Statistique et traitement des données

1. Introduction aux statistiques

1.1. Généralités

A l'origine les statistiques concernaient tout ce qui avait trait aux affaires de l'État (ressources, populations, inventaires, etc...) et donc le mot statistique au pluriel a toujours ce sens actuellement.
La statistique est l'étude plus globale de tout ce qui permet de présenter, décrire, résumer des données, mais aussi de prédire des phénomènes (par exemple des faits sociaux) dont une étude exhaustive de tous les facteurs est impossible (trop grand nombre, trop grande complexité, inconnus,etc...)

L'ensemble sur lequel on étudie ces données est appelé une population statistique.
Un individu est élément individuel de cette population.
Un échantillon correspond à une partie de la population statistique.

Deux cas de figures se présentent:
-> on cherche à décrire au mieux un certain nombre de grandeurs caractéristiques de la population: c'est le rôle de la statistique descriptive.
-> ou au contraire on cherche à tirer des conclusions sur la population à partir d'un échantillon: on parle alors d'inférence statistique.

Chaque individu peut-être décrit par un ou des caractères statistiques.
Ces caractères peuvent être:

- quantitatifs: discrets (par exemple le nombre d'enfants) ou continus (ex: taille, âge, poids)
- qualitatifs: nominaux (sexe, profession, département) ou ordinaux (taille vestimentaire, préférence plus ou moins grande,...)

Les valeurs que peut prendre un caractère s'appellent les modalités du caractère.

En fonction de leur nature le traitement qui pourra en être fait sera plus ou moins poussé, à l'aide de tableaux, graphiques, paramètres-clés, etc...

On notera N la taille de la population et n la taille de l'échantillon.
On notera $X(,Y,Z)$ les variables afférentes à cette population, correspondant aux différents caractères statistiques.
On notera $x_1, x_2, ..., x_k$ les modalités possibles de X .
On notera $x_1, x_2, ..., x_n$, sans confusion possible, les modalités observées du caractère attaché à la variable X pour un échantillon donné. On dira alors que les $(x_i)_{1 \leqslant i \leqslant n}$ sont n observations de X. La donnée de n observations de X définit une série statistique.

1.2. Effectifs et fréquences

Définition : On se donne une série statistique $(x_i)_{1 \leqslant i \leqslant n}$.
On appelle effectif de la modalité x_i , que l'on note n_i le nombre d'observations de la modalité x_i .
On appelle fréquence relative d'apparition de la modalité x_i , notée f_i , le rapport $\frac{n_i}{n}$.

Remarques : 1. Une fréquence relative est une valeur comprise entre 0 et 1.
2. On peut également l'exprimer en pourcentage: $F_i = \frac{n_i}{n} \times 100$.

Remarque : Pour les distributions à caractères qualitatifs, on est limité aux outils précédents et l'intérêt de telles séries est donc limité. En pratique on ne s'intéressera désormais qu'aux distributions à variables quantitatives.

Effectifs et fréquences cumulées :
Pour calculer un effectif (ou une fréquence) cumulé(e), on commence par ordonner les valeurs de la série statistique par ordre croissant, puis on fait la somme depuis le départ.

Classes : Dans le cas d'une distribution avec un grand nombre de valeurs, celles-ci sont regroupées en classe.

Exemple :
Par exemple, voici le tableau donnant la répartition par tranches d'âge des femmes en France et son évolution:

Année	15-19 ans	20-49 ans	50-69 ans	> 70 ans	Total
1990	2 117 163	12 001 800	6 066 714	3 284 514	23 470 191
1999	1 941 673	12 663 523	6 146 597	4 043 415	24 795 208
2010	1 913 401	12 375 310	7 451 960	4 808 952	26 549 623
2020	1 844 486	11 982 593	8 362 367	5 192 976	27 382 422

Le regroupement s'est fait par classe, du type $[c_i; c_{i+1}[$.
Le centre d'une classe est noté m_i et se calcule par: $m_i = \frac{c_i + c_{i+1}}{2}$.
La largeur d'une classe est: $c_{i+1} - c_i$.
Les valeurs à l'intérieur d'une classe sont alors remplacées par la valeur du centre de la classe.

