325,22.4429,22.5736,22.4963,20.3173,19.9359,21.2419,21.7201,23.042,20.3581,20.9891,21.2
053,22.2722,24.6599,19.905,21.1525,20.5289,19.6892,23.0116,21.9344,21.0272,24.4215,22.8209,24.1801,19.3281,24.1139,22.1272,22.908,22.7461,21.1971,21.4168,22.8275,25.3526,2
1.8064,23.8301,23.8895,22.0251,20.9252,22.1108,21.618,22.8304,21.9762,19.8371,21.5658,21.3926,24.5148,22.4824,24.2166,19.2272,22.9035,23.3487,21.316,20.0615,22.5887,23.257
7,19.9404,19.7292,25.3688,24.5148,21.0435,17.7924,19.5564,21.9192,21.4246,20.4099,21.9664,22.0153,23.1949,23.688,20.4001,23.595,21.6944,23.8715,23.9287,22.1128,21.2395,21.3
415,20.7356,22.1463,22.6697,25.8613,22.1273,22.6262,20.6173,24.2977,22.0638,23.1456,20.8304,21.0723,22.2104,24.3353,22.211,21.1548,21.46,22.2447,22.5356,22.6603,21.8923,22.
7876,22.2851,19.7228,22.1659,17.3757,23.6163,22.9588,18.0035,20.4234,24.0915,21.5887,20.8979,23.3121,22.378,22.1082,24.4689,22.4979,20.8285,22.5576,20.3502,21.2214,20.4756
,20.6882,23.0354,19.7939,20.5854,22.4876,20.1091,20.2713,21.9091,22.6516,20.5307,22.7725,24.2102,20.964,21.4984,21.5315,23.864,20.3351,24.5685,22.2156,22.0523,23.7483,23.98
73,20.7396,20.7079,23.5191,23.186,21.6912,24.2278,20.9696,21.6345,20.3914,23.5741,21.4248,23.628,21.819,22.6735,20.5614,22.5691,24.4183,20.9214,21.2813,22.8205,22.2456,23.1
665,20.9889,21.3892,20.3035,21.6413,20.1294,21.0504,21.9741,21.7855,20.2595,26.0748,21.7209,21.0298,22.4466,25.8659,24.2525,19.6907,25.3698,25.0583,20.2276,18.7441,22.9444
,23.6711,23.8524,22.9369,24.3076,20.4901,22.2702,21.5294,22.3349,21.422,22.4909,19.3753,20.1752,21.4638,21.0669,22.1744,20.8167,19.0778,24.08,23.1761,23.1448,20.5345,22.032
4,21.7268,20.4471,22.2989,23.579,22.1572,22.5754,24.7945,20.5282,21.0903,20.9776,21.3348,22.0319,19.8234,20.4558,21.5364,23.3289,22.3635,20.0738,21.8143,21.212,19.9764,24.5
195,22.4117,21.8558,21.5459,21.0608,20.7732,20.5677,21.5437,21.3101,19.4854,21.8906,23.9355,20.0243,21.8718,20.1095,25.1581,19.9342,21.9627,18.6053,21.7336,18.2667,22.0895
,21.895,23.8318,21.5301,20.4751,21.2308,23.4565,24.0173,20.4257,23.1766,23.2098,21.2889,22.7591,20.5131,22.4229,20.538,23.4348,23.6758,23.9163,20.8103,20.1809,23.0546,22.52
63,19.8644,22.6126,22.9867,21.0105,19.4691,22.9463,21.9663,20.0169,24.6067,21.3345,21.2907,21.6247,22.0517,23.5207,21.0188,21.7794,17.9944,19.4068,21.26,21.2571,19.9833,22.
