MASTER : Traitement de l'Information et Exploitation des Données
Fouad Badran, Cécile Mallet, Carlos, Mejia,
Charles Sorror, Sylvie Thiria

TPA02 : Statistiques descriptives unidimensionnelles

I - Objectif

Le but de ce TP est de traiter de données statistiques unidimensionnelles, pour lesquelles il faudra déterminer des indicateurs élémentaires (moyenne, écart type, médiane, ...) et donner des représentations graphiques comme par exemple des camemberts, des histogrammes, et des boîtes à moustaches). L'objectif des exercices proposés est d'apporter une première compréhension de l'information véhiculée par les données à travers ces indicateurs et représentations.

Pour certains rappels de statistiques, vous pouvez vous rendre sur le site de statnet à l'adresse : http://www.agro-montpellier.fr/cnam-lr/statnet/cours.htm

II - Les Données

Trois sujets sont proposés dans ce TP : le 1er porte sur le lancé de dés, le second aborde la représentation par histogramme de données simulées (selon une loi de probabilité) et une illustration du théorème de la limite centrale, le 3ème et dernier sujet traite de données réelles qui concernent la pluviométrie au Sénégal. Celles-ci seront décrites lors de la présentation du sujet.

III - Eléments pour la réalisation du TP

La plupart des fonctions nécessaires à la réalisation de ce TP sont des fonctions Matlab. Souvent, nous vous donnerons des indications sur les fonctions à utiliser sans pour autant entrer dans le détail de la programmation qu'il vous incombera de prendre à votre charge. Nous avons par ailleurs jugé utile de fournir 2 fonctions ad hoc supplémentaires pour ne pas compliquer d'avantage le TP. Il s'agit des fonctions :

- **hncfd** : Représentation d'un histogramme normalisé et des courbes des fonctions de densité empirique et théorique associées.

- **moustache** : produit une figure des boîtes à moustaches de chaque colonne de la matrice passée en paramètre. Cette fonction à été réalisée pour être compatible avec Octave, ceux qui travaille avec Matlab, peuvent si il le souhaite utiliser la fonction **boxplot** à la place.

Vous pouvez vous référer à la fonction d'aide (**help**) de ces fonctions pour avoir des précisions sur leur utilisation.

IV - Statistiques descriptives unidimensionnelles

Pour le travail demandé, vous aller être amenés à produire des résultats souvent sous forme de figures graphiques qu'il conviendra de présenter. Cette présentation doit répondre à deux préoccupations :

- décrire factuellement les visualisations et les conditions dans lesquelles elles ont été obtenues
- Rédiger des commentaires synthétiques mettant en avant l'interprétation. Ce 2ème point ne doit pas être négligé car c'est celui qui apporte une réelle plus value au travail effectué.

1) Lancés d'un dé :

♦ Pour cet exercice, on vous demande de représenter par un camembert la répartition des valeurs prises par un dé lancé **n** fois. Chacune des valeurs de **1** à **6** constitue une classe. On admet que cette valeur suit une loi uniforme, on pourra donc simuler les lancés de dé avec la fonction **rand**. Vous devrez mener des expériences pour les différentes valeurs **n** du nombre de lancés du dé suivantes : **n=10**, **n=20**, **n=50**, **n=100**, **n=500** et **n=2000** ; et présenter les camemberts en indiquant la fréquence d'apparition de chacune des classes pour chacune de ces expériences. Comparez les résultats.

Indication pour la réalisation d'un camembert :

Soit **X** le résultat d'un tirage aléatoire de **n** lancés de dé. Pour en connaitre la répartition dans chaque classe vous pouvez utiliser la fonction **hist** : **N = hist(X,6)**. Vous pourrez ensuite utiliser la fonction **pie** pour dessiner le camembert : **pie(N)**. En plus du paramètre **N** vous pouvez aussi passer des labels à associer à chaque part du camembert (faites un **help pie** ou aidez-vous d'internet pour avoir plus de détail), par exemple :

```
for k=1:6;  pielab{k} = sprintf('%d %.0f%c',k,Pc(k),'%'); end  %avec Pc les pourcentages par classe.
```

♦ On vous demande également, pour chaque expérience de calculer les indicateurs suivants sur la répartition des classes :

- le pourcentage minimum
- le pourcentage maximum
- la différence entre le pourcentage maximum et le pourcentage minimum
- l'écart type des pourcentages.