1.3. Représentations graphiques

Les deux types de diagrammes adaptés aux variables qualitatives sont:

- les diagrammes en bâtons
- les diagrammes circulaires (et semi-circulaires)

Dans un diagramme circulaire, l'angle α_i formé par un secteur est proportionnel à la fréquence d'apparition. Ainsi, on a:

$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$$

Pour les variables quantitatives, outre les représentations précédentes, on peut utiliser :

- les histogrammes :
  A noter que lorsque les rectangles représentant la classe sont d'aires proportionnelles aux effectifs de la classe. Aussi, si les classes n'ont pas la même amplitude, la hauteur correspond à la densité d'effectif : $\frac{\text{effectif de la classe}}{\text{amplitude de la classe}}$. Dans ce cas, on n'oubliera pas de donner l'unité d'aire en légende.
- les polygones d'effectifs cumulés ou de fréquences cumulées

1.4. Méthodologie d'étude d'une série statistique.

Afin d'étudier des distributions à variables quantitatives, on va tout d'abord s'intéresser à les décrire, puis ensuite à définir des grandeurs mathématiques permettant de les caractériser.

Si on veut avoir une première approche d'une distribution, les points suivants seront intéressants à regarder

- 1/ Où se situe le centre de la distribution ? Et tous les éléments qui concernent la position de la distribution.
- 2/ La distribution est-elle dispersée ou au contraire resserrée ? Et tous les éléments qui concernent la dispersion de la distribution.
- 3/ La distribution est-elle symétrique ou au contraire très antisymétrique ? Et tous les éléments qui concernent la dissymétrie.
- 4/ La distribution est-elle uni-modale ou bi- ou multi-modales ?

Il va s'agir d'essayer de quantifier ces différents points. Dans toute la suite, on notera: x_1, x_2,..., x_n des observations de la variable X.

2. Quantités permettant de mesurer la position

2.1. Mode, classe modale

Définition : La modalité la plus fréquente d'une série statistique est appelée le mode de la distribution. Dans le cas d'une série à valeurs continues, on parle de classe modale.

2.2. Moyenne arithmétique

Définition: La moyenne arithmétique d'une série statistique $(x_i)_{1\leqslant i\leqslant n}$ issue de n observations de la variable X est donnée par:

$$m(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n}.$$

En pratique, on utilise souvent la moyenne pondérée par les effectifs, lorsque la modalité x_i est présente n_i fois dans la série statistique.

$$m(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n}.$$

Remarque: Dans le cas de variables à valeurs continues, on remplace la valeur de la modalité par le centre de la classe. Donc plus grande seront les classes, plus grande sera l'erreur.

Propriété: Lorsqu'on dispose de deux séries statistiques d'une même variable X, de moyennes respectives $\bar{x_1}$ et $\bar{x_2}$, d'effectif respectif n_1 et n_2, alors la moyenne $\bar{x}$ de la série réunissant les deux séries est donnée par : $\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2}$.

2.3. Médiane.

Définition : La médiane de la d'une série statistique $(x_i)_{1\leqslant i\leqslant n}$ issue de n observations de la variable X est une valeur telle que si l'on ordonne les observations de X, celles-ci se retrouvent coupées en deux sous-ensembles d'observations de même effectif.
On la note med(X).

En pratique la médiane est la plus petite valeur telle que 50 % des valeurs de la série statistique lui sont inférieure.

