Guide-Formulaire Statistiques L1 Sociologie

I Statistiques descriptives à une variable. Vocabulaire :

Population. On appelle *population* $\mathcal{P}$ tout ensemble étudié par la statistique. On notera N le nombre d'éléments de $\mathcal{P}$. Il s'agit en générale d'objets de "même nature".
Les éléments de la population sont appelés *individus*.

Examples. La population des étudiants de L_1 de sociologie - un individu est un(e) étudiant(e).
L'ensemble des villes de France - un individu est une ville.

Échantillon. Un *échantillon* est un sous-ensemble d'une population choisi au hasard.

Caractère - Variable statistique. Une *caractère* (ou *variable statistique*) est une quantité ou qualité définie sur les individus de la popuation $\mathcal{P}$ qui peut varier d'un individu à l'autre.

Examples. Sur la population des étudiants de L_1, l'âge X est un caractère.
Sur la population des villes de France, le nombre d'habitants X par ville est un caractère.

Indicateurs de tendance centrale :

Moyenne. La *moyenne arithmétique* est la somme des valeurs divisée par le nombre d'éléments :

1. Petit échantillon (ou série de données brutes) : $m(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$

2. Grand échantillon : $m(X) = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$

3. Variable regroupée par classes : $m(X) = \frac{1}{n}\sum_{i=1}^{n} n_i c_i$ avec c_i centre de la i-ème classe.

Exemple. La variable X prend les valeurs suivantes sur 10 individus :
$x_1 = 3$, $x_2 = 6$, $x_3 = 9$, $x_4 = 4$, $x_5 = 17$, $x_6 = 1$, $x_7 = 10$, $x_8 = 12$, $x_9 = 9$, $x_{10} = 11$. On obtient donc

$m(X) = (1/10)\sum_{i=1}^{10} x_i = (3+6+9+4+17+1+10+12+9+11)/10 = 82/10 = 8,2$

Mode. Le *mode* d'un caractère est la valeur la plus fréquente du caractère dans l'ensemble de données.

Médiane. La *médiane* d'une variable quantitative est la valeur qui partage l'échantillon en deux ensembles d'effectifs égaux : 50 % des valeurs lui sont supérieures et 50 % lui sont inférieures.

Quartiles. Les *quartiles* sont les trois valeurs qui découpent la distribution en quatre classes d'effectifs égaux, on les note Q_1, Q_2 et Q_3. Bien évidement $Q_2 =$ médiane.

Indicateurs de dispersion. Les indicateurs de dispersion absolue montrent de combien les valeurs d'une distribution s'écartent en général de la valeur de référence - un indicateur de tendence centrale.

Étendue. C'est égale à la différence entre la plus grande et la plus petite valeur de la distribution.

Écart interquartiles. C'est la longueur de l'intervalle de Q_1 à Q_3, c-à-d $Q_3 - Q_1$.

Variance : La variance est un indicateur de la dispersion des données par rapport à la moyenne.
C'est la moyenne des carrés des écarts $X - m(X)$ de la moyenne : $V(X) = m((X - m(X))^2)$.
Il est plus simple de calculer la variance par la formule "moyenne des carrés moins le carré de la moyenne" :

$$\textbf{Variance :}\quad V(X) = m(X^2) - m(X)^2\,.$$

Selon les cas considérés on aura

1. Petit échantillon (ou série de données brutes) : $V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (m(X))^2$.

2. Grand échantillon : $V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} n_i x_i^2\right) - (m(X))^2$.

3. Variable regroupée par classes : $V(X) = \left(\frac{1}{n}\sum_{i=1}^{n} n_i c_i^2\right) - (m(X))^2$.

Exemple. En prenant les valeurs du caractère X de l'exemple précédent on calcule
$m(X^2) = (1/10)\sum_{i=1}^{10} x_i^2 = (3^2+6^2+9^2+4^2+17^2+1^2+10^2+12^2+9^2+11^2)/10 = 87,8.$
Alors la variance est donnée par $V(X) = 87,8 - 8,2^2 = 20,56$

Écart-type. L'*écart-type* est la racine carrée de la variance :

$$\textbf{Écart-type :}\quad s(X) = \sqrt{V(X)}\,.$$

Exemple. De l'exemple pécédent on obtient $s(X) = \sqrt{20,56} = 4,5343$.

