**Test Analyse de données avec Matlab- M1 Sup-info 2020-21**

**Année : avril 2021**

**Formateur: Dr Cheikh Amadou Bamba Dath, Tel: 775230226/705024203**

**E-mail:** [**bambadath@gmail.com**](mailto:bambadath@gmail.com)

**Date limite de remise du projet par e-mail: 19 juin 2021 à 15 heures**

**Prénom : Mouhamadou Moustapha**

**Nom : Diop**

**Groupe : RTPR G16**

**Matricule : SIDK005088**

**Exercice** 1:

1-On suppose que 60% des habitants d’une ville sont des femmes et 40% des hommes. On veut former un échantillon de 200 individus en respectant ces strates. Quelle sera la part de chaque strate dans cet échantillon ?

2-cette ville compte 30 000 habitants, cet échantillon de 200 individus est-il représentatif pour effectuer un sondage le niveau de vie ou sur la prévalence d’une maladie ?

3-Si en plus, il est établi que 70% de la population est jeune avec moins de 35 ans et que 30% de la population chôment. Ces informations sont-elles doivent-elles être intégrer dans le choix de l’échantillon? Justifier

**Réponse 1:**

1. La part de chaque strate dans cet échantillon :

Hommes : 200 \* 40% = 80

Femmes : 200 \* 60% = 120

1. Population :

Hommes : 30 000 \* 40% = 12 000

Femmes : 30 000 \* 60% = 18 000

Oui notre échantillon 200 individus est représentatif dans le cas où ces proportions de 60 % des femmes et 40% des hommes sont appliquées.

1. Oui ces informations doivent être intégrées dans le choix de l’échantillon afin que ce dernier soit plus précis, fiable et représentatif.

**Exercice 2 : Questions** :

1. Rappel des principales méthodes de constitution de l’échantillon
2. Comment déterminer la taille de l’échantillon ? Donner la formule, et expliquer tous les termes.
3. Pourquoi prendre un échantillon dans une étude?
4. Donner les notions de paramètres de position et de dispersion et donner deux exemples pour chacune des notions.

**Réponse 2 :**

1. Les principales méthodes de constitution de l’échantillon sont :
2. **Echantillonnage aléatoire simple**
3. **Echantillonnage par grappes**
4. **Echantillonnage par la méthode des quotas**
5. Détermination de la taille de l’échantillon :
6. **La taille de la population mère** : Plus la population n’est importante, plus on a besoin d’un échantillon de plus grande taille. Cependant, lorsqu’il s’agit de très grandes populations, la taille de la population n’a plus d’influence sur la taille de l’échantillon.
7. **La variabilité des caractéristiques de la population mère** : Plus la population mère qui vous intéresse est diverse et présente des caractéristiques variées, plus il faudra interroger de personnes. A l’inverse, plus cette population est homogène et moins il faudra interroger de personnes.

**Formule de la taille de l’échantillon**

La formule de la taille de l’échantillon est :

**n = tα² ∗ p (1 - p) ∗ N / tα² ∗ p (1 - p) + (N - 1) ∗ m²**

**n** = taille de l’échantillon ;  
**tα** = intervalle de confiance selon la loi normale centrée réduite (pour un niveau de confiance de 95%, **tα = 1.96**, pour un niveau de confiance de 99%, **tα = 2.575**, pour 80 % **tα =1,28** pour 85 % **tα =1,44** , pour 90 % **tα = 1,65**) ;  
**p** = proportion estimée de la population qui présente la caractéristique (lorsque inconnue, on utilise **p = 0.5** ce qui correspond au cas le plus défavorable c’est-à-dire la dispersion la plus grande).  
**m** = marge d’erreur tolérée (par exemple on veut connaître la proportion réelle à 5% près)  
Cette formule détermine le nombre de personnes **n** à interroger en fonction de la marge d’erreur **m** que l’on peut tolérer sur une proportion de réponses **p**.

**N** est la population à étudier ou base de sondage utilisée.  
Si **N** est très grand on peut réduire la formule de calcul de l’échantillon à :

**n = z² x p (1 – p) / m²**

1. On utilise des échantillons parce qu’une étude approfondie de toute la population serait impossible, trop longue et trop chère pour le niveau de précision exigé.

Ainsi, à partir d’une population source, on va sélectionner un échantillon de sujets respectant différents caractères.