(nb : les pourcentages pourront être arrondis)

Et de représenter ces informations sur une même figure que vous devrez commenter (aussi en liaison avec les camemberts).

Incidemment pour cette figure, selon que vous ayez indiqué ou pas des valeurs en abscisse, vous pouvez souhaiter modifier les labels de l'axe des x. Cela peut ce faire avec l'instruction **set**, sur la variable **gca** qui est le pointeur (handle) sur les axes courants, exemple :

```
set(gca,'Xtick',[1:6],'XTickLabel',{'10','20', '50', '100', '500', '2000' }); .
```

♦ Pour finir, vous devrez constater que les moyennes des lancés tendent de plus en plus vers l'espérance lorsque **n** augmente.

2) histogramme normalisé et fonction de densité

On appelle histogramme normalisé la représentation de la distribution empirique des données : le rectangle construit sur chaque classe a une surface proportionnelle à la fréquence d'apparition de la classe. La surface totale d'un histogramme normalisé vaut 1 ce qui est utile d'une part pour ajuster un histogramme avec une courbe de densité et d'autre part pour comparer des histogrammes entre eux.

1°) Utilisation de données unidimensionnelles simulées

1.1°) Pour simuler un jeu de données, vous devez créer une fonction **datagen(Loi,a,b,N)** qui :
Si Loi=='norm'

Génère N données selon la loi de probabilité $N(a,b^2)$ (densité de probabilité gaussienne de d'espérance E = a et de variance $V = b^2$)
Vous devrez utiliser la fonction **randn** qui renvoie des valeurs selon une distribution N(0,1). Ces valeurs devrons donc être transformés pour correspondre à une distribution $N(a,b^2)$

Si loi=='unif'

Génère N données selon la loi de probabilité U(a,b) Uniforme entre a et b
Vous devrez utiliser la fonction **rand** qui renvoie des valeurs uniformes dans [0,1]. Ces valeurs devrons donc être transformés pour correspondre à une distribution U(a,b)

1.2°) Avec la fonction **datagen** que vous aurez créée, on vous demande de simuler un jeu de 500 données selon une loi N(3,5) puis selon une loi U(3,5). Dans chacun de ces 2 cas, vous devrez vérifier l'adéquation entre les paramètres empiriques et théorique puis produire :

- un histogramme des données brutes avec la fonction **hist**
- et, un histogramme normalisé.

Pour produire l'histogramme normalisé, nous mettons à votre disposition la fonction ad hoc **hncfd** qui en plus de produire l'histogramme normalisé trace, en fonction de la loi choisie, une courbe de la fonction de densité empirique et une courbe de la fonction de densité théorique. Faire un « **help hncfd** » pour connaitre les paramètres à passer à la fonction **hncfd**.

2°) Illustration du théorème de la limite centrale (TCL)

Le TCL rappelé dans l'encadré ci-dessous indique que la moyenne d'une suite de variables aléatoires tend vers une loi normale. Pour illustrer cela vous devrez :

- Simuler **p** tirages de taille **n** selon une loi uniforme **U(a,b)** avec **a=3** et **b=5**. Vous pourrez utiliser la fonction **datagen** à condition de l'adapter avec un paramètre supplémentaire pour le nombre **p** de tirages que la fonction devra prendre en compte.
- Calculer pour chaque tirage sa moyenne empirique (X_bar), et utiliser la fonction **hncfd** pour représenter l'histogramme normalisé des **p** moyennes obtenues, elles suivent une loi $N(\mu,\sigma^2/n)$. On rappelle qu'une loi uniforme U(a,b) a une densité de probabilité d'espérance $\mu=(a+b)/2$ et de variance $\sigma^2=(b-a)^2/12$.

  Vous devrez utiliser les valeurs suivantes : pour **p** : **p=10** et **p=200**, pour **n** : **n=5** ; **n=50** et **n=500** ce qui représente 6 cas différents. Pour une meilleure visualisation, nous vous suggérons d'utiliser la fonction **axis** pour avoir seulement des abscisses à la même échelle mais en conservant telles quelles les ordonnées. Cela peut par exemple se faire de la façon suivant (après l'appel à hncfd) : axis([3.2 4.8 ylim]); où 3.2 et 4.8 sont des valeurs choisies par expérimentation et ylim est une fonction qui renvoie les coordonnées min et max de

l'ordonnée de l'axe courant.

Qu'observez vous lorsque **n** augmente selon les valeurs de **p**.

