Mini-Glossaire de Statistique Descriptive - Jean VAILLANT

Amplitude d'une classe (ou d'un intervalle) : C'est la longueur de l'intervalle. L'amplitude de la classe $]a_{i-1}\,;a_i]$ est $a_i - a_{i-1}$. Exemple : la classe]16;43] est d'amplitude $43 - 16 = 27$ (unités de mesure).

Caractère qualitatif : Un caractère statistique est qualitatif si ses valeurs, ou modalités, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme,$\cdots$, n'ont pas de sens. Exemples : *Sexe de la personne interrogée, Situation familiale, Numéro de son département de naissance; Etat du temps constaté à une station expérimentale chaque jour; Variété de la plante observée, Etat sanitaire, numéro de Site.*

Caractère quantitatif : Un caractère statistique est quantitatif si ses valeurs sont des nombres sur lesquels des opérations arithmétiques telles que somme, moyenne,$\cdots$, ont un sens. Exemples : *Taille, Poids, Salaire, Rendement, Note à un examen, PNB/habitant, Espérance de vie, Nombre d'habitants, Taux d'infestation.*

Caractère statistique (ou variable statistique) : C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative.

Classe modale : C'est la classe correspondant au maximum de l'histogramme (plus grand effectif par unité d'amplitude). Dans le cas d'une classe modale unique, on parle de distribution continue unimodale.

Classes statistiques : Intervalles de valeurs d'une variable statistique. L'ensemble des classes forment une partition de l'ensemble des valeurs possibles de la variable. Par exemple, si tous les salaires des employés d'une entreprise se situent entre 1000 et moins de 20000 EUR, on peut construire (par exemple) les classes :

$$]1000;3000],]3000;5000],]5000;7000],]7000;20000]$$

Les classes statistiques sont exclusives c'est-à-dire une valeur observée appartient à une classe et une seule.

Remarque : on peut utiliser une distribution en classes statistiques pour une variable discrète pouvant prendre beaucoup de valeurs distinctes. Exemple : it nombre d'insectes par unité d'échantillonnage dans le cas de pullulation.

Coefficient de corrélation (linéaire) : Le coefficient de corrélation entre deux variables statistiques X et Y sur les mêmes individus est le nombre r vérifiant : $r = \dfrac{s_{xy}}{s_x s_y}$

où s_{xy} est la covariance entre X et Y, et s_x, s_y les écarts-types de X et Y.

Ce coefficient est toujours compris entre -1 et + 1.

S'il est proche de + 1 ou - 1 , X et Y sont bien corrélées linéairement, c'est-à-dire qu'elles sont liées entre elles par une relation presque affine ; le nuage de points est presque aligné le long d'une droite (croissante si $r = +1$, décroissante si $r = -1$). S'il n'y a aucun lien entre X et Y, ce coefficient est nul, ou presque nul.

Coefficient de Spearman (ou coefficient de corrélation des rangs) : C'est, dans le cas de deux variables ordinales X et Y mesurées sur les mêmes individus, le coefficient de corrélation entre le rang des individus pour X et le rang des individus pour Y.

Coefficient de variation : C'est le rapport écart-type sur la moyenne. Il est calculé pour des variables statistiques positives :taille, durée, poids. C'est un nombre sans dimension (c'est-à-dire qu'il est indépendant du choix des unités de mesure). Il permet de comparer la dispersion autour de la moyenne de variables statistiques ayant des échelles ou des unités de mesure différentes.

Courbe cumulative : On l'utilise quand la variable quantitative est continue. Il s'agit d'une fonction continue, affine par morceaux. Pour la tracer, on relie les points $(x_i, F(x_i))$, pour les points distincts x_i de la série statistique.

Diagramme circulaire (ou à secteurs angulaires ou camembert) : Il s'agit d'un disque divisé en sections angulaires. Chaque section correspond à une modalité de la variable qualitative et a un angle proportionnel à la fréquence de cette modalité.

Diagramme cumulatif : C'est le tracé de la fonction qui à tout x associe $F(x) =$ proportion d'observations $\leq x$. Il s'obtient au moyen des effectifs cumulés croissants. On a une fonction dite en escalier. On l'utilise dans le cas d'une variable quantitative discrète.

Diagramme figuratif : Chaque modalité de la variable qualitative est représentée par une image (ordinateur, maison, plante, avion,...) rappelant la variable (ou la population) statistique étudiée, et de taille proportionnelle à la fréquence de cette modalité.