1. **Paramètres de position**: ils donnent l'ordre de grandeur des observations et sont  
   liées à la tendance centrale de la distribution.

**Exemples :**

* La moyenne
* La médiane

**Paramètres de dispersion**: ils montrent la manière dont les observations  
Fluctuent autour de la tendance centrale.

**Exemples :**

* L’écart-type
* La variance

**Exercice3 :**

L’ensemble des PME et startups en numérique et transport utilisant les rayonnements enregistrées dans un pays est de 8500 unités économiques.

La famille du secteur transport et télécommunication, d’après le recensement général. de la population et de l’habitat de ce pays, estime l’ensemble de ces entreprises égales à 0,4% de l’effectif total de l’ensemble des entreprises de ce pays. Le nombre d’entreprise total d’entreprises de ce pays était de 807 8820.

**Etape 0 : compléter le tableau de répartition des 8500 entreprises (base de sondage) suivant :**

|  |  |  |
| --- | --- | --- |
| *Stratification de la population entreprise du numérique* | | |
| Zone Géographique | Nombre Entreprise | Proportion |
| Autres régions | 425 | 5% |
| Da | 7 140 | 84% |
| Dl | 255 | 3 % |
| SL | 170 | 2% |
| T | 425 | 5% |
| Z | 85 | 1% |
| Total | 8 500 | 100% |

**Etape 1** **: donner la méthode d’échantillonnage et calculer de la proportion p**

:

**Méthode d’échantillonnage : Par grappes ou quotas**

**Calcul de la proportion p :**

Nombre total de PME/Startup du numérique transport utilisant le rayonnement = 8 500

Nombre total d’entreprise du secteur numérique et transport = 0.4% \* 8 078 820 = 32 315

p = 8 500 / 32 315

**p = 0.26**

**AN :**

***Etape 2 : déterminer la taille de l’échantillon :***

La taille de notre échantillon est choisie avec les spécifications suivantes :

* La marge d’erreur, notée m = 7%
* Le risque maximum, noté tα : sa valeur est issue d’une loi de probabilité (loi de Student ; pour une population de taille notée N < à 30, loi normale : pour une population de taille N > à 30 le cas de cette étude). Soit tα = 1,65 pour un intervalle de confiance de 90%.
* La connaissance statistique de la proportion attendue d’une réponse de la population ou proportion réelle.

**Détermination de la taille de l’échantillon : n**

p = 0.26

N = 8 500 > 30 (base de sondage)

tα = 1,65

n = (1.65)² \* 0.26 (1 – 0.26) \* 8500 / (1.65)² \* 0.26 (1 – 0.26) + (8500 – 1) \* (0.07)²

n = 4452,3765 / 595,453809

n = 105.58

**n est aux environs de 106 donc (n = 106)**

**Taux de sondage = 106 / 8500**

**Taux de sondage = 0.012 soit 1.2 %**

**AN**:

On peut dire implicitement que nous avons fixé un taux de réponse (ou taux de sondage)

Qui est donné par :   **taille n de l’échantillon / base de sondage**,

Ce taux de réponse doit être supérieur ou **égale à 20 %, donc il est acceptable**

**AN : taille n de l’échantillon / base de sondage**

* **Etape 3 : Tirage de l’échantillon**

On peut subdivise la population en strates selon la zone et l’échantillon est choisi en tirant au sort dans chacune des strates (on réalise un sondage sur chacune des strates).

**Selon les cas, on peut choisir d’autres types de tirage de l’échantillon : échantillonnage aléatoire simple, échantillonnage par quota, échantillonnage systématique. Echantillonnage par grappe.**

Compléter le tableau suivant pour la répartition des éléments de l’échantillon

|  |  |  |
| --- | --- | --- |
| *Stratification de la population entreprise de l’échantillon* | | |
| Zone Géographique | Nombre Entreprise | Proportion |
| Autres régions |  | 5% |
| Da |  | 84% |
| Dl |  | 3 % |
| SL |  | 2% |
| T |  | 5% |
| Z |  | 1% |
| Total |  | 100% |

**Etape 4 : élaboration, validation du questionnaire (enquête)**

**En exemple élaborer un ensemble de 4 questions :**

**-1 statut : PME/Start Up**

**2 principales activités : réseaux, développement**

**3-Zone d’implantation : rurale, urbaine**

**4-Principaux partenaire : public/privé**

**Etape 5 : recueil et traitement des données : -tri à plat, -Tri croisé, -Modélisation et approximation**

Recueillir et traiter les réponses, faire une analyses des données.