Théorème Central Limite (TCL) :

Soient X_1, ..., X_i, ..., X_n une suite de variables aléatoires indépendantes et identiquement distribuées d'espérance μ et de variance σ^2 finie. On note $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Théorème central limite : $\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{n\to\infty} N(0,1)$ *soit* $\overline{X} \xrightarrow{n\to\infty} N(\mu, \sigma^2/n)$

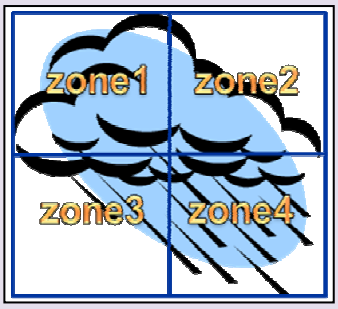
La distribution normalisée d'une moyenne tends asymptotiquement vers une loi normale

- Pour compléter vos résultats on vous demande de :
  - représenter pour chaque tirage p, et sur un même repère, les valeurs moyennes de X_bar en fonction de n. On y représentera également la moyenne théorique.
  - représenter pour chaque tirage p, et sur un autre repère, la variance de X_bar en fonction de n, ainsi que la variance théorique.

3°) Convergence de la densité empirique vers la densité théorique

Nous vous demandons de faire apparaître graphiquement la convergence de la densité empirique vers la densité théorique. Pour cela vous devez réaliser **12** expériences de **p** tirages aléatoires, selon une loi uniforme, dont, comme lors de l'exercice précédent, vous prendrez la moyenne pour construire des variables X_bar. Pour les 12 expériences, vous devrez successivement utiliser les valeurs suivantes de **p : 5, 10 20, 30, 40, 50, 100, 200, 500, 1000, 2000, 5000**. Vous n'utiliserez qu'une seule valeur pour la taille des données fixée à **n=10** pour tous les tirages. En utilisant la fonction **hncfd**, présentez pour chaque valeur de p un repère avec uniquement cette fois, les deux courbes de densité, théorique et empirique (c'est-à-dire, sans faire apparaitre d'histogramme). Pour pouvoir mieux apprécier la convergence, définir des axes identiques pour tous les repères.

3) Statistiques descriptives (sur des données réelles)



Nous vous proposons maintenant de travailler avec des données environnementales réelles qui sont des relevés de précipitation ; il s'agit de cumuls annuels effectués sur 4 zones géographiques du Sénégal pour 51 années de 1950 à 2000. Ces 4 zones qui constituent nos 4 différentes variables, seront notées avec un indice : zone1, zone2, zone3 et zone4 correspondant respectivement au nord-ouest, nord-est, sud ouest et sud-est.

Les données sont disponibles dans le fichier **pluie.txt** dont la 1ère colonne indique l'année et les colonnes suivantes les moyennes annuelles de la pluviométrie en mm dans l'ordre des 4 zones sus mentionnées.

Les résultats que nous vous demandons de produire ci-après devront faire l'objet d'une description permettant de synthétiser les informations qui émergent. Même si cela n'est pas précisé à chaque fois, il sera judicieux d'habiller les figures avec des éléments facilitant leur compréhension (choix des couleurs, légende, titre, ...).

1°) Après avoir chargé les données, nous vous demandons de présenter un tableau dans lequel, devront être indiqués, pour chaque zone : la moyenne, l'écart type, le minimum, le maximum et l'étendue.

Le tableau devra être accompagné d'une figure des courbes de chaque variable, avec des couleurs (ou marqueurs) différentes (et) une légende (et un titre). L'abscisse devra être labellisée par les années.

2°) Pour chaque zone, on vous demande de présenter, sur des figures différentes cette fois, les courbes des séries chronologique de pluie et leur encadrement à plus ou moins 1 écart type de la moyenne. Pour matérialiser l'encadrement, vous pouvez utiliser la fonction **line** pour tracer les lignes correspondantes à :

- la moyenne plus l'écart type,
- la moyenne moins l'écart type.
- la ligne de la moyenne pourra également apparaitre.

Vous devrez également mentionner, soit dans le rapport, soit directement sur la figure, les pourcentages de points inclus dans l'encadrement.

3°) Nous vous demandons de calculer, pour chaque zone, les quartiles (Q1, Q2, Q3) ainsi que l'écart interquartile (Q3-Q1), (fonction à utiliser : **prctile**). Vous devrez également en faire une représentation graphique en boîtes à moustaches avec la fonction **moustache**.