Dispersion : Un indicateur statistique est dit de dispersion s'il s'agit d'un nombre clé caractérisant la variabilité des observations dans la série statistique. Ainsi l'étendue donne l'écart entre la plus petite et la plus grande valeur dans la série statistique; l'écart interquartile donne la plage de variation des observations situées dans le second et troisième quarts de la série statistique réordonnée.

Distribution statistique : Ensemble des modalités, valeurs, ou classes d'une variable, avec les effectifs observés correspondants.

Ecart interquartile : C'est la différence I entre le 1er et le 3ème quartile : $I = Q_3 - Q_1$.

Ecart-type : pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i,, l'écart-type noté s_x est donné par la formule :

$$s_x = \sqrt{\frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)}$$

où $\bar{x}$ est la moyenne de la série.

Etendue : C'est l'écart entre la plus petite et la plus grande valeur dans la série statistique.

Fractiles (ou quantiles) : On appelle fractiles des valeurs divisant une série en plusieurs parties. Pour une valeur α comprise entre 0 et 1, le fractile d'ordre α noté q_α est par définition tel que la proportion de valeurs inférieures à q_α vaut α. On a donc $F(q_\alpha) = \alpha$. Les fractiles divisant la série en k parties d'effectifs égaux ont parfois une dénomination commune : Les 3 quartiles divisent la série en 4 parties d'effectifs égaux, les 9 déciles en 10, les 99 centiles en 100. Les 3 quartiles sont notés Q_1, Q_2, Q_3 (Q_2 étant la médiane).

Fréquence (ou fréquence relative) : C'est la proportion (ou le pourcentage) d'individus pour lesquels une variable statistique a pris une valeur donnée. Si, sur 150 familles, 50 ont 2 enfants, on dira que la fréquence f_i correspondant à la valeur $x_i = 2$ de la variable *nombre d'enfants*, est : 0.33 ou 1/3 ou 33.33%.

Fréquence cumulée : Résultat de l'addition, de proche en proche, des fréquences d'une distribution observée, soit en commençant par le 1er :

$F_1 = f_1$, $F_2 = f_1 + f2, \cdots$, $F_i = f_1 + f_2 + \cdots + f_i$ (fréquences cumulées croissantes),

soit en commençant par le dernier :

$F_K^* = f_K, F_{K-1}^* = f_K + f_{K-1}, \cdots, F_i^* = f_K + f_{K-1} + \cdots + f_i$ (fréquences cumulées décroissantes).

Histogramme : Graphique permettant de représenter une distribution continue regroupée en classes : rectangles juxtaposés dont les bases sont les classes, et les surfaces sont proportionnelles aux effectifs (ou fréquences) associés.

Indépendance : Deux variables statistiques X et Y sont dites indépendantes si la distribution de Y conditionnelle à $X = x$, pour tout x, est constante (c'est-à-dire ne dépend pas de x). Cela signifie que les profils des lignes du tableau de contingence sont identiques, ou de façon équivalente que les profils des colonnes du tableau de contingence sont identiques, et donc que la distribution de fréquences conditionnelle est égale à la distribution de fréquences marginale.

Indicateur statistique (ou résumé numérique) : C'est un nombre permettant de résumer numériquement les traits principaux d'une distribution statistique. On parle aussi de résumé numérique. On distingue principalement deux types d'indicateurs :

- les indicateurs de position (ou de tendence centrale) qui donne une idée de l'ordre de grandeur de la série;
- les indicateurs de dispersion qui donnent une idée de la variabilité dans la série.

Inégalité de (Bienaymé)-Tchébichev : Pour toute série statistique $x_1, \cdots, x_n$ de moyenne $\bar{x}$ et d'écart-type s_x , la proportion de valeurs dans l'intervalle $[\bar{x}-k\times s_x; \bar{x}+k\times s_x]$ est supérieure à $1 - \dfrac{1}{k^2}$, pour tout nombre $k \geq 1$. Par exemple, 75% des valeurs au moins appartiennent à : $[\bar{x} - 2s_x; \bar{x} + 2s_x]$, c'est-à-dire s'écartent de moins de 2 écart-types de la moyenne.

Intervalle interquartile : C'est l'intervalle dont les bornes sont le 1er et le 3ème quartile : $[Q_1, Q_3]$. Il contient 50% des observations; rappelons que 25% des valeurs de la série statistique sont inférieures à Q_1 et 25% sont supérieures à Q_3.

Intervalle médian : C'est l'intervalle dont toutes les valeurs vérifient la propriété de la médiane pour la série statistique étudiée.