Remarque: Médiane ou moyenne ?
Si la distribution est symétrique et que les valeurs sont resserrées, on aura: $m(X) \approx \text{med}(X)$
Si la distribution est symétrique et qu'il y a des valeurs erratiques, par exemple:
Série A : 10, 11, 12, 12, 12 , 13, 14 Série B : 10, 11, 12, 12, 12, 13, 1400
Les séries A et B ont la même médiane, mais des moyennes sensiblement différentes, et pourtant une seule valeur diffère.
La médiane résiste mieux aux effets de bords que ne fait la moyenne et est alors nettement plus significative.
Suivant les informations que l'on veut obtenir, la moyenne peut-être intéressante quand il s'agît ensuite de faire des totaux (exemple: nombre moyen d'une certaine maladie dans de multiples échantillons, pour prédire le nombre total de cas dans une population), ou au contraire la médiane lorsque l'on veut avoir une idée de comment est réparti une population.

2.4. Quantiles.

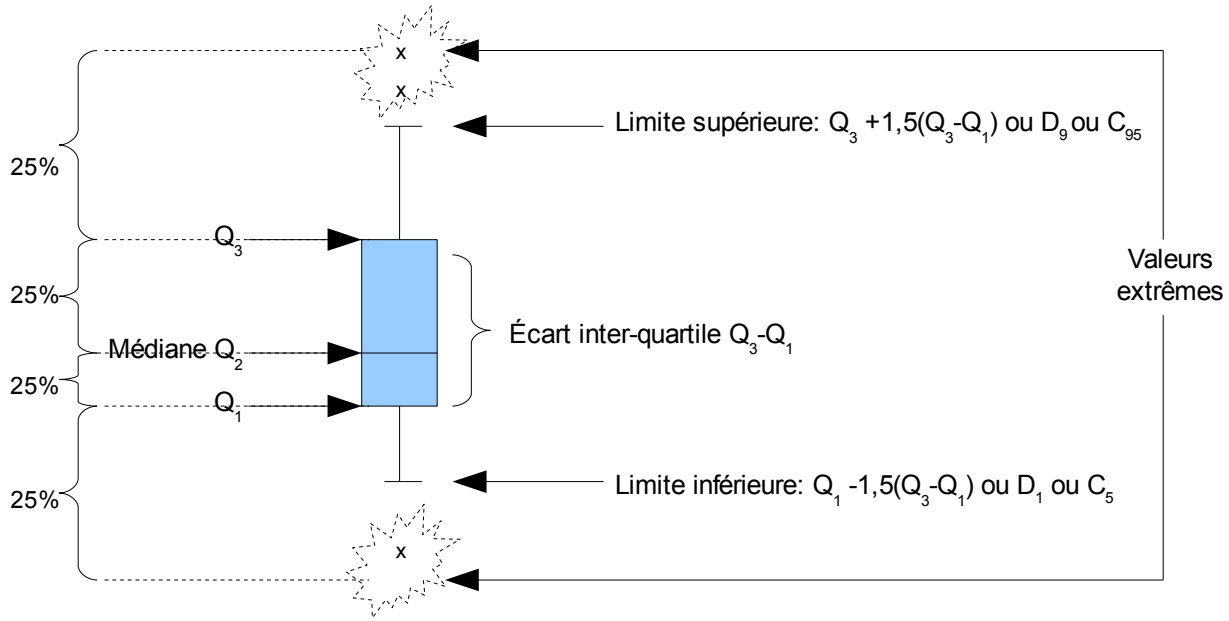
Pour préciser cette idée de répartition, on dispose d'autres outils:
On généralise l'idée de médiane, en utilisant la notion de quantile où l'on va partager la distribution en n sous distributions de même taille.
Les valeurs de n classiques sont:

- n=4 --> quartiles
- n=10 --> déciles
- n=100 --> centiles

Ainsi la médiane est le deuxième quartile, le cinquième décile, le 50-ième centile (ou percentile 50%), le quantile 0,5.
Le premier quartile est le percentile 25%, le troisième quartile est le percentile 75%.

En pratique, pour la recherche des quartiles, on adopte la définition suivante :
Le premier quartile Q_1 est la première observation de la série statistique ordonnée telle que 25 % des valeurs observées de la série statistique soit inférieure à cette valeur.
Le troisième quartile Q_3 est la première observation de la série statistique ordonnée telle que 75 % des valeurs observées de la série statistique soit inférieure à cette valeur.

