MASTER : Traitement de l'Information et Exploitation des Données

Fouad Badran, Cécile Mallet, Carlos Mejia,
Charles Sorror, Sylvie Thiria

Master TRIED

TPA02 : Rapport

Sujet :

Statistiques descriptives unidimensionnelles

Réalisé par :

Exemple de Rapport

Année universitaire : ----/----

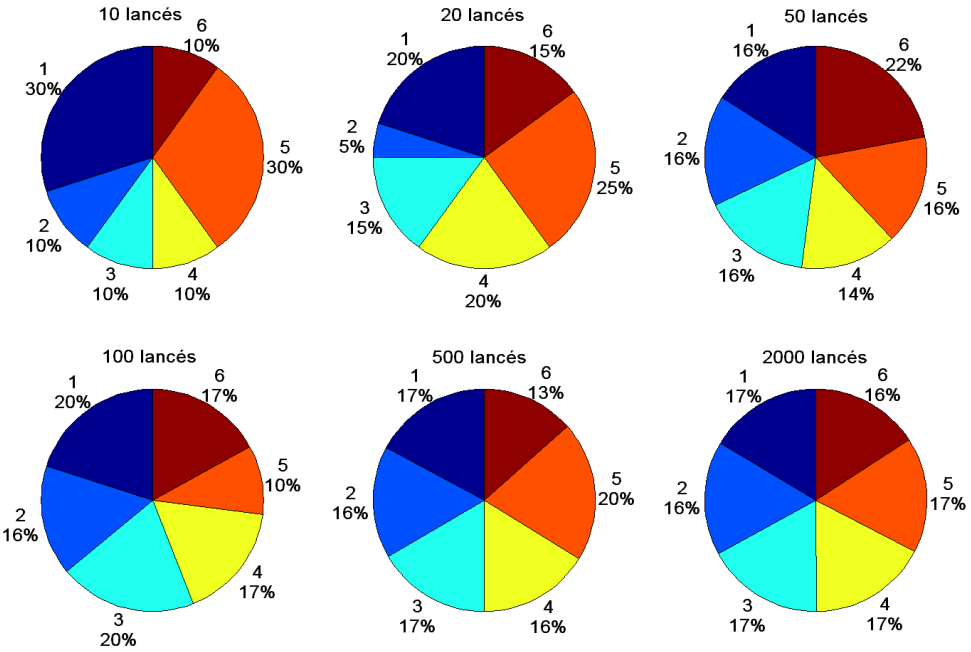
Objectif :

Etude de données statistiques unidimensionnelles à l'aide d'indicateurs élémentaires et de représentations graphiques. Le but des exercices proposés est d'acquérir une première compréhension de l'information véhiculée par les données à travers ces indicateurs et représentations.

1) Lancés d'un dé :

Nous avons effectué plusieurs expériences de lancés d'un dé avec un nombre de lancés différents à chaque fois. Les figures ci-dessous rendent compte des résultats obtenus selon le nombre de lancés.

♦) La première figure montre, les camemberts représentant la fréquence d'apparition de chacun des chiffres selon le nombre de lancés qui est indiqué pour au dessus de chaque camembert. En regard de chaque portion, nous avons indiqué 2 nombres : Le 1er correspond au chiffre tiré, qui représente en quelque sorte une classe, qui va donc de 1 à 6 dans le sens inverse des aiguilles d'une montre. Le second, est le pourcentage (arrondis) de l'effectif qu'il représente ; et qui est donc proportionnel à la surface de la part.



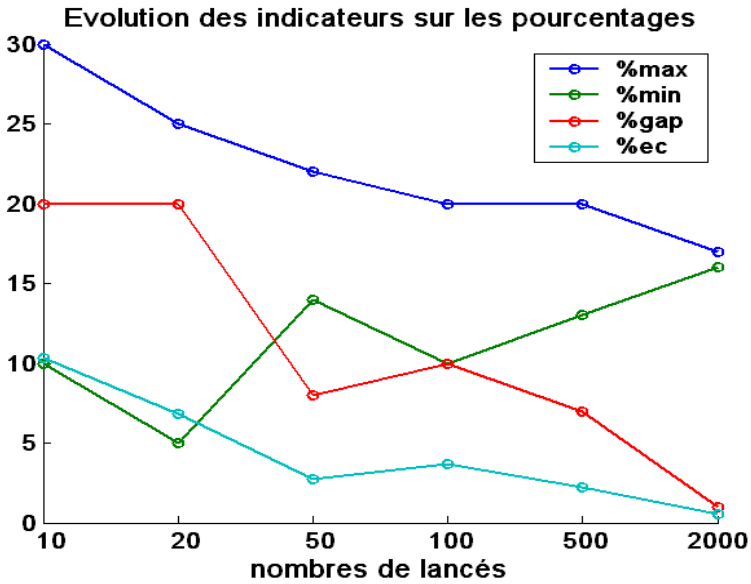
Etant donné qu'un lancé de dés suit une loi uniforme, on s'attend à ce que la répartition des effectifs entre les classes s'équilibre peu à peu avec l'augmentation du nombre de lancés. C'est effectivement ce que nous pouvons constater en menant différentes expériences avec 10, 20, 50, 100, 500 et enfin 2000 lancés :

