INTRODUCTION AU COURS DE STATISTIQUE

”Dans ce pays, une loi fixe à l'avance combien de personnes se marieront et à quel âge, combien de jeunes filles épouseront des hommes âgés, combien de jeunes hommes des femmes âgées, pour combien de couples la différence d'âge sera de tel ordre, pour combien elle sera de tel autre ordre, combien de veufs épouseront des veuves, combien de divorces seront prononcés par les tribunaux, etc....

Ce qui ne pourrait jamais être réalisé de cette manière par la volonté et par la force des hommes, s'accomplit merveilleusement, sans aucune intervention, grâce à l'organisation naturelle de la société humaine.....

En examinant les mariages, les suicides, les crimes et en dégageant leurs lois, nous pouvons prévoir avec une très grande exactitude combien de mariages, de divorces, de suicides et de crimes se produiront pendant une année et comment ils se répartiront.

L'examen futur des résultats de cette année révélera qu'ils sont tout aussi conformes aux prévisions que si nous nous trouvions dans ce Etat imaginaire. Le plus étonnant est que nous soyons nous-mêmes les éléments d'un grand mécanisme et que nous gardions l'entière liberté de nos mouvements sans pour autant empêcher le fonctionnement préétabli de ce mécanisme.” Adolph Wagner 1864

A l'origine, la statistique consistait en une simple collecte de chiffres, ce qui correspond à la signification première de state-istique, ensemble d'informations concernant la population et l'économie, indispensable à l'Etat. La tâche initiale de la statistique a été, comme le dit Cournot : ” le recueil des faits auxquels donne lieu l'agglomération des hommes en sociétés politiques ”.

Actuellement, la statistique s'est développée comme une méthode scientifique d'analyse s'appliquant à l'économie et à l'ensemble des sciences sociales et de la nature.

L'assurance vie est la première grande application du calcul des probabilités. Avant la fin du 17ème siécle, Halley calcula les premières table de mortalité, qui lui permirent d'estimer la durée de vie. Se prémunir contre l'adversité a un prix depuis deux millénaires, ce prix étant calculé auparavant sans règle précise, en s'appuyant sur l'expérience. En 1637, avant la naissance de Halley, un savant crétois, du nom de Canopius, se prépare un breuvage que l'on goûte pour la première fois en angleterre : un café..Le succès est tel que s'ouvrent bientôt des centaines de ”coffee house”, dont l'un donnera naissance à la célèbre Lloyd's. Halley, quant à lui, utilisa l'outil de prédilection des assureurs, la loi des grands nombres, loi indispensable à l'exercice de l'activité d'assurance. Cette loi fait le lien entre les statistiques et les probabilités, puisqu'elle met en relation fréquence statistique et probabilité. On devra s'y résoudre ou s'en réjouir : le hasard est une notion qui ne s'analyse pas sans mathématiques...

Comme toute science, la Statistique a son vocabulaire propre et nécessite beaucoup de rigueur. On devra d'abord définir la population concernée par notre étude (par exemple l'ensemble des étudiants inscrits à ce cours au premier semestre 2009-2010, l'ensemble des villes de France de moins de 20000 habitants, les 131.2 millions de votants à l'election présidentielle américaine de 2008, etc.), puis ensuite le caractère ou la variable statistique étudié (âge, taille, nombre d'habitants, sexe, etc.).

I Statistique descriptive et statistique inférentielle

Le présent cours est essentiellement un cours de statistique descriptive, en ce sens que l'on va étudier les graphiques,les paramètres, les outils permettant de présenter et d'analyser l'information de façon synthétique et utilisable.

Cependant on voit aisément qu'en général on n'a pas accès à toute la population et qu'on doit travailler sur un échantillon. Si l'on veut savoir si la soupe est correctement salée, on ne mangera pas toute soupe contenue dans la soupière, mais on en goûtera une cuillère. La statistique inférentielle a pour but de généraliser cette information basée sur des échantillons, à la population complète, ou selon l'expression consacrée de faire des inférences, par exemple d'estimer le pourcentage d'électeurs prêts à voter pour tel candidat, à partir de l'étude d'un échantillon de 1000 électeurs.