**Résolution**

1 – Statut :

PME START UP

2 – Principales activités :

Réseaux Développement

3 – Zones d’implantation :

Rurale Urbaine

4 – Principaux partenaires :

Public Privé ONGs et autres

**Recueil et Traitement des données brutes :**

On a recueilli 800 réponses lors de cette étude.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Questions | Statut | | Principales activités | | Zones d’implantation | | Principaux partenaires | | | **Total** |
| Réponses | 100 | | 200 | | 200 | | 300 | | | **800** |
| Choix | PME | Start up | Réseaux | Dév. | Rurale | Urbaine | Public | Privé | ONGs et autres |  |
| Réponses | 70 | 30 | 50 | 150 | 20 | 180 | 20 | 200 | 80 | **800** |
| Taux de réponse (%) | 8.75 | 3.75 | 6.25 | 18.75 | 2.5 | 22.5 | 2.5 | 25 | 10 | **100%** |

Nombre des choix de réponse = 9

Effectif total = 800

Moyenne arithmétique = 800 / 9

**Moyenne arithmétique = 88,89**

**Calcul Moyenne - Variance et Ecart-type :**  **Code Matlab**

clear all

% Analyse de données recueil des données du questionnaire

donnees\_questionnaire = [70, 30, 50, 150, 20, 180, 20, 200, 80];

% moyenne

moy = mean(donnees\_questionnaire) % moy = 88.8889

% variance

variance = var(donnees\_questionnaire) % variance = 4.9111e+03

% écart-type

ecart\_type = sqrt(variance) % ecart\_type = 70.0793

**Médiane = 200**

**Modes :** 300

**Code Matlab de l’analyse de données : Cf fichier matlab (analyse\_questionnaire.m) – dans dossier du projet**

clear all

% Analyse de données recueil des données du questionnaire

donnees\_questionnaire = [70, 30, 50, 150, 20, 180, 20, 200, 80];

% taux de réponse en %

Taux\_reponse = [8.75, 3.75, 6.25, 18.75, 2.5, 22.5, 2.5, 25, 10];

% choix de réponse

choix\_reponse = {'PME', 'Start up', 'Réseau', 'Développement', 'Rurale',...

'Urbaine', 'Public', 'Privé', 'ONGs et autres'};

% moyenne

moy = mean(donnees\_questionnaire) % moy = 88.8889

% variance

variance = var(donnees\_questionnaire) % variance = 4.9111e+03

% écart-type

ecart\_type = sqrt(variance) % ecart\_type = 70.0793

subplot(1,2,1)

plot(donnees\_questionnaire, Taux\_reponse, '-\*b')

%Le titre de la courbe

title('Taux de réponse', 'FontSize', 10, 'color', 'b')

xlabel('Réponses/choix', 'FontSize', 12)

ylabel('Taux de réponse', 'FontSize', 10, 'color', 'red')

legend('Taux de réponse')

hold on

%gradriller la fenetre avec grid on

grid on

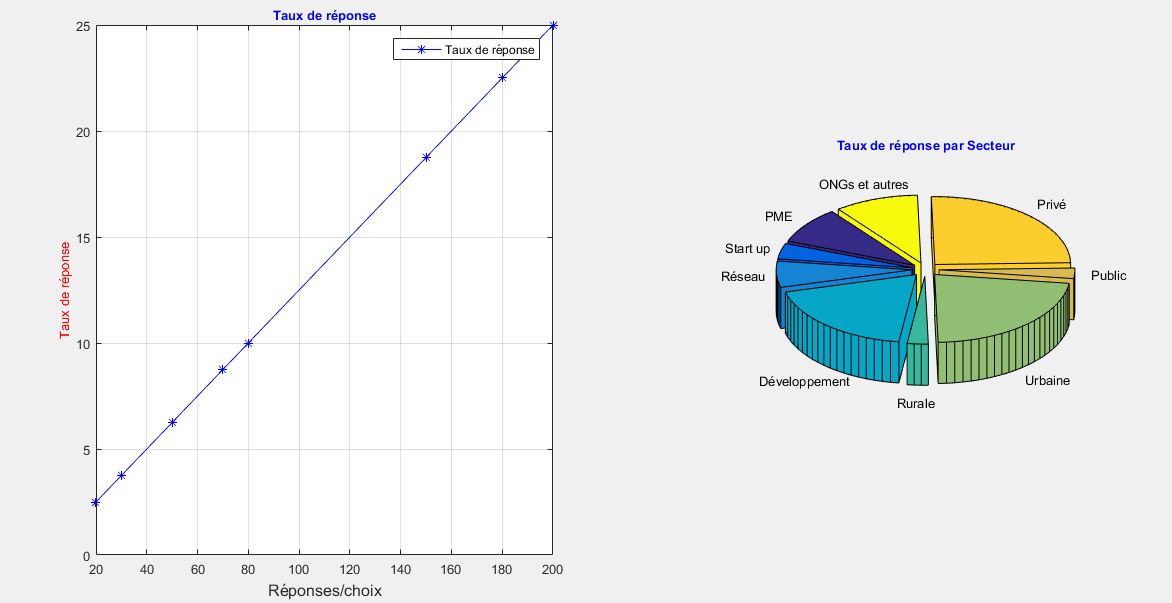
subplot(1,2,2);