Avec moins de 20 lancés les disparités entre classes sont bien marquées. Par exemple, pour le cas des 10 lancés, les classes 1 et 5 représentent chacune 30% de l'effectif alors que ceux des autres (classes 2, 3, 4 et 6) ne n'établissent qu'à 10%.

A partir de 50 ou 100 lancés, on voit qu'une répartition plus équitable commence à s'établir mais il subsiste encore des irrégularités, puisque par exemple la classe 5 de l'expérience à 100 lancés ne recueille que 10% alors que la moyenne théorique est de 100/6=16.67%.

Avec 500 lancés un équilibre entre les classes est atteint, Mais c'est lors de l'expérience à 2000 lancés qu'on a constaté une répartition entre classe vraiment homogène.

♦) La figure suivante fait état de 4 indicateurs qui ont été calculés pour chaque expérience. Le 1er (%max en, bleu) et le second (%min en vert) sont respectivement les pourcentages maximums et minimums obtenus. La courbe rouge est un peu redondante puisqu'elle correspond à la différence des 2 premières (%max-%min). La dernière (%ec en cyan) est l'écart-type des pourcentages.



Les informations apportées par cette figure recoupent celle de la figure précédente en apportant une vision différente des résultats. Les deux premiers indicateurs convergent vers l'équilibre obtenu pour 1/6 ; les deux autres vers 0 qui est une conséquence directe de l'équilibre atteint. L'expérience à 2000 lancés nous permet en particulier de mieux apprécier ce phénomène de la convergence des courbes qui se produit pour les grands nombres.

♦) Nous pouvons vérifier par ailleurs que la moyenne empirique des valeurs tirées tend vers l'espérance lorsque n augmente. S'agissant d'un lancé de dés, les expériences ayant été simulées selon la loi uniforme, on sait que pour cet exemple, l'espérance est de 3.5 $\left(\sum_{i=1}^{n} p_i x_i = \sum_{i=1}^{6} \frac{1}{6} i\right)$.

Les moyennes empiriques pour les différentes expériences, s'établissent ainsi :

Pour n = :	10	20	50	100	500	2000
Moyenne :	3.300	3.700	3.640	3.320	3.466	3.489

L'espérance est d'autant mieux approchée par la moyenne que n est grand.

Cette mise en pratique expérimentale montre l'importance de disposer d'un jeu de données sinon de grande taille, mais pour le moins de taille assez importante pour que des phénomènes sous-jacents représentés par un échantillonnage statistiques puissent être interprété avec une confiance suffisante.