Evidemment si la soupe est mal mélangée, cet échantillon ne sera pas représentatif de la population (échantillon biaisé) et ne donnera pas une idée exacte quant à la salaison de la soupe. Le prélèvement d'échantillons, que nous ne traiterons pas ici, doit être effectué suivant des méthodes rigoureuses assurant l'obtention d'un échantillon représentatif. Si l'on s'intérresse aux notes d'un partiel dans un EC donné, on ne peut prendre comme échantillon les étudiants du premier rang, ni ceux du dernier rang, ni ceux dont le nom commence par un A, tous ces échantillons sont biaisés, contrairement au mode le d'échantillonnage le plus important, le mode aléatoire simple, qui consiste à effectuer des tirages aléatoires et indépendants.

En plus du mode d'échantillonnage, il est important de s'intérresser à un élément important : la taille de l'échantillon. On retiendra le résultat, à priori surprenant : c'est la taille de l'échantillon qui conditionne la qualité des résultats, et ce sans rapport avec la taille de la population totale. Avec un échantillon représentatif de 100 étudiants parmi les 21487 étudiants inscrits à l'université Paris 8 en 2007-2008 on représente la population étudiante de l'université avec une qualité équivalente à celle obtenue par un échantillon représentatif de 100 votants parmi les 131.2 millions de votants à l'election présidentielle américaine de 2008.

II EXERCICE-1

La liste suivante est composée de prénoms d'un groupe d'étudiants suivis entre parenthèses du nombre de livres lus pendant le mois :

Pierre (3), Paul (2), Jacques (2), Ralph (3), Abdel (1), Sidonie (2), Henri (0), Paulette (1), Farida (2), Laure (2), Kevin (0), Carole (3), Marie-Claire (0), Jeanine (3), Julie (2), Ernest (3), Cindy (3), Vanessa (2), José (1), Aurélien (1).

1. Déterminer la population et le caractère étudiés.
2. Préciser la nature et les modalités du caractère.

3. A partir des données brutes, compléter le tableau statistique suivant représentatif de la distribution :

Modalités x_i	Effectifs n_i	Fréquences f_i	$n_i cc$	$n_i cd$
0				
1				
2				
3				

4. Représenter la distribution par un diagramme en bâtons.
5. Représenter la distribution par un secteur circulaire.
6. Calculer le nombre moyen de livres lus par les étudiants de ce groupe.
7. Calculer les effectifs cumulés croissants et décroissants.
8. Calculer les fréquences cumulées croissantes et décroissantes.
9. Combien d'étudiants ont lu au moins 1 livre ? au plus 2 livres ?

III EXERCICE-2

Les données ci-contre correspondent aux distances parcourues, en milliers de km, avant la première panne importante d'une flotte automobile de 50 véhicules.

Milliers de km	effectif
[40-50[	1
[50-60[	2
[60-70[	2
[70-80[	3
[80-90[	4
[90-100[	4
[100-110[	6
[110-120[	9
[120-130[	7
[130-140[	5
[140-150[	4
[150-160[	3
Total	50

1. Déterminer la population et le caractère étudiés.
2. Préciser la nature du caractère.
3. Elaborer un tableau statistique qui sera complété au fur et à mesure des questions.
4. Calculer les centres de classes, placer les dans le tableau statistique et calculer la moyenne de la série.
5. Calculer les effectifs cumulés croissants et décroissants.
6. Combien de voitures ont parcouru au moins 130 000 km avant la première panne? au plus 90 000 km ?
7. Tracer l'histogramme des effectifs.

