CHAPITRE 1

STATISTIQUES DESCRIPTIVES - RAPPELS

I Définitions

La statistique recouvre un ensemble de méthodes d'obtention puis de traitement et d'analyse de données numériques sur des ensembles nombreux et souvent à effectif important.
Les premières statistiques furent élaborées à partir de recensements démographiques. On en a gardé le vocabulaire de base : les ensembles étudiés sont nommés "populations" et leurs éléments "individus".
Les éléments que l'on étudie sur une population sont appelés "variables" ou "caractères".

1) Récupération de données

Les données relatives à une variable peuvent être obtenues par deux méthodes principales :
- la méthode exhaustive qui consiste à étudier ou mesurer la variable pour chaque individu de la population,
- l'échantillonnage, dans lequel on n'étudie qu'une partie de la population, l'"échantillon". Cet échantillon doit être représentatif de la population complète, et pour cela les individus qui le composent doivent être prélevés de façon aléatoire.

Remarque : dans certains cas plus complexes on utilisera des échantillonnages stratifiés, par exemple lorsque la population étudiée est divisée en plusieurs sous-ensembles distincts.

2) Les variables étudiées

Mathématiquement, une "variable" ou un "caractère" est une application X de la population $\mathcal{P}$ dans l'ensemble des nombre réels $\mathbb{R}$. On la note $X : \mathcal{P} \longrightarrow \mathbb{R}$.
La variable peut ainsi prendre différentes valeurs appelées "modalités".

Exemples : la taille, l'âge, la catégorie socioprofessionnelle, etc ...

On distingue deux types de variables :

a) les variables qualitatives qui ne peuvent pas être représentées par des nombres mais doivent être codées pour s'y ramener. Dans ce cas, elles peuvent être simplement nominales, sans notion de hiérarchie, ou au contraire traduire une hiérarchie entre les modalités.
On parlera d'échelle nominale ou ordinale.

b) les variables quantitatives qui sont représentées directement par des nombres.
Une variable quantitative est dite "discrète" si elle ne prend qu'un nombre fini de valeurs ou un ensemble de valeurs que l'on peut numéroter, c'est-à-dire mettre en bijection avec $\mathbb{N}$ ou un sous-ensemble de $\mathbb{N}$. Elle peut résulter dans ce cas du codage d'une variable qualitative.

Une variable quantitative est dite "continue" si elle peut prendre n'importe quelle valeur d'un intervalle donné. Elle peut aussi correspondre à une variable discrète qui prend un grand nombre de valeurs et est alors regroupée en différentes classes.

II Représentation des données

1) Série statistique

Une série statistique est une énumération de l'ensemble des valeurs rencontrées, sans classement particulier.

Exemple : la taille de 20 personnes, exprimée en cm

167 172 154 168 173 182 149 162 175 164 155 174 181 163 172 193 178 166 164 174

2) Tableau statistique et représentation graphique

Le tableau statistique correspond à une mise en forme des données, et dépend du type de variable. On exprime les résultats en termes de fréquences :

- A chaque modalité de la variable X peuvent correspondre plusieurs individus dans l'échantillon étudié.
Si l'échantillon a un effectif de n individus, on notera n_i le nombre d'individus pour lesquels la variable prend une modalité x_i donnée. Ce nombre n_i est l'effectif de la modalité. On appelle parfois cette quantité "fréquence absolue".

- On appelle fréquence relative, ou plus simplement fréquence, de la modalité x_i la quantité $f_i = \frac{n_i}{n}$.
Le pourcentage correspondant à cette modalité est alors égal à $100 * f_i$.

- On appelle fréquence cumulée croissante au niveau x_i la somme des fréquences de toutes les modalités inférieures ou égales à x_i.

Voir exemples dans le TD.

a) Variables quantitatives discrètes

Le tableau se compose des colonnes "modalités", "effectifs", "fréquences", et si nécessaire "fréquences cumulées croissantes". On lui associe un "diagramme en bâtons" dans lequel la hauteur des bâtons correspond à l'effectif de la modalité ou à sa fréquence.

b) Variables quantitatives continues

On effectue une répartition des modalités en "classes". Les amplitudes des différentes classes ne sont pas nécessairement égales. Dans le tableau, la colonne "modalités" est remplacée par une colonne "classe", les autres colonnes restant inchangées.
Par contre, le diagramme associé est un "histogramme", dont la **surface** est proportionnelle à l'effectif ou à la fréquence, et non à la hauteur. Il faut donc définir un rectangle unité qui représente un effectif ou une fréquence de référence.

Remarque : un diagramme en "secteurs" (camembert) contenant des pourcentages peut parfois remplacer les diagrammes cités.

III Indices numériques

1) Indices de position

a) La moyenne

Sur un échantillon de n mesures x_i d'une variable quantitative X. La moyenne de cette variable s'écrit $\bar{X} = \frac{1}{n}\sum_i x_i$ ou $\bar{X} = \frac{1}{n}\sum_i n_i x_i$.