Application : box-plot ou boîte à moustache ou diagramme de Tuckey



Intérêt : Comparer rapidement deux séries statistiques.

3. Quantités permettant de mesurer la dispersion

3.1. Mesures de dispersion issues des quartiles.

Dans une série statistique, on s'intéresse de savoir comment se situent les valeurs les unes par rapport aux autres.
Par exemple, les deux séries suivantes :
Série A : 10 11 12 13 14 15 16 Série B : 6 13 13 13 13 13 20
ont bien la même moyenne mais ne sont pas aussi dispersées l'une que l'autre.

L'espace inter-quartile $Q_3 - Q_1$ est une première mesure de dispersion qui prend en compte 50% des valeurs, celles qui sont centrales. Elle élimine les effets de bords.

Afin de s'affranchir de toute unité, on peut calculer l'espace inter-quartile relatif : $\frac{Q_3 - Q_1}{Q_2}$.

3.2. Mesures de dispersion issues d'une comparaison avec la moyenne.

On considère une série statistique constituée de n observations $x_1, ..., x_n$ de la variable X.
On note n_i le nombre d'observations correspondant à la modalité x_i.

3.1.1. Écart absolu moyen

On cherche à mesurer l'écart à la moyenne. Le premier calcul consiste à calculer l'écart relatif moyen : on constate qu'il est nul. D'où l'idée d'introduire l'écart absolu moyen.
C'est la moyenne des valeurs absolues des écarts à la moyenne.

$$e_a = \frac{n_1 |x_1 - \bar{x}| + ... + n_k |x_k - \bar{x}|}{n}$$

Par exemple avec la série A :

x	10	11	12	13	14	15	16
$x - \bar{x}$	3	2	1	0	1	2	3

Et donc $e_a = \frac{3+2+1+0+1+2+3}{7} = \frac{12}{7} \approx 1,7.$

Remarque : L'écart absolu moyen correspond bien à la moyenne des distances à la moyenne.

3.1.2 Variance. Écart-type

La variance observée de X est la moyenne des carrés des écarts entre les observations et la moyenne.
Si on note $V(X)$ cette variance, on a:

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))^2 = \frac{1}{n}\left((x_1 - m(X))^2 + ... + (x_n - m(X))^2\right).$$

Remarque : la variance s'exprime dans le carré de l'unité des valeurs de la série. On veillera sur les calculatrices à bien utiliser la touche σ_n .

On définit alors l'écart-type comme la racine carrée de la variance observée, on le note $s(X)$.

$$s(X) = \sqrt{V(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))^2}$$

Remarques : 1. On fera attention sur la calculatrice à bien utiliser le bon écart-type.
2. L'écart-type donne de l'importance aux valeurs loin de la moyenne.

Théorème de Koenig :

Toujours sous les mêmes hypothèses : $V(X) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (m(X))^2$.

Autrement dit la variance est égale à la moyenne des carrés moins le carré de la moyenne.

4. Effet d'une transformation affine des données

Propriété : On considère une série statistique S constituée de n observations $x_1, ..., x_n$ de la variable X .
Soit a et b deux réels fixés. On considère la série statistique S' constituée de n observations $y_1, ..., y_n$ telles que pour $1 \leqslant j \leqslant n$, on a : $y_j = ax_j + b$.
Alors les paramètres de S' se déduisent de ceux de S :

	Moyenne	Ecart-type	Médiane	Quartile	Écart inter-quartile
S	$\bar{X}$	$s(X)$	$\text{med}(X)$	Q	$Q_3 - Q_1$
S'	$a\bar{X} + b$	$\lvert a\rvert s(X)$	$a\text{med}(X) + b$	Pour $a > 0$: $a \times Q + b$	$\lvert a\rvert (Q_3 - Q_1)$