IV EXERCICE-3 :

Le tableau ci-dessous indique le temps mis par 200 fleurs pour s'ouvrir

temps (en min)	[0 ; 6[	[6 ; 10[	[10 ; 12[	[12 ; 16[	[16 ; 20[	[20 ; 24[
effectifs n_i	15	40	45	60	30	10

1. Déterminer la population, le caractère et sa nature.
2. Représenter cette série statistique par un histogramme et déterminer son mode. Interpréter votre résultat.
3. Calculer la moyenne de cette série statistique.

a. Calculer la médiane de cette série statistique. Interpréter votre résultat.

b. Représenter la médiane sur l'histogramme. Expliquez comment retrouver graphiquement la signification de la médiane.

V Dictionnaire (premières notions)

1. Une **population** est l'ensemble des éléments auquels se rapportent les données étudiées (étudiants d'une université, habitants d'un pays, entreprises d'un secteur...).
2. Dans une population donnée, chaque élément est appelé un "**individu** " ou une "unité statistique".
3. En fait, la collecte d'informations sur une population est rarement effectuée de façon exhaustive (enquête sur la totalité des individus) ; on a souvent recours à des enquêtes par sondage qui portent sur une partie de la population, appelée **échantillon**.
4. Il existe deux types de caractères statistiques : les caractères **quantitatifs**, c'est à dire qui prennent des valeurs numériques (taille, salaire, etc...) et les caractères **qualitatifs** (sexe, métier, couleur des yeux, situation matimoniale, etc..), ceux dont les modalités ne sont pas numériques.
5. **Caractère quantitatif discret** : qui prend des valeurs isolées. Exemple : nombre d'enfants; les valeurs sont des entiers naturels.
6. **Caractère quantitatif continu :** prend des valeurs quelconques dans un intervalle ; les données sont regroupées en classes ; exemple : salaires, taille ...
7. **Modalités** : les modalités d'un caractère sont ses différentes "valeurs" ; exemple : l'état matrimonial comporte souvent cinq modalités : célibataire, marié, pacsé, veuf, divorcé.
8. **Effectif :** l'effectif d'une modalité, en général noté n_i, représente le nombre d'individus correspondant à cette modalité. En général, on note $N = \sum n_i$, l'effectif total de l'échantillon étudié.($\sum$: lu "Sigma" et signifie somme) .
9. **Effectifs cumulés croissants : notation :** $n_i cc$. On effectue la somme des effectifs des modalités inférieures ou égale à une modalité donnée ; exemple des notes à un devoir : si l'on suppose que les notes sont des entiers de 1 à 10, l'effectif cumulé croissant correspondant à 8, consiste à compter le nombre de personnes ayant une note inférieure ou égale à 8.(cf exemple)
10. **Effectifs cumulés décroissants** : **notation :** $n_i cd$. On effectue la somme des effectifs des modalités supérieures ou égales à une modalité donnée; pour le même exemple que précédemment, on trouvera pour 8, le nombre total de personnes dont la note est supérieure ou égale à 8.
11. **Fréquence (relative) :** la fréquence d'une modalité, notée f_i, est donnée par: $f_i = \dfrac{n_i}{N}$; elle représente la proportion d'individus se rapportant à une modalité par rapport à l'effectif total. On a :

    $\boxed{0 \leq f_i \leq 1 \text{ et } \sum f_i = 1}$; enfin une fréquence peut être donnée en pourcentage et alors, la somme des fréquences donne 100%.
12. **Fréquences cumulées croissantes ou décroissantes : notations :** $f_i cc$ et $f_i cd$

    $f_i cc = \dfrac{n_i cc}{N}$ et $f_i cd = \dfrac{n_i cd}{N}$.
13. **Amplitude de classe** : l'amplitude de la classe $[a\,;b[$ est $b-a$, c'est la longueur de l'intervalle.
14. **Centre de classe :** C'est le milieu de l'intervalle, donc le centre de $[a\,;b[$, soit,la moyenne arithmétique x de a et b : $x = \dfrac{a+b}{2}$.
15. **Densité de classe** (continu) : on appelle densité de la classe $[a_i\,;b_i[$, le nombre d_i défini par : $d_i = \dfrac{n_i}{b_i - a_i}$, qui représente le nombre d'individus par unité de classe.
16. **Mode**
    a. Caractère discret : un mode est une valeur du caractère ayant l'effectif maximal ; on notera qu'une série statistique peut avoir un mode (unimodale) ou des modes (plurimodale). Dans l'exercice 1 le mode est 2.

    b. Caractère continu : une classe modale est une classe ayant la densité maximale. Dans l'exemple 2, il s'agit de la classe $[110\,;120[$ et dans l'exercice 3 de la classe $[10\,;12[$.