Propriétés :
Si X et Y sont deux variables quantitatives et a et b deux réels, on a : $\overline{aX + bY} = a\bar{X} + b\bar{Y}$.
Si de plus X et Y sont deux variables indépendantes, alors $\overline{X * Y} = \bar{X} * \bar{Y}$.

Remarque :
Si $\bar{X} = 0$, on dit que la variable$X$ est centrée.
Réciproquement, si $\bar{X} \neq 0$, en posant $Y = X - \bar{X}$, on a $\bar{Y} = 0$. Cette variable Y est appelée "variable centrée associée à X". On a ainsi réalisé le centrage de X.

b) La médiane

La médiane M est la valeur de la variable pour laquelle la fréquence cumulée croissante est égale à 0.5. C'est le centre de la série statistique classée par ordre croissant ou décroissant : il y a de part et d'autre de la médiane 50% des valeurs prises par la variable.

Par convention, si l'on trouve une valeur comprise entre deux valeurs atteintes par la variable, mais non atteinte elle-même, on prend pour médiane la moyenne des deux valeurs encadrant la valeur obtenue (la "vraie" médiane n'existe pas).

Exemple : si une variable prend pour valeurs 2, 8, 8, 11, 15, 20, on prendra pour médiane la valeur 9.5, moyenne de 8 et 11.

On définit de même les quartiles Q_1 et Q_3 comme étant les valeurs de la variable pour lesquelles la fréquence cumulée croissante prend respectivement les valeurs 0.25 et 0.75.
L'intervalle interquartile est la quantité $Q_3 - Q_1$.

c) Le mode

Le mode d'une variable quantitative X est la valeur pour laquelle le diagramme en bâtons passe par un maximum.

Attention :
(1) le mode n'est pas forcément unique. On parle de variable "unimodale" ou "plurimodale".
(2) Dans le cas d'une variable continue, on parlera de classe modale plutôt que de mode.

Remarque : on résume souvent ces caractéristiques d'une série statistique à l'aide d'un schéma nommé "boîte à moustache".

2) Indices de dispersion

a) Ecart absolu moyen

Définition : l'écart absolu moyen d'une variable quantitative X est la quantité $\overline{|X - \bar{X}|}$. Ce nombre est assez peu utilisé, on lui préfère en général l'écart quadratique moyen.

b) Ecart quadratique moyen ou écart-type

Définitions : (1) on appelle variance d'une variable quantitative X le réel noté $Var(X)$ ou $V(X)$ tel que $V(X) = \overline{(X - \bar{X})^2}$;

(2) on appelle écart quadratique moyen ou écart-type de X le réel noté $\sigma(X)$ tel que $\sigma(X) = \sqrt{V(X)}$.

Méthode de calcul : $V(X) = \frac{1}{n}\sum_i n_i(x_i - \bar{X})^2 = \overline{X^2} - \bar{X}^2$ avec $\overline{X^2} = \frac{1}{n}\sum_i n_i x_i^2$.

c) La covariance

Définition : on appelle covariance entre les variables quantitatives X et Y la quantité $cov\,(X,Y) = cov(Y,X) = \overline{(X - \bar{X})(Y - \bar{Y})} = \overline{X * Y} - \bar{X} * \bar{Y}$.

Remarque : si X et Y sont indépendantes, $cov\,(X,Y) = 0$.

d) Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire entre deux variables quantitatives X et Y est le réel noté

$$\rho(X,Y) = \frac{cov\,(X,Y)}{\sigma(X) * \sigma(Y)}$$

On montre que : $-1 \leq \rho(X,Y) \leq 1$.

e) Variable réduite, variable centrée réduite

Une variable quantitative X est dite réduite si sa variance est égale à 1.

Si $V(X) \neq 1$, on la réduit en posant $Y = \frac{X}{\sigma(X)}$. Alors $V(Y) = 1$.

Si X n'est ni centrée ni réduite, on définit la variable centrée réduite associée à X par :

$$Z = \frac{X - E(X)}{\sigma(X)}$$

Alors $E(Z) = 0$ et $V(Z) = 1$.

f) Propriétés des variances et covariances

Si X est une variable aléatoire et a et b deux réels, on montre que $V(aX+b) = a^2V(X)$ et donc $\sigma(aX+b) = |a|\sigma(X)$.
Si X et Y sont deux variables aléatoires, $V(X+Y) = V(X) + V(Y) + 2cov\,(X,Y)$ et de même $V(X-Y) = V(X) + V(Y) - 2cov\,(X,Y)$.

Enfin, la covariance est bilinéaire, c'est-à-dire que :
$cov\,(aX+bY, cZ+dT) = ac * cov\,(X,Z) + ad * cov\,(X,T) + bc * cov\,(Y,Z) + bd * cov\,(Y,T)$.

IV Méthode des moindres carrés

1) Généralités

On suppose pour appliquer cette méthode que l'on dispose de deux variables quantitatives X et Y mesurées pour n individus et fournissant des couples de valeurs (x_i, y_i) que l'on a représentés dans le plan par des points $M_i(x_i, y_i)$.

La méthode des moindres carrés consiste à rechercher une droite telle que la somme de ses "distances" aux différents points M_i soit minimale.