5368,23.2949,21.9595,21.3267,22.6941,20.7432,20.548,19.4708,22.5494,21.9142,23.161,25.7488,20.5259,20.9114,20.2596,22.1095,21.4771,21.5185,19.8107,23.3829,23.7289,22.227,2
0.5466,24.135,20.71,22.0961,22.0898,22.5179,23.1235,23.9336,20.1968,20.6304,18.7286,21.2689,21.7275,21.9655,22.7759,19.5995,21.0997,21.1973,21.6092,21.2409,21.8347,19.2128,
22.2421,22.5038,20.1877,20.1844,20.5964,24.1974,21.1802,20.7839,22.9073,21.2892,22.7387,21.4673,19.9925,22.1209,25.2032,18.6975,23.7998,23.668,24.7017,21.6635,26.1271,22.3
527,22.8087,19.3223,22.0819,22.2764,20.6712,21.8832,21.2443,22.1878,24.1806,20.9122,20.3846,21.1686,22.3988,22.7618,22.7579,21.5288,22.886,20.1,20.6161,20.9904,22.9211,23.2
372,21.227,22.7128,21.6631,23.8136,18.804,22.7697,22.368,22.9973,22.6328,19.9602,20.5321,20.4211,21.3748,22.1425,22.3464,23.8862,23.9191,23.916,21.1812,22.4054,18.47,22.498
,22.1336,18.3914,24.644,20.8943,22.8967,23.0052,21.221,20.1163,23.4182,23.681,21.7957,21.9606,20.1827,21.4934,22.6522,21.1667,21.9489,20.0733,21.0286,22.184,24.2663,21.9717
,21.1658,21.4923,21.0373,20.6469,21.8694,21.9656,22.3248,20.4588,23.1018,22.1371,23.7947,22.3761,21.3187,24.5719,23.4047,20.9357,21.0657,20.1016,24.0218,19.7302,21.0334,19.
5877,20.5576,21.7585,21.7665,19.3971,22.6879,23.7813,21.4816,24.1564,21.8068,20.9749,23.3734,21.5154,20.5968,24.8296,23.8651,21.2518,21.7664,20.6756,21.46,23.1122,21.8433,
23.7105,19.2936,21.2794,21.2478,23.8183,23.4663,22.7491,19.1856,22.4687,20.2621,21.6354,25.6577,20.1915,24.8094,22.271,21.5782,24.9883,22.4761,25.0282,21.6149,21.8915,22.8
939,21.9566,21.9277,21.7137,19.7483,23.1157,25.9199,19.6739,24.551,22.3357,21.7305,22.5141,21.5372,21.9909,23.7764,24.0368,21.6641,20.4484,20.1564,21.6626,26.0219,20.1486,
21.1446,19.3143,20.5329,21.1744,21.2135,24.6693,22.3155,21.4881,22.7594,21.0238,21.9381,22.3569,21.5383,23.1648,22.8024,23.0135,23.0195,23.047,22.9107,22.0266,21.5515,23.6
725,23.062,21.7267,22.5698,21.1491,22.2937,23.9639,22.6225,20.367,23.6538,19.9048,23.4702,20.6345,22.6483,20.2301,23.5522,21.0213,21.978,23.231,22.4052,24.6921,22.4821,20.2
181,20.5363,24.1186,21.0321,22.4778,21.1647,23.2732,23.3368,23.6434,24.7221,22.4561,18.7182,21.0228,23.0752,22.7622,20.9806,22.3628,21.7922,22.3199,20.6507,23.4565,23.0084
,23.1772,21.5719,22.0424,23.8243,21.9652,21.247,20.963,18.5714,19.9313,22.0888,22.6335,20.7793,22.7425,21.4737,22.5596,22.0874,22.0245,21.6814,19.674,20.5234,19.989,21.7555
,22.9398,20.5391,25.1479,22.4826,22.3957,21.41,22.5439,23.3521,23.4195,23.8252,22.5564,21.5931,21.0718,22.3873,24.3194,23.168,21.8941,22.6341,20.6939,21.3538,23.0697,23.585
4,24.8795,20.5299,20.527,18.8383,20.2159,22.4024,21.4963,19.8543,24.0245,23.5167,23.3124,23.7125,20.7166,25.0092,22.6131,19.8207,20.3695,19.3765,20.9595,23.6904,22.5561,19.