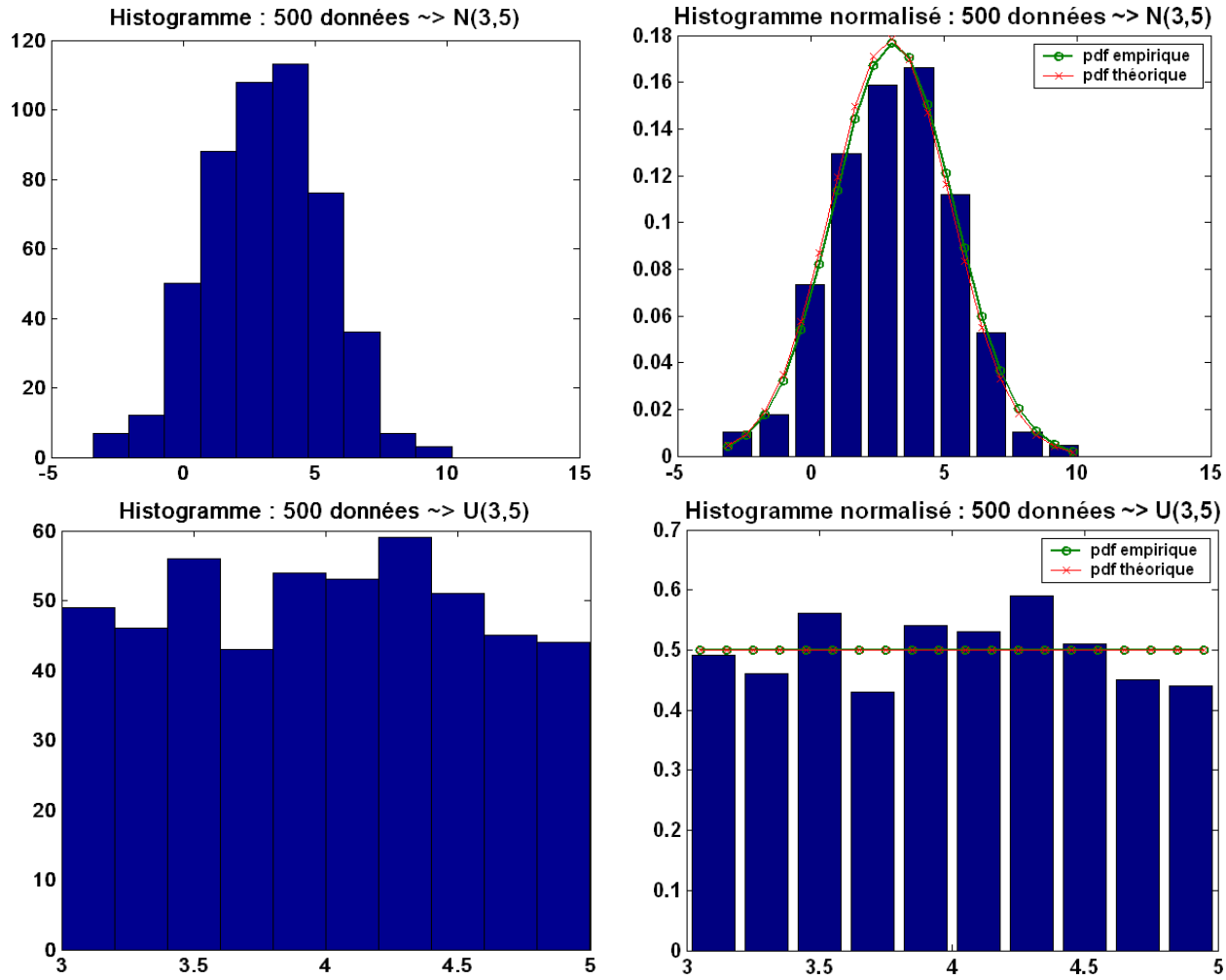
2) histogramme normalisé et fonction de densité

2.1) Données unidimensionnelles simulées

Un histogramme normalisé représente une distribution de probabilité empirique telle que la surface totale soit égale à 1. Le rectangle représente chaque classe par une surface proportionnelle à la fréquence de la classe. Un histogramme ainsi construit peut aisément être comparé à des courbes de fonction de densité théorique ou empirique.

Nous avons généré deux ensembles de 500 données simulées, l'un selon une loi normale N(3,5), l'autre selon une loi uniforme U(3,5). Pour le 1er ensemble on obtient une moyenne empirique de 3.113 et une variance empirique de 5.087, pour le 2ème, le minimum est à 3.001 et le maximum 4.999. On a donc une bonne adéquation entre les valeurs empiriques et théoriques.

Les figures ci-dessus représentent les histogrammes de ces données. Sur les histogrammes normalisés (ceux de droite), nous avons ajouté les tracés de la fonction de densité associée. Pour la première ligne, il s'agit des fonctions de densité obtenues à partir des équations de la loi normale N(3,5) (courbe en rouge) et celle obtenue en utilisant l'échantillon avec comme paramètres ceux obtenus par maximum de vraisemblance N(3,113, 5,087) (courbe verte). Même chose pour la deuxième ligne et la loi uniforme : La courbe rouge correspond à la loi U(3,5) celle en vert, à la loi U(3.001, 4.999).