explode = [1 1 1 1 1 1 1 1 1]; % espacement

pie3(Taux\_reponse , explode ,choix\_reponse)

title('Taux de réponse par Secteur', 'FontSize', 10, 'color', 'b')

**Visualisation sous forme de graphes**



**Exercise 4** :

La fonction densité de probabilité pour une distribution normale de moyennehttps://ch.mathworks.com/help/examples/matlab/win64/DetermineUnderlyingProbabilityDistributionExample_02.png, d’écart type https://ch.mathworks.com/help/examples/matlab/win64/DetermineUnderlyingProbabilityDistributionExample_03.png, et de variance https://ch.mathworks.com/help/examples/matlab/win64/DetermineUnderlyingProbabilityDistributionExample_04.pngest

https://ch.mathworks.com/help/examples/matlab/win64/DetermineUnderlyingProbabilityDistributionExample_05.png

**1-Compléter et interpréter le code suivant pour le tracé  de cette fonction.**

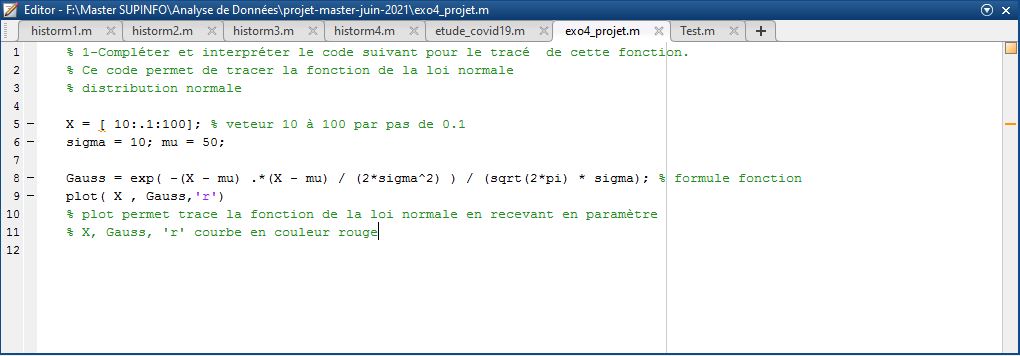
**X = [ 10:.1:100];**

**sigma = 10; mu = 50;**

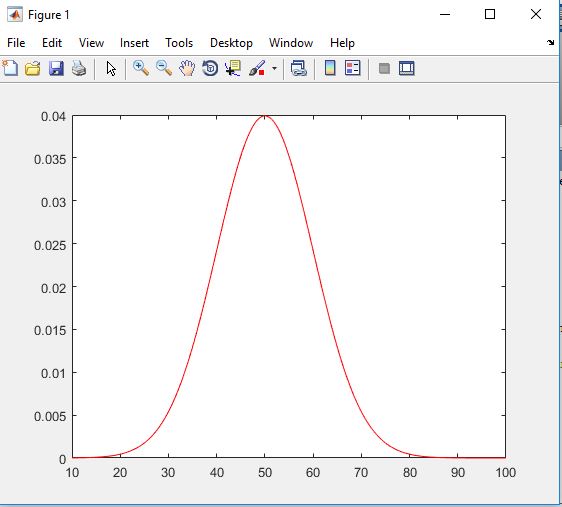
**Gauss = exp( -(X - mu) .\*(X - mu) / (2\*sigma^2) ) / (sqrt(2\*pi) \* sigma);**