8882,22.0978,22.0569,21.7759,21.6121,24.7844,22.8878,22.4278,20.5413,21.3695,20.1034,22.6173,20.1615,23.5357,21.8098,23.0309,22.1929,22.2984,21.6047,23.1317,22.2018,22.995
7,22.3716,18.8849,24.2612,22.2362,24.4121,20.2801,22.2344,21.8722,25.0942,24.0571,21.2429,20.9966,21.5104,21.8233,23.7242,23.9598,21.6121,22.2509,22.3253,23.1573,23.4725,2
3.5167,19.7608,21.6317,20.7811,20.7797,22.2168,22.3282,22.0939,24.0034,22.7831,21.7058,23.9918,22.3965,22.3228,19.9754,23.334,21.8537,22.1331,22.022,22.0998,23.2531,23.515
5,20.7031,25.002,21.9785,21.8599,22.4707,19.0499,20.3292,22.2958,24.2614,21.0234,23.0702,21.4293,23.9044,22.172,22.3877,20.1865,22.9999,22.1148,21.5033,21.191,21.2026,20.53
53,22.6357,22.2867,21.7287,18.6785,22.2657,20.8521,20.5066,23.6796,23.8732,21.6798,20.7733,22.343,21.5978,23.2381,21.8587,19.7267,20.3767,22.0657,22.9008,21.6795,19.8407,2
1.4753,25.0363,25.2243,22.4303,21.8969,21.7314,21.8288,24.527,19.8974,21.8043,23.7155,21.5858,20.0689,23.421,19.3536,21.3267,21.4792,23.8718,21.6614,22.5946,20.431,24.1783,
21.9369,21.032,18.1097,23.7653,21.1169,21.7924,21.3202,19.3734,22.1729,20.3198,24.7289,22.6108,21.5809,22.8568,23.3367,20.6199,22.3791,19.9845,22.4064,23.5702,23.0427,24.0
171,23.8055,21.069,21.5328,22.3381,25.6669,21.9666,23.0975,24.0473,20.2229,21.9153,21.8838,20.6783,24.6992,22.6926,21.7503,21.8061,21.6929,22.0405,21.2098,17.7611,19.4642,
22.2131,22.4586,21.9787,22.0795,19.1402,22.2243,21.2474,20.7135,21.1398,21.7658,21.1281,22.5209,19.8459,18.2876,20.9062,21.3941,21.4738,20.5873,24.574,21.397,24.3165,22.92
14,21.1312,22.1411,22.0009,22.8106,19.4808,23.5313,21.7071,23.7969,22.9652,23.0863,23.4027,20.1414,21.933,22.3857,24.4394,19.6958,21.6122,19.9767,20.1868,20.0054,20.7894,2
3.0571,20.5731,18.8505,23.1973,21.6618,23.0218,20.1564,22.169,23.9645,20.3494,20.352,20.3481,23.035,19.9036,22.502,22.1544,21.3049,20.8047,21.5806,21.1112,21.3372,20.1476,2
0.0045,23.7986,22.1854,20.6344,20.4993,21.2421,24.2056,23.4244,21.3185,21.3429,22.5092

FIGURE 7 – Data issues d'une statistique d'âge relative à une population étudiante (1000 étudiants) d'un établissement universitaire.

6-6. Représenter à l'aide de l'instruction "***cdf()***" l'allure de la fonction de distribution cumulative liée à ces data collectées.

6-7. Sachant que 80% des étudiants de cet établissement universitaire qui ont validé la 3ème année durant l'année scolaire du sondage sont âgés de 21 à 22 ans. Compléter le tableau suivant :

Situations	Probabilités de validation	Probabilités de non validation
Étudiant de 19 ans		
Étudiant de 20 ans		
Étudiant de 21,4 ans		
Étudiant de 23 ans		
Étudiant de 24 ans		
Étudiant de 26 ans		