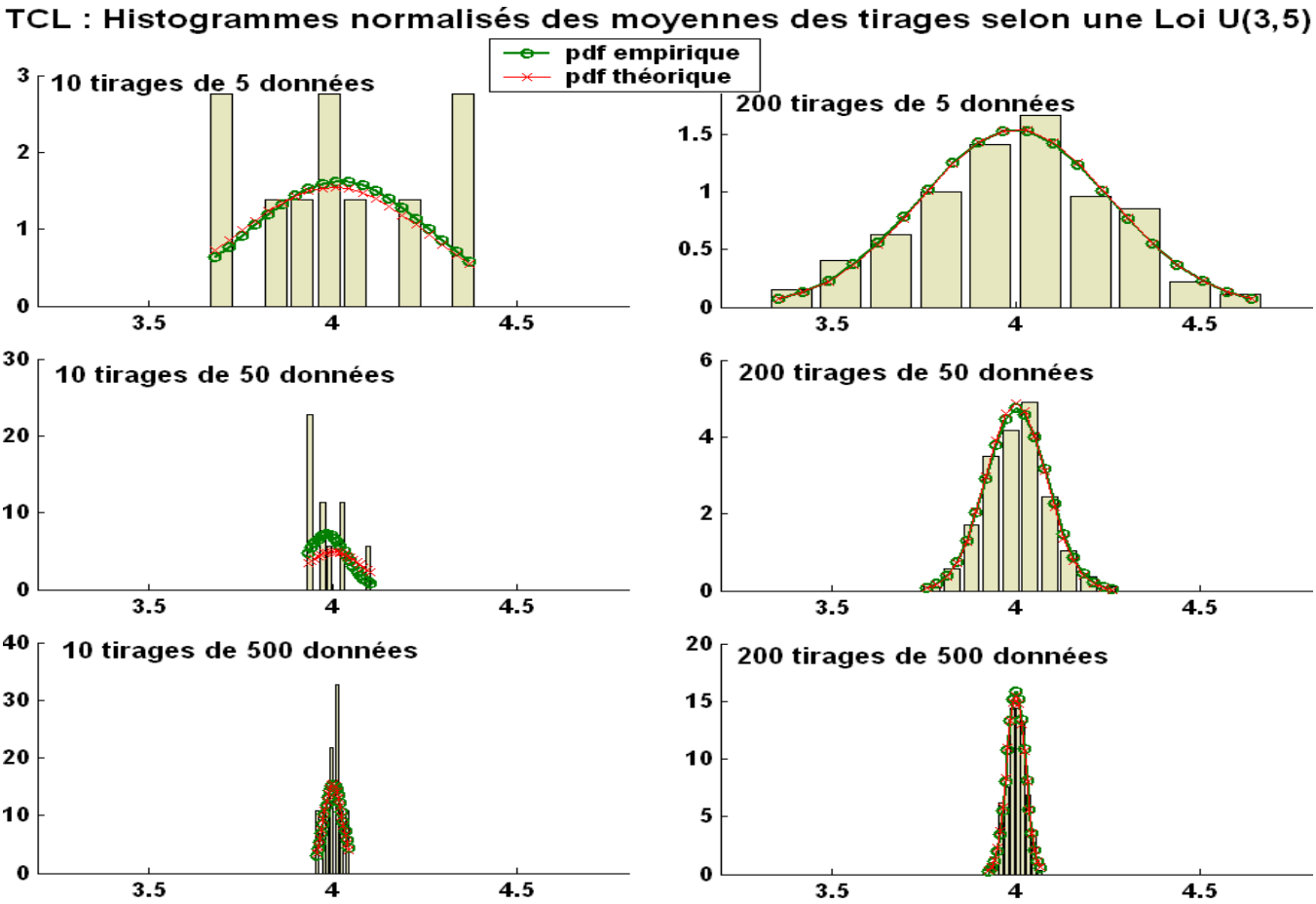
Nous pouvons remarquer que la forme de la distribution d'un histogramme normalisé n'est pas modifiée par rapport à l'histogramme des données de départ.

Dans cet exemple, les deux courbes (théorique et empirique) sont quasiment confondues. C'est la taille de l'échantillon qui régit l'ajustement des courbes entre elles et aux histogrammes.

2.2) Illustration du théorème de la limite centrale (TCL)

♦) Soit $\overline{X}$ la moyenne d'une suite de n variables aléatoires (va) indépendantes et identiquement distribuées d'espérance μ et de variance σ^2 finie, Le TCL (Theorem Central Limit ou théorème de la limite centrale) énonce que, lorsque n tend vers l'infini, $\overline{X}$ tend vers une loi $N(\mu, \sigma^2/n)$.

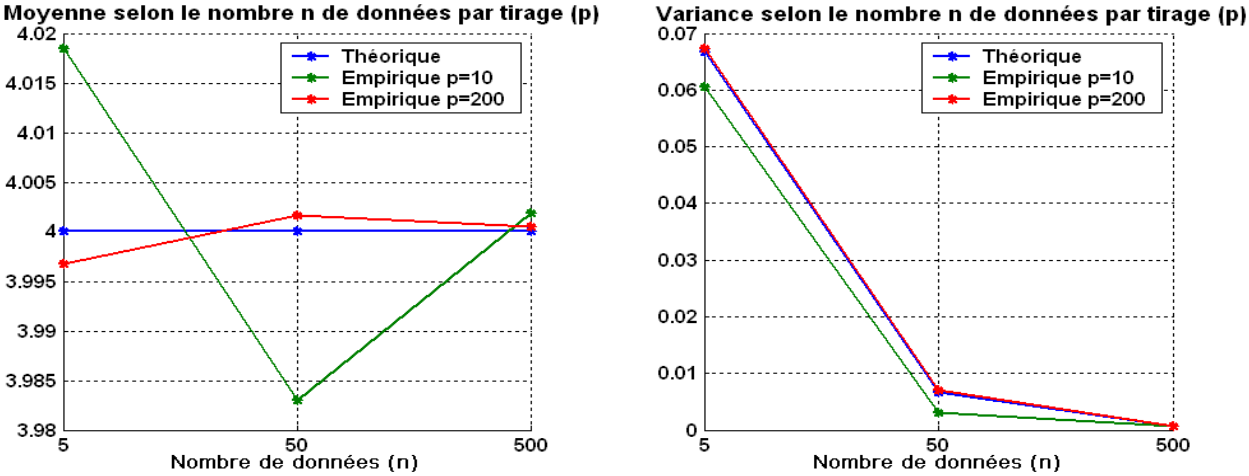
Les histogrammes normalisés de la figure ci-dessous illustre ce théorème. Ces histogrammes sont construis à partir des moyennes de différents tirages de n données simulées selon une loi uniforme U(3,5). Ce sont ces moyennes qui sont des instanciations de la variable aléatoire $\overline{X}$ évoquée ci-dessus. Les histogrammes de gauche correspondent à 10 échantillons, ceux de droite à 200. Pour chacune des lignes (d'histogramme) nous avons calculé la moyenne des échantillons pour n égal 5 puis 50 et 500 données simulées.