**plot( X , Gauss,'r')**

**Interprétation du code :**

****

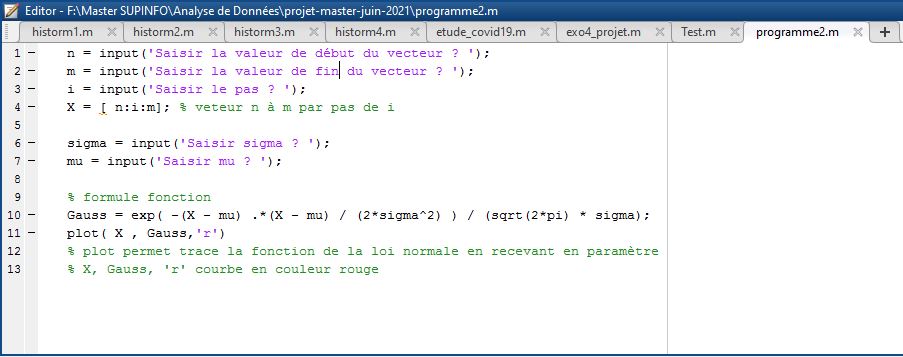
**Résultat du programme :**

****

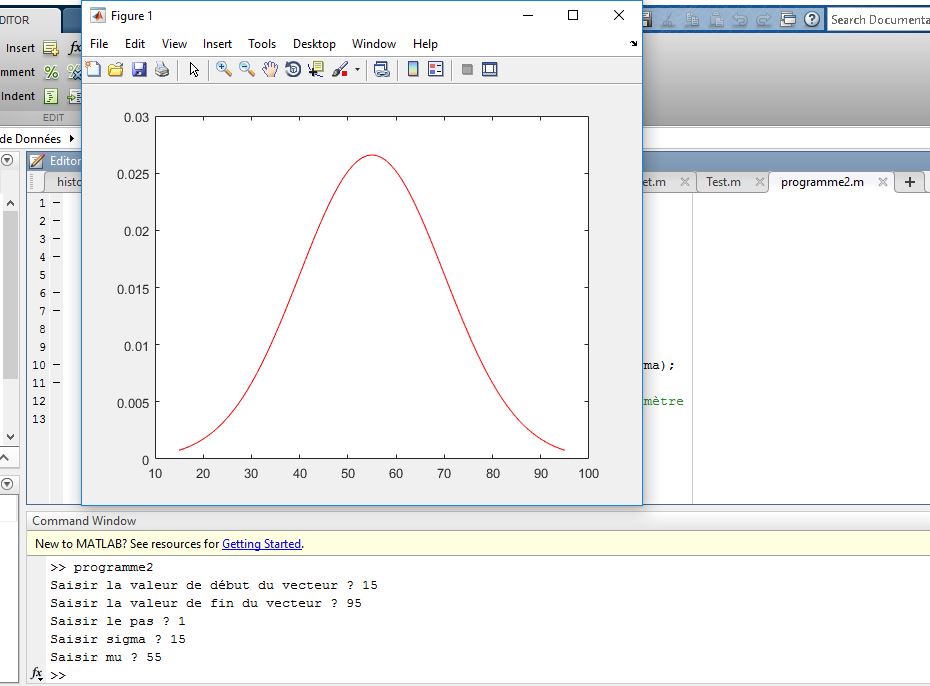
**2-Compléter ce code suivant pour tracer une gaussienne d’une autre façon en entrant les données au clavier.**

**Réponse :**

**Programme 2:**

****

**Résultat du programme 2 :**

****

**Exercice 5** : **On donne deux vecteurs suivants, représentant un ensemble d’observations :**

Nb\_tests\_real=[1,2,2,2,2,2,2,6,1,8,6,11,3,2,21,44,27,22,56,31,85,60,59,142,130,144,98,151,87,97,195,127,153,177,92,95,110,108,199,195,173,104,144,272,227,435,287,226,313,281,466,482,387,528,703,579,677,873,800,116,901,1190,810,807,702,1182,1027,872,848,1356,917,1099,920,992,1128,989,816,1172,930,1042,1251,1199,934,889,1091,551,961,991,1231,1370,1024,1820,1339,1549,1348,1700,1166,1266,1010,1402,1205,1164,1005,1378,1223,1254,999,970,1406,1617,1247,978,1040,1045,987];

Nb\_tests\_posit=[1,1,2,0,0,0,0,0,0,1,5,11,3,2,1,4,5,2,9,9,11,12,7,13,6,14,11,12,20,13,15,5,12,12,3,4,11,7,6,15,13,2,11,8,15,21,7,8,17,10,35,30,37,66,69,57,64,87,59, 51,91,91,67,89,58,104,59,59,83,75,177,109,110,84,119,...