Attention : le terme de "distance" est employé au sens large, il s'agit ici d'une expression notée d vérifiant les critères suivants :

1) pour tous points A et B du plan, $d(A,B) \geq 0$,
2) pour tous points A et B du plan, $d(A,B) = 0 \Leftrightarrow A = B$,
3) pour tous points A et B du plan, $d(B,A) = d(A,B)$
4) pour tous points A, B et C du plan, $d(A,B) \leq d(A,C) + d(C,B)$.

La distance utilisée ici est la somme des carrés des différences entre les ordonnées y_i de chaque point M_i et $\widehat{y_i}$ du point de la droite ayant la même abscisse x_i que M_i.

2) Equation de la droite "des moindres carrés"

a) Calcul de l'équation

On cherche une équation de la forme $\hat{y} = ax + b$, et plus précisément ici a et b tels que la somme :

$$S = \sum_{i=1}^{n} (\widehat{y_i} - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

soit minimale.
En considérant cette expression comme un trinôme en b puis comme un trinôme en a, on obtient :

$$S = n * b^2 + 2\left(\sum_{i=1}^{n} (ax_i - y_i)\right) * b + \sum_{i=1}^{n} (ax_i - y_i)^2$$

$$S = \left(\sum_{i=1}^{n}({x_i}^2)\right) * a^2 + 2\left(\sum_{i=1}^{n} x_i(b - y_i)\right) * a + \sum_{i=1}^{n}((b - y_i)^2)$$

Rappel : le trinôme $\alpha t^2 + \beta t + \gamma$ possédant un coefficient $\alpha > 0$ passe par un minimum en une valeur $t = -\frac{\beta}{2\alpha}$ (résultat connu ...).

En utilisant successivement la première puis la seconde forme de S, on obtient donc :

$$b = -\frac{2(\sum_{i=1}^{n}(ax_i - y_i))}{2n} \text{ et } a = -\frac{2(\sum_{i=1}^{n} x_i(b - y_i))}{2\sum_{i=1}^{n}({x_i}^2)}$$

Ces égalités donnent après transformation :

$$b = \bar{Y} - a * \bar{X} \text{ et } a = \frac{\mathrm{Cov}(X,Y)}{V(X)}$$

L'équation de la droite de régression ou droite d'estimation de Y en X s'écrit alors :

$$\hat{y} = ax + \bar{Y} - a * \bar{X} = a(x - \bar{X}) + \bar{Y}$$

ce qui permet de constater qu'elle passe par le point moyen de coordonnées $(\bar{X}, \bar{Y})$, qui représente le centre de gravité des points.

b) Validité de la droite de régression

La droite de régression existe pour n'importe quelle donnée d'un couple de variables (sauf quelques cas particuliers). Cependant elle n'est utilisable dans la pratique que lorsque le coefficient de corrélation linéaire entre X et Y est assez proche de 1 en valeur absolue.

On considère plus précisément que l'ajustement est valide et acceptable lorsque :

$$0.7 \leq |\rho(X,Y)| \leq 1$$

Dans les autres cas, on devra chercher une expression de l'estimation de Y en fonction de X à l'aide d'autres fonctions (carré, racine carrée, exponentielle, logarithme, etc...).

c) Droite d'estimation de X en Y

A moins d'un raisonnement particulier sur les données, il est en principe aussi logique de calculer une droite d'estimation de X à l'aide de Y que le contraire.
En reprenant des calculs analogues à ceux du a), on trouve l'équation suivante pour cette droite :

$$\hat{x} = a'y + b' = a'(y - \bar{Y}) + \bar{X} \text{ avec } a' = \frac{Cov(Y,X)}{V(Y)} \text{ et } b' = \bar{X} - a'\bar{Y}$$

La droite obtenue est différente de la droite de régression de Y en X, dans presque tous les cas.

Remarques :

- On constate que $a * a' = \rho^2(X,Y)$. Si les deux droites sont identiques, $a' = \frac{1}{a}$ et par suite $|\rho(X,Y)| = 1$: l'ajustement est valide, par contre, si $\rho(X,Y)$ est assez loin de 1 en valeur absolue, les deux coefficients directeurs sont loin d'être inverses l'un de l'autre, et les deux droites sont sensiblement différentes.

- Dans les cas où la droite de régression n'est pas pertinente, on devra chercher une estimation de Y en fonction de X en utilisant des fonctions numériques. Les plus courantes sont les fonctions carré, racine carrée, logarithme, exponentielle.

3) Lissage

Le lissage a pour but de réduire les irrégularités ou les singularités dans une série statistique. Il existe de nombreuses techniques.

En particulier, le lissage par moyenne mobile consiste à regrouper p points consécutifs et à remplacer chaque groupe de points par son point moyen. On commencera par lisser le tableau avant de calculer une droite de régression.

Exemple :
Le tableau

x_i	1	2	3	4	5	6	7	8
y_i	3	4	5	2	8	10	4	5

traité avec des regroupements par 3 deviendrait :

x_i	2	3	4	5	6	7
y_i	4	3.67	5	6.67	7.33	6.33

On effectuerait donc ici les calculs sur le nouveau tableau.