Thème 0 : Statistique Descriptive

Introduction

Statistique descriptive : permet de décrire les données à l'aide de graphiques et de paramètres d'une façon compréhensible et utilisable
Probabilité : permet de modéliser efficacement les phénomènes étudiés en statistiques
Statistique inférencielle : permet de faire des prévisions ou généralisations à toute une population à partir d'échantillons
Régression linéaire : permet d'étudier la relation existante entre deux variables. Met en place des modèles de prévisions et des outils pour valider ceux ci

Vocabulaire

Ensembliste	Statistique
Ensemble	Population (Ω)
Application	Variable / Caractère
Elément	Individu / unité statistique
Sous-Element	Sous-population
Cardinal	Effectif

Fréquence d'une sous population E de Ω : $f(E) = \frac{Card(E)}{Card(\Omega)} \in [0,1]$

Variables ou caractères

Variables **qualitatives** : appartenance à une catégorie
Variables **quantitative**/**numériques** : taille, poids, volume...
Variables **discrète** : nombre fini ou indéfini dénombrable de valeurs observées
Soit une variable discrète X, l'ensemble des valeurs (modalités) prises par X est l'ensemble :
$X(\Omega) = \{x_1, x_2, \dots, x_n \dots\} = \{x_i, i \in \mathbb{N}\}$

Loi d'une variable quantitative, fonction de répartition

La **loi** ou **distribution empirique** d'une variable X sur Ω est la donnée de la fréquence de chaque classe définie par la variable X

— Si X quantitative ou qualitative discrète, sa loi est définie par la fréquence de chaque sous-population du type $\{X = x_i\} = \{\omega \in \Omega, X(\omega) = x_i\}$;

— Si X continue et si les valeurs possibles de X sont réparties en classes C_i, la loi est la donnée de chaque fréquence des sous-populations $\{X \in C_i\} = \{\omega \in \Omega, X(\omega) \in C_i\}$

La **fonction de répartition empirique** de X est la fonction, notée F_x, qui à $x \in \mathbb{R}$ associe la fréquence de la sous-population $\{X \leq x\}$:

$$F_X : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto F_X(x) = \frac{Card\{\omega \in \Omega, X(\omega) \leq x\}}{Card\Omega}$$

Grandeurs statistiques usuelles

La **moyenne** du caractère X est la quantité
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
La moyenne est une statistique *peu robuste* (sensible aux valeurs extrêmes)
Proposition : si $Y = aX + b$ avec $a, b \in \mathbb{R}$, alors : $\bar{y} = a\bar{x} + b$
Le **mode** ou **classe modale** d'une distribution statistique est la valeur ou la classe du caractère qui correspond à la plus grande fréquence.
La **mediane** du caractère X est la valeur M_e telle que, en notant $f(\dots)$ la fréquence : $f(\{X \leq M_e\}) \geq \frac{1}{2} et f(\{X \geq M_e\}) \geq \frac{1}{2}$
Les **quartiles** $Q_1, Q_2 et Q_3$ sont les valeurs permettant de diviser la population en quatre sous-populations d'effectif égaux, représentant chacune 25% de la population totale.
L'**étendue** est la différence entre les valeurs extrêmes du caractère : $\omega = x_{max} - x_{min}$
La **variance** de la variable X est la quantitié $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ représentant la moyenne des carrés des écarts entre les observations et leur moyenne
Proposition : transformation affine sur la variance : si on pose $Y = aX + b$, $\sigma_Y^2 = a^2\sigma_X^2$
L'**écart-type** de X est la racine carrée σ de la variance

Distributions à deux caractères

L'**effectif marginal en** X et la **fréquence marginale en** X de la classe C_i :
$n_{i.} = \sum_{j=1}^{s} n_{ij}$ et $f_{i.} = \frac{n_{i.}}{n} = \sum_{j=1}^{s} f_{ij}$
La **loi conditionnelle de** Y **sachant** $X \in C_i$ est la donnée, pour tout $j \in \{1, \dots, s\}$ des fréquences relatives des classes D_j par rapport à C_i : $f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$
Les deux variables X et Y sont dites **indépendantes** si la loi conditionnelle de Y sachant $X \in C_i$ ne dépend pas de i

Cas de deux variances quantitatives

La **covariance** de deux variables quantitatives X et Y est : $\mathbb{C}\text{ov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
La covariance permet de quantifier la liaison entre les deux variables (positive = même sens = **liaison positive**, négatif = sens contraires = **liaison négative**)
Propriétés de la covariance :
(1) La covariance est symétrique : $\mathbb{C}\text{ov}\ (X,Y) = \mathbb{C}\text{ov}\ (Y,X)$
(2) Covariance de X avec elle-même : $\mathbb{C}\text{ov}\ (X,X) = \mathbb{V}(X)$
(3) Transformation affine : $\mathbb{C}\text{ov}\ (aX + b, cY + d) = ac\mathbb{C}\text{ov}\ (Y,X)$
(4) Variance d'une somme : $\mathbb{V}(X+Y) = \mathbb{V}(X) + 2\mathbb{C}\text{ov}(X,Y) + \mathbb{V}(Y)$
(5) Inégalité de Cauchy-Schwartz : $|\mathbb{C}\text{ov}\ (X,Y)| \leq \sigma(X)\sigma(Y)$ avec égalité si et seulement si il existe une relation affine entre X et Y : $Y = aX + b$ ou $X = cY + d$
(6) Cas de variables indépendantes : si X et Y sont indépendantes, leur covariance est nulle. La réciproque est fausse
Lorsque deux variables ont une covariance nulle, on dit qu'elles sont **décorrélées**.