119,51,64,73,97,98,97,67,71,83,31,92,95,81,106,110,84,97,96,89,134,94,79,99,89,124,119,92,145,94,83,74,122,106,164,144,105,82,64,95];

1-Les deux vecteurs représentent respectivement le nombre de tests réalisés et le nombre de test positifs sur la covid-19 au Sénégal du 02 mars 2020 au mois de juin 2020.

1. Donner le nombre de jours testés avec Matlab
2. Données le vecteur taux de positivité journalière à partir de données des deux premiers vecteurs

2-representé le taux de positivité et le nombre de cas positifs par jour

3-legender, titrer nominer les axes et ajuster les couleurs et traits pour affiner la représentation graphique.

Représenter le nombre de test réalisés dans une partie différente de la fenêtre graphique et l’affiner

4—Représenter le diagramme en moustache des trois vecteurs dans une fenêtre séparée et l’interpréter.

**Réponse 5 : Fichier : etude\_covid19.m (A voir pour plus de détails sur cet exercice)**

1. Le nombre de jours testés avec Matlab :

**Code :**

% longueur du vecteur Nb\_tests\_real est égale au nombre de jours

nbre\_jours\_testes = length(Nb\_tests\_real) % nbre\_jours\_testes = 115 jours

b - Donne le vecteur taux de positivité journalière à partir de données des deux premiers vecteurs

**Code :**

% b - Données le vecteur taux de positivité journalière à partir de données des deux premiers vecteurs

%vect\_taux\_positivite vecteur taux de positivité

vect\_taux\_positivite = Nb\_tests\_posit ./ Nb\_tests\_real

2-representé le taux de positivité et le nombre de cas positifs par jour

% 2 - representé le taux de positivité et le nombre de cas positifs par jour

jours\_testes = 1:nbre\_jours\_testes % jours\_testes = 1:115

%Partage fenetre en 2

subplot(1,4,1);

plot(jours\_testes, vect\_taux\_positivite, '-\*b');

3 –

%Le titre de la courbe

title('Taux de positivité journalier', 'FontSize', 10, 'color', 'b')

xlabel('Jours testés', 'FontSize', 12)

ylabel('Taux de positvité', 'FontSize', 10, 'color', 'red')

legend('Taux positivité journalier')

hold on

%gradriller la fenetre avec grid on

grid on

subplot(1,4,2);

plot(jours\_testes, Nb\_tests\_posit, '-\*r');

%Le titre de la courbe

title('Nombre de tests positifs journalier', 'FontSize', 10, 'color', 'b')

xlabel('Jours testés', 'FontSize', 12)

ylabel('Nombre de tests positifs', 'FontSize', 10, 'color', 'red')

legend('Tests positifs-journalier')

% Représenter le nombre de test réalisés dans une partie différente de la fenêtre graphique et l’affiner

hold on

grid on

subplot(1,4,3);

plot(jours\_testes, Nb\_tests\_real, '-\*y')

title('Nombre de tests réalisés', 'FontSize', 10, 'color', 'b')

xlabel('Jours testés', 'FontSize', 12)

ylabel('Nombre de tests réalisés', 'FontSize', 10, 'color', 'red')

legend('Nombre de test réalisé')

4 –

% 4—Représenter le diagramme en moustache des trois vecteurs dans une fenêtre séparée et l’interpréter