On remarque que lorsque le nombre n de la taille des tirages augmente (de haut vers le bas), l'écart type diminue. Ceci quelque soit le nombre p de tirages. Ce comportement était prévisible puisque, comme l'indique le TCL, la valeur de l'écart type ($\sigma/\sqrt{n}$) décroit avec l'augmentation de n puisque σ de sont coté reste constant.

On remarque que si le nombre p de tirage n'est pas assez grand, il est impossible de voir la forme de la distribution de la variable $\overline{X}$. La colonne de droite qui utilise déjà 200 tirages permet de mieux faire apparaître la courbe en cloche caractéristique de la loi normale.

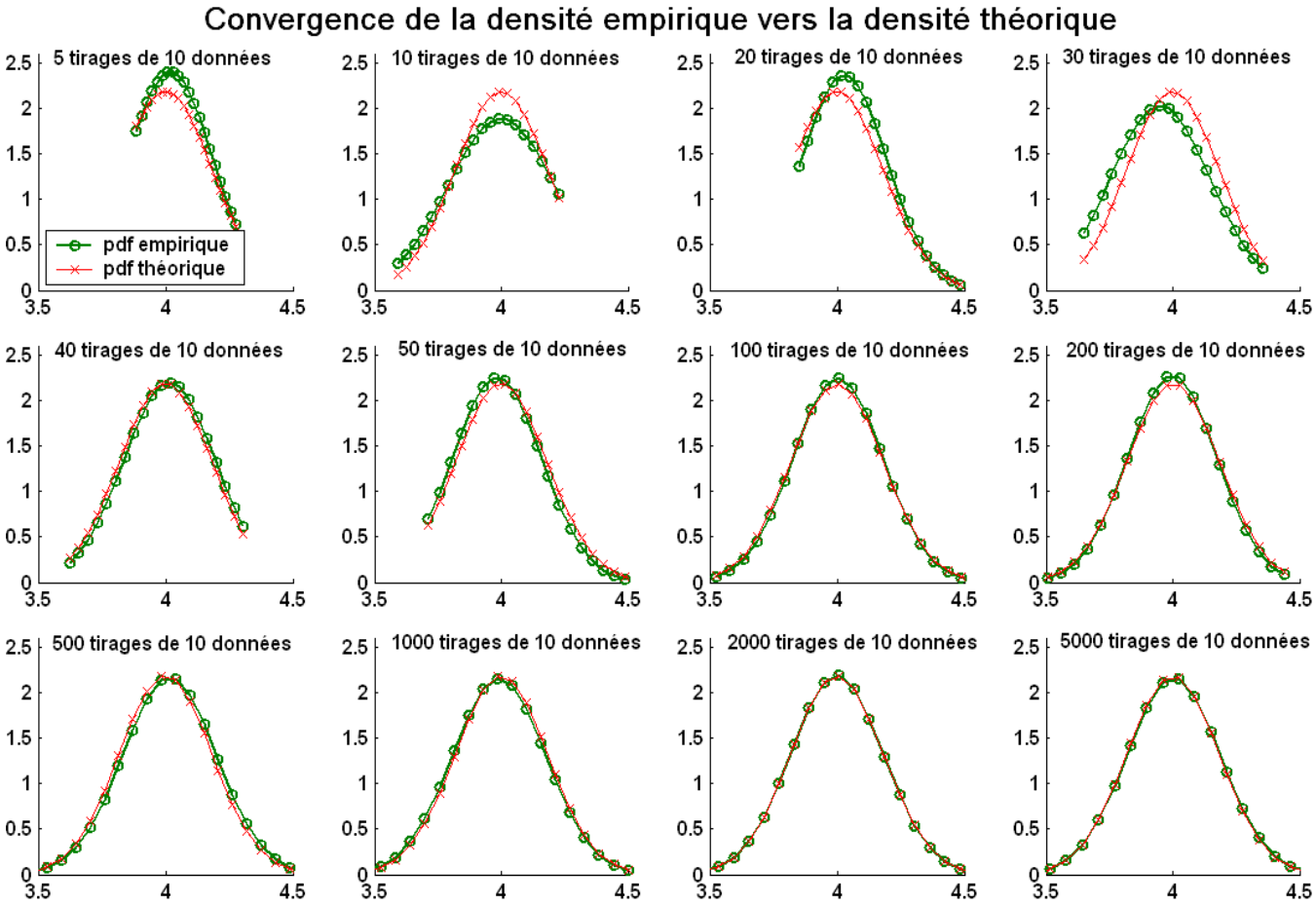
♦) La figure qui suit, montre, pour la variable aléatoire normale, dans les 2 cas de tirage utilisés (p=10 et p=200) et en fonction de la taille n des données, les moyennes ainsi que la moyenne théorique sur le repère de gauche ; les variances et la variance théorique sur celui de droite.