17. **Médiane**
La médiane est à priori la valeur du caractère qui partage la série ordonnée (de la plus petite modalité à la plus grande) en deux groupes de même effectif ; on doit cependant distinguer plusieurs situations. On utilisera les effectifs ou fréquences cumulées croissantes.
a. Caractère discret
   i. Nombre d'observations impair ($n = 2p+1$) : alors il y a un terme central, c'est la médiane. Si par exemple la série comporte 201 observations, la médiane est la 101^{eme} observation de la série ordonnée.
   ii. Nombre d'observations pair : il n'y a pas de terme central, donc à priori pas de médiane ; par convention on prend l'intervalle médian constitué par les deux valeurs centrales et la médiane est la moyenne arithmétique de ces deux valeurs. Si par exemple la série comporte 100 observations, aprés l'avoir ordonné, on prend la 50^{eme} et la 51^{eme} valeur et la médiane est leur moyenne arithmétique.

b. Caractère continu
On procède par interpolation linéaire, pour déterminer la valeur qui correspond à une fréquence cumulée croissante de 50%.

18. **Quartiles**
Il y a trois quartiles : Q_1, Q_2,et Q_3 ; le principe est le même que pour la médiane, mais il s'agit de partager la série en quatre groupes comprenant 25% de la population. Q_1 est la plus petite valeur telle qu'il y ait au moins 25% des valeurs de la série inférieures ou égales à Q_1. Q_2 est la médiane et Q_3 est la plus petite valeur telle qu'il y ait au moins 75% des valeurs de la série inférieures ou égales à Q_3. On distingue comme pour la médiane le cas discret et le cas continu et on utilise les effectifs ou fréquences cumulées croissantes.
Boîte à moustaches ou Box plot : ce graphique sur lequel nous reviendrons utilise les quartiles et est très précieux pour comparer diverses séries (salaires dans différents pays européens par exemple).

19. **Moyenne**
La moyenne d'un caractère statistique quantitatif x est notée $\overline{x}$ et définie par :

$$\boxed{\overline{x}=\frac{1}{N}\sum n_i x_i = \sum f_i x_i}$$

.On notera que : $\sum n_i x_i = N\overline{x}$.

20. **Variance et écart-type**
Pour mesurer les fluctuations d'un caractère autour de sa moyenne, c'est-à dire fournir un indicateur de dispersion, on définit la variance et l'écart-type, définis respectivement par :

$$\boxed{V(x) = \frac{1}{N}\sum n_i(x_i-\overline{x})^2 = \frac{1}{N}\sum n_i x_i^2 - \overline{x}^2}$$

($MC - CM$: moyenne des carrés moins carré de la moyenne ; on note que $V(x) \geq 0$).
$\sigma(x) = \sqrt{V(x)}$.

21. **Covariance**
La covariance concerne les séries bivariées (deux caractères quantitatifs, par exemple salaire et âge) et mesure les fluctuations simultannées des deux caractères par rapport à leurs moyennes respectives. La formule de la covariance est en fait un dédoublement de celle de la variance, et nous donnerons la formule dans le cas particulier d'observations uniques (effectifs égaux à 1) :

$$\boxed{Cov(x;y) = \frac{1}{N}\sum(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{N}\sum x_i y_i - \overline{xy}}$$

($MP - PM$: moyenne des produits moins produit des moyennes ; on notera que contrairement à la variance, la covariance peut être négative).
A suivre.....