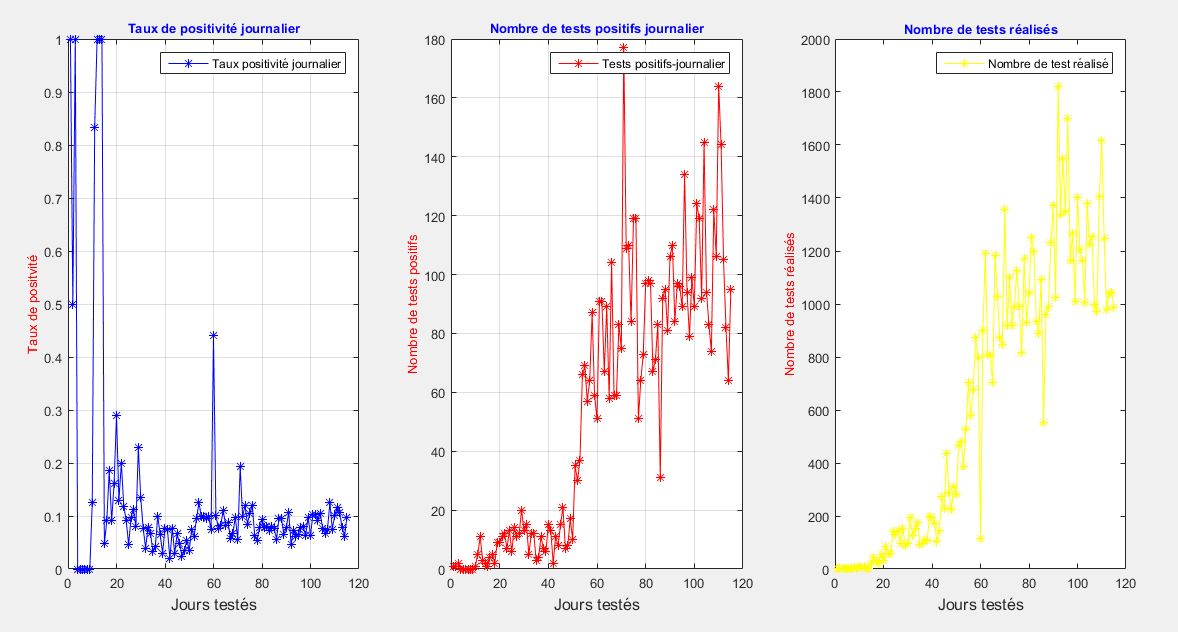
hold on

grid on

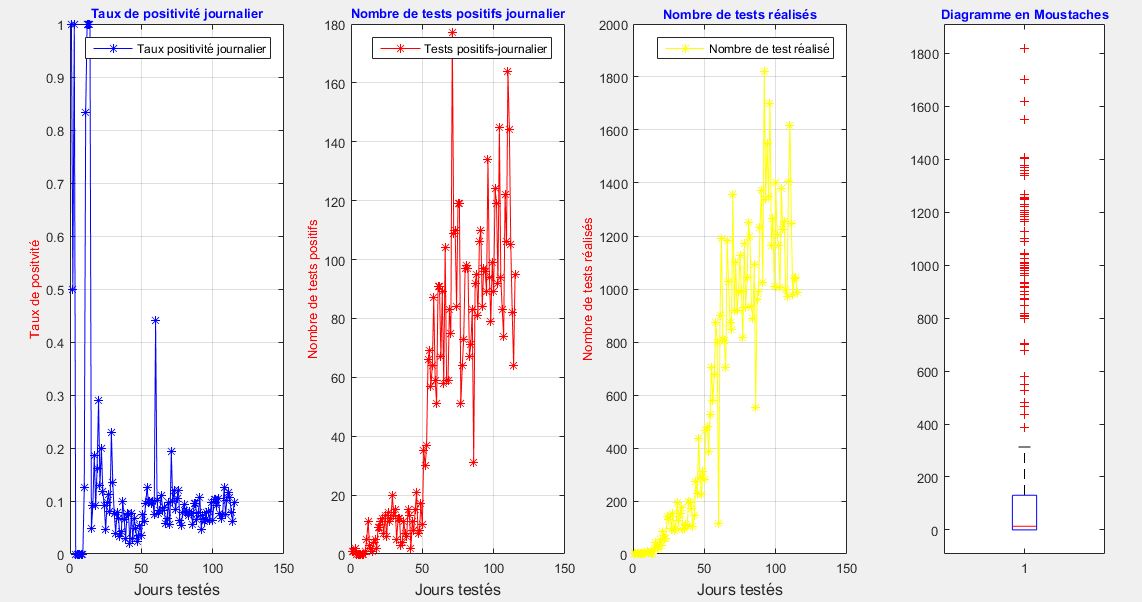
subplot(1,4,4);

boxplot([Nb\_tests\_real Nb\_tests\_posit vect\_taux\_positivite])

**Résultat étude covid-19 au Sénégal**



**Résultat étude covid-19 au Sénégal : Diagramme en Moustache**

****