Ces 2 figures nous permettent de mieux visualiser la correspondance entre ces indicateurs empiriques et théoriques selon n et p. Dans les 2 cas (moyenne et variance) les indicateurs empiriques se rapprochent d'autant plus près des indicateurs théoriques que n est grand. On voit aussi que cette convergence est d'autant plus rapide que p est grand. On remarque d'ailleurs, s'agissant de la variance, que pour p=200, celle-ci est d'emblée bien approximée même avec n petit. On observe de nouveau, que, comme on l'a vu précédemment en évoquant le TCL, la variance décroit et tend vers 0 lorsque n augmente.

2.3) Convergence de la densité empirique vers la densité théorique

Nous avons mis à l'épreuve la convergence de la densité empirique vers la densité théorique lorsque le nombre des tirages aléatoires (selon une loi indépendante et identiquement distribuée (iid)) augmente. Comme pour la question précédente, nous avons utilisé la moyenne de tirages selon une loi uniforme pour construire des variables aléatoires normales. Nous avons procédé à 12 expériences en augmentant à chaque fois le nombre **p** de tirages. Quelques soient les tirages, ils ont tous été réalisés avec le même nombre n = 10 de données. La figure qui suit montre, pour chaque expérience, dont le nombre p de tirages est mentionné, les courbes de densité théorique (en rouge) et empirique (en vert) obtenues.



Il apparait de façon assez évidente que la densité empirique se confond de mieux en mieux avec la densité théorique lorsque p augmente. Lorsque p est petit (5, 10, 20, 30), la courbe empirique, ne parvient pas à se former complètement sur l'étendue de la courbe en cloche caractéristique de la loi normale. Après p=30, la courbe empirique s'ajuste de mieux en mieux ; avec p=100 et au-delà, elle s'ajuste très bien à la courbe théorique, même si l'on décèle encore visuellement de légers écarts. Les deux courbes se confondent quasiment avec p=2000.

3) Statistiques descriptives (sur des données réelles)

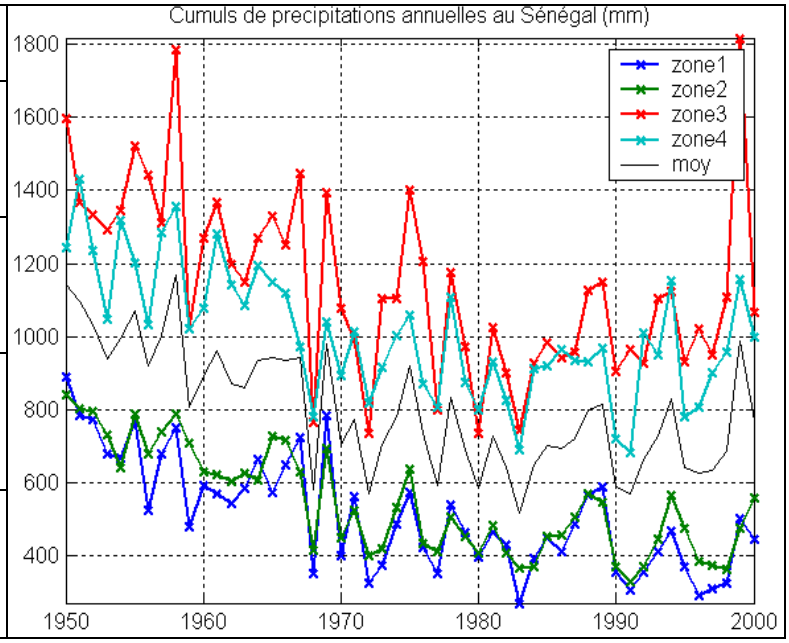
Nous nous intéressons maintenant à une étude de la pluviométrie au Sénégal. Pour cela nous disposons de données environnementales réelles qui ont permis d'établir des cumuls annuels de pluviométrie sur 4 zones géographiques du pays pour 51 années de 1950 à 2000. Ces 4 zones sont identifiées par un indice de 1 à 4 associés au nord-ouest, nord-est, sud ouest et sud-est.