II Couple de variables statistiques.

Nuage statistique. Pour un couple de variables (X, Y) chaque individu de l'échantillon est représenté dans le plan par un point de coordonées (X_i, Y_i) qui sont les valeurs des variables X, Y pour le i-ème individu. L'ensemble de ces points s'appelle *nuage statistique* de l'échantillon.

Barycentre. Un indicateur de tendence centrale du nuage statistique, analogue de la moyenne, est le *barycentre* (appelé aussi *point moyen*). Ses coordonnées sont les moyennes des caractères X, Y :

$$G = (m(X), m(Y)) .$$

Droites de régression : Une relation entre les caractères X, Y est linéaire si le nuage de points peut "s'ajuster" à une droite. Une telle droite sera décrite par une équation de la forme $Y = aX + b$.

Le nombre a, appelé *pente* ou *coefficient directeur*, mesure l'inclinaison de la droite et décrit le sens de l'inclinaison (si a est positif, la droite monte quand on la parcourt de la gauche vers la droite, et la droite descend si a est négatif). On obtient la pente à partir de deux points (x_1, y_1), (x_2, y_2) de la droite :

$$\textbf{Pente d'une droite :} \quad a = \frac{y_2 - y_1}{x_2 - x_1} .$$

Le nombre b est "l'ordonnée à l'origine", c-à-d l'ordonnée de l'intersection de la droite avec l'axe Y.

$$\textbf{Équation de la droite :} \quad Y = a(X - x_1) + y_1 .$$

Exemple. Ces formules donnent l'équation de la droite de Mayer quand $G_1 = (x_1, y_1)$ et $G_2 = (x_2, y_2)$.

$$\textbf{Covariance :} \quad Cov(X, Y) = m(XY) - m(X)m(Y) .$$

Droites obtenues par la métohode des moindres carrés :

$$D_{Y|_X} : \; Y = aX + b ; \quad a = \frac{Cov(X, Y)}{V(X)} ; \quad b = m(Y) - a \cdot m(X)$$

$$D_{X|_Y} : \; X = \hat{a}Y + \hat{b} ; \quad \hat{a} = \frac{Cov(X, Y)}{V(Y)} ; \quad \hat{b} = m(X) - \hat{a} \cdot m(Y)$$

Les deux droites $D_{Y|_X}$ et $D_{X|_Y}$ passent par le point moyen $G = (m(X), m(Y))$.

$$\textbf{Coefficient de corrélation linéaire :} \quad r(X, Y) = \frac{Cov(X, Y)}{s(X) \cdot s(Y)} .$$

III Caractères regroupés par classes.

Médiane d'un caractère regroupé par classes. On repère la classe j qui contient la médiane. On note $F^{\text{cum}}(a_j)$ la fréquence cumulée à sa borne inférieure a_j, $F^{\text{cum}}(a_{j+1})$ la fréquence cumulée à sa borne supérieure a_{j+1} et $A_j = a_{j+1} - a_j$ son amplitude. En application du thèorème de Thalès on trouve

$$\text{Médiane} = a_j + \frac{A_j}{F^{\text{cum}}(a_{j+1}) - F^{\text{cum}}(a_j)} \left(0.5 - F^{\text{cum}}(a_j)\right) .$$

Si les fréquences cumulées sont exprimées en pourcentages on a :

$$\text{Médiane} = a_j + \frac{A_j}{F^{\text{cum}}_{\%}(a_{j+1}) - F^{\text{cum}}_{\%}(a_j)} \left(50 - F^{\text{cum}}_{\%}(a_j)\right) .$$

Pour calculer le premier et troisième quartiles, on repère la classe j qui contient le quartile cherché et on utilise la même formule, mais on remplace respectivement $0,5$ (ou 50%) par $0,25$ (25%) pour Q_1 et par $0,75$ (75%) pour Q_3.