Médiane : C'est le fractile d'ordre 0.5. La médiane est notée M_e et vérifie $F(M_e) = 0.5$. Il y a autant de valeurs inférieures à M_e que supérieures à M_e dans la série statistique.

Mode : C'est la valeur la plus fréquente dans la série statistique. Le mode n'est pas forcément unique. Quand il existe plusieurs modes, la distribution statistique est dite multimodale.

Moyenne : Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la moyenne notée $\bar{x}$ est la somme des valeurs divisée par le nombre de valeurs. Elle est donné par la formule : $\dfrac{1}{n}(n_1x_1 + \cdots + n_kx_k)$.

Nuage de points : Ensemble de points isolés représentés dans un graphique cartésien. Une séries à deux caractères quantitatifs $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$ peut être représentée par les n points $M_1, M_2, \cdots, M_n$ de coordonnées $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$.

Population statistique : Une population statistique est un ensemble d'éléments sur lesquels porte une étude. Exemples : ensemble des électeurs d'une région; ensemble des accidents de la route dans une zone, pendant une période; ensemble de parcelles cultivées sur lesquelles on peut mesurer un rendement; ensemble de pays pour lesquels on dispose de données géographiques ou économiques, ...

Position : Un indicateur statistique est dit de position (ou de tendance centrale) s'il s'agit d'un nombre clé permettant de préciser où se répartit une certaine fraction des observations. Ainsi les quartiles permettent de situer le quart inférieur, la moitié, le quart supérieur des observations.

Profil : C'est une distribution conditionnelle de fréquences (et non d'effectifs). Dans un tableau de contingence à I lignes et J colonnes, le profil de la ligne i est obtenu en divisant les effectifs $n_{i1}, n_{i2}, \cdots, n_{iJ}$ de cette ligne par la somme $n_{i.}$ de ces effectifs. On obtient : $\dfrac{n_{i1}}{n_{i.}}, \dfrac{n_{i2}}{n_{i.}}, \cdots, \dfrac{n_{iJ}}{n_{i.}}$. De même, le profil de la colonne j est : $\dfrac{n_{1j}}{n_{.j}}, \dfrac{n_{2j}}{n_{.j}}, \cdots, \dfrac{n_{Ij}}{n_{.j}}$. où $n_{.j}$ est la somme des effectifs de cette colonne.

Quartiles : Ce sont les 3 fractiles d'ordre 0,25, 0.5 et 0,75. Ils sont notés dans l'ordre croissant Q_1, Q_2, Q_3. Ils divisent la distribution statistique en quatre parties d'égale fréquence. Q_1 est le premier quartile, Q_3 le troisième. Q_2 est la médiane. (voir fractiles).

Résumé numérique : Voir indicateur statistique.

Série statistique (ou distribution observée) : Séquence des modalités, ou valeurs d'une variable statistique. L'ordre correspond souvent à l'ordre chronologique de recueil des observations.

Statistique Descriptive : Ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer des données nombreuses et variées.

Statistique Descriptive univariée : La Statistique Descriptive univariée consiste en la description de chacun des caractères statistiques, un par un, et non des liens éventuels existant entre eux.

Statistique Descriptive multivariée : La Statistique Descriptive multivariée consiste en la description d'un nombre $k > 1$ de variables mesurées ou observ'ees simultanément sur les mêmes individus. Elle permet de mettre en évidence le type de lien existant éventuellement entre ces variables. Si $k = 2$, on parle de Statistique Descriptive bivariée.

Statistique Inférentielle : La Statistique Inférentielle utilise la théorie des probabilités pour extrapoler à toute la population statistique, des résultats observés sur des échantillons. Elle inclut *l'Estimation Statistique* d'une part, et la *Théorie des Tests d'hypothèses* d'autre part.

Tableau de contingence : C'est le tableau d'effectifs obtenu par tri croisé d'une série bivariée (ou multivariée).

Tri à plat d'une série statistique brute : C'est l'inventaire des modalités ou valeurs rencontrées dans la série, avec les effectifs correspondants.

Tri croisé d'une série bivariée : C'est l'inventaire des modalités ou valeurs rencontrées conjointement dans une série comportant deux variables mesurées pour chaque individu statistique, avec les effectifs correspondants.

Variable statistique (ou caractère statistique) : C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative.

Variance : Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la variance notée s_x^2 est donnée par la formule :

$s_x^2 = \frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)$. La variance est le carré de l'écart-type.