3.1) Indicateurs statistiques élémentaires

Nous avons tout d'abord procédé à des calculs d'indicateurs statistiques élémentaires sur ces données. Les valeurs en sont présentées dans le tableau de gauche ci-dessous.

zone	moyenne	écart type	min	max	étendue
1	513	153	271	889	618
2	546	144	331	840	510
3	1146	250	735	1817	1083
4	1008	176	682	1432	750

(valeurs arrondies)



A la droite du tableau nous montrons une figure qui représente les courbes de pluviométrie (en mm) de chaque zone sur la période. Les valeurs numériques du tableau sont des informations synthétiques et précises des données étudiées. En cela, l'approche par des valeurs numériques de l'étude d'un phénomène est incontournable. Cependant, la représentation graphique par des courbes nous apporte une visualisation qui en facilite la compréhension et l'interprétation. Ces 2 approches sont complémentaires, les chiffres du tableau sont une synthèse des données illustrées par la représentation graphique.

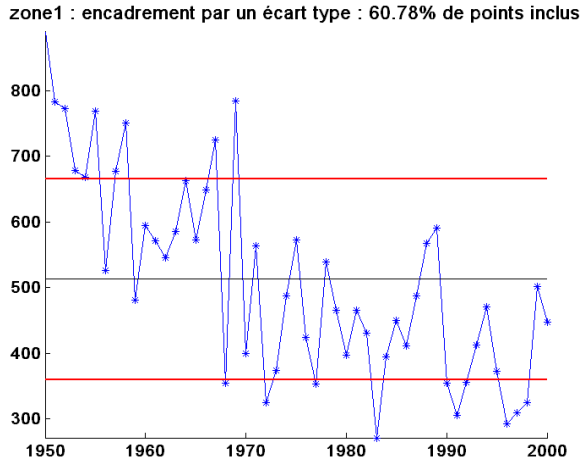
Globalement, on remarque une tendance à la décroissance de la pluviométrie des années 1950 aux années 1980 environ, années à partir desquelles elle semble s'être stabilisée (en faisant abstraction d'un pic, peut être exceptionnel, sur la zone 3 pour l'année 1999).

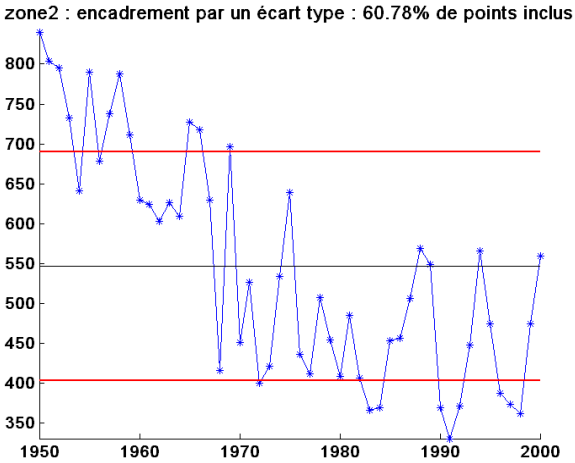
On voit également une différence nette de la pluviométrique entre les zones du nord et celles du sud qui est en moyenne 2 fois plus élevée.

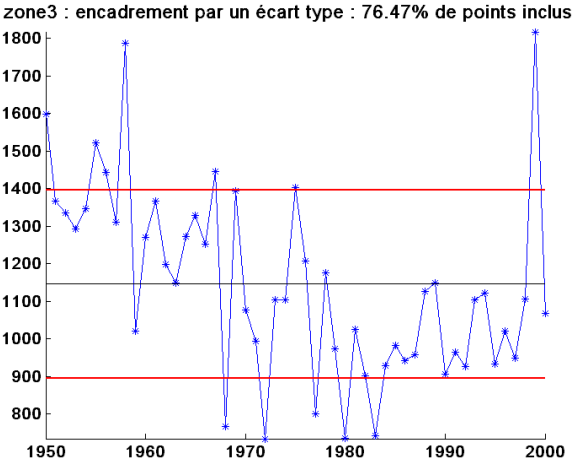
La colonne de l'écart type attire notre attention sur le fait que la zone 3 est le lieu d'un régime de pluie de qui connait une plus grande variabilité que les autres zones. Cela est d'ailleurs visible quand on regarde l'allure de la courbe rouge.

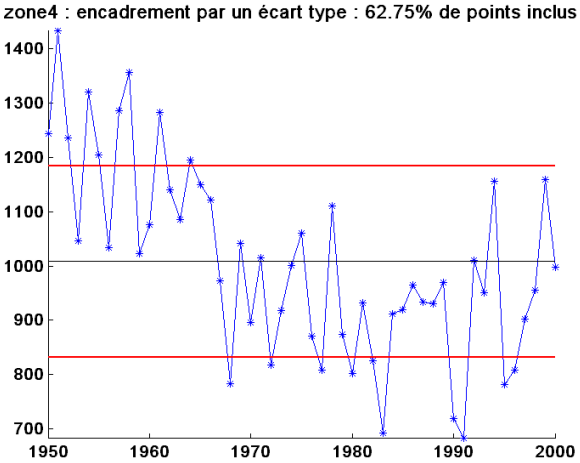
3.2) Encadrement à un écart type

Nous avons représenté, dans les figures ci-dessous et pour chaque zone, l'encadrement des données à plus ou moins un écart-type de la moyenne. Cet encadrement est matérialisé par les lignes rouges. Nous avons également tracé une ligne grisée pour situer la valeur moyenne.









Si l'on calcule le pourcentage de point inclus dans l'encadrement on trouve, dans l'ordre des zones, 60.78%, 60,78%, 76.47% et 62.75%. Ce sont ces valeurs arrondies qui sont mentionnées sur chacune des figures

Nous pouvons décider d'utiliser l'encadrement réalisé pour déterminer 3 régimes (ou classes grossières) de précipitation qui seraient un régime moyen, élevé ou faible selon que le niveau de précipitation se situe dans l'encadrement, au-delà ou en deçà d'un écart type à la moyenne.

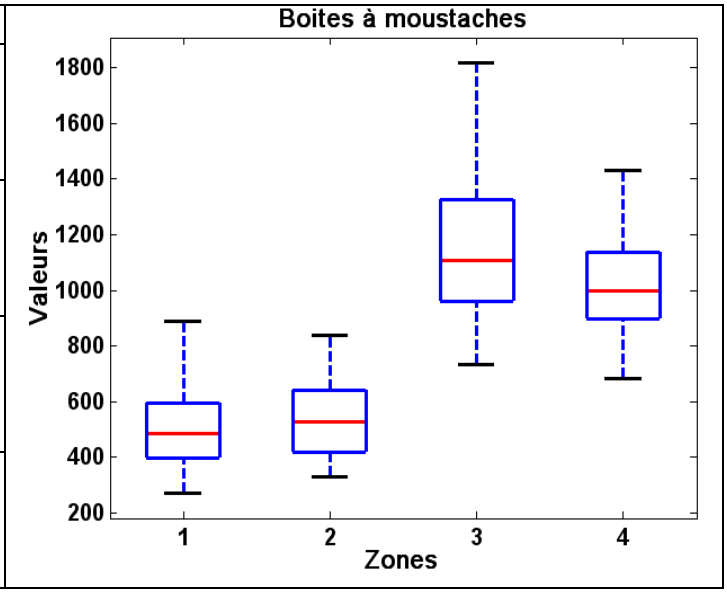
3.3) Quartiles

Nous donnons dans le tableau de gauche ci-dessous, et pour les 4 zones, les quartiles qui sont des indicateurs de tendance centrale ou position. On rappelle que 25% des données ont une valeur inférieure à Q1 (=>75% sont supérieures), 25% des données ont une valeur supérieure à Q3 (=>75% sont inférieures), et Q2 est la médiane qui partitionne en 2 parts égales l'ensemble des données.

On peut comparer les médianes et les valeurs moyennes (qui sont dans l'ordre des zones : 513, 546, 1146 et 1008). Les valeurs moyennes sont pour les quatre zones un peu plus élevées que les médianes, indiquant une distribution un peu décalée vers la droite. On a également indiqué l'écart (ou intervalle) interquartile qui est

égal à Q3-Q1. A l'instar de l'écart type ou de l'étendue, c'est un indicateur qui permet de rendre compte de la dispersion.

zone	Q1	Q2	Q3	Q3-Q1
1	396	487	593	198
2	417	526	641	224
3	959	1107	1324	365
4	897	998	1136	238



(valeurs arrondies)

La figure de droite est une représentation graphique de ces mêmes quartiles appelée « boîtes à moustache ». Le bord inférieur d'une boîte correspond à Q1, le bord supérieur à Q3 et le trait rouge à la médiane Q2. On constate de nouveau les différences entre les régions du nord (zone 1 et 2) et celles du sud (zone 3 et 4) ; ces dernières présentant un niveau de pluviométrie plus élevé, y figure aussi le minimum et le maximum qui font apparaître les valeurs extrêmes.

Grâce aux boîtes à moustaches, on peut de plus remarquer que les zones du nord ont une répartition des données plus équilibrée autour de la médiane, en particulier pour celles qui tombent dans l'intervalle interquartile. Les zones du sud sont plutôt plus ou moins décalées de ce point de vu.