Cours : statistiques univariées

I. Vocabulaire statistique

Le but de la statistique est de dégager des significations pertinentes à partir de données obtenues au cours de l'étude d'un phénomène (social, biologie, géographique, historique, physique, économique...).

On peut distinguer la statistique descriptive qui permet d'ordonner et de réduire les données afin de pouvoir en donner un résumé et en faire une interprétation et la statistique inférentielle qui tente de faire des prévisions et de déterminer des relations de corrélation (C'est une méthode de raisonnement inductif).

Définition : On appelle population l'ensemble E des unités ou individus sur lequel on effectue une analyse statistique.

Définition : On appelle échantillon de E un sous-ensemble fini de E.

Remarque :On effectue un échantillonnage de E lorsque le cardinal de E est beaucoup trop important pour que E soit étudiée directement. C'est la différence entre un recensement et un sondage.

La question de l'échantillonnage est fondamental en pratique : comment réaliser un prélèvement qui permette de généraliser les résultats obtenus sur l'échantillon à la population entière. Différentes techniques d'échantillonnage ont été théorisées (par quotas, par grappes...).

Définition : On appelle variable, ou caractère, un trait déterminé présent chez tous les individus de la population.

Un caractère est quantitatif si il est mesurable et qualitatif sinon.

Définition : On appelle modalités les différentes valeurs possibles de la variable.

Pour une variable qualitative, les modalités sont les étiquettes.

Pour une variable quantitative, les modalités sont les valeurs des mesure de ce caractère, c'est à dire des réels qui peuvent être compris dans un intervalle (on dit que la variable est continue) ou dans un ensemble discret de nombres isolés (on dit la variable est discrète).

Remarque :Dans le cas des valeurs continues, on pourra être amené à regrouper les valeurs par classe d'intervalle afin d'en faire une étude discrète.

On peut toujours passer d'une variable quantitative à une variable qualitative en identifiant les modalités par leur dénomination. On perd alors de l'information.

Les modalités doivent être incompatibles et exhaustives : chaque individu de la population doit présenter une et une seule modalité.

II. Statistiques univariées

Dans ce chapitre on s'intéresse à une population (ou un échantillon de cette population) que l'on note E et à un caractère C de cette population. On suppose que $Card(E) = N$ et $Card(C) = p$. On note $E = \{e_1, ...e_N\}$ et $C = \{x_1, ...x_p\}$.

2.1 Série statistique

Définition : Une série statistique est l'ensemble des valeurs d'un caractère prises spécifiquement par les individus de E observés.

Définition : Soit S une série statistique sur E pour le caractère C. Pour $i \in [|1, p|]$, on note A_i l'ensemble des éléments de E qui ont la modalité x_i. On appelle effectif de la modalité x_i le cardinal n_i de A_i. On a donc $n_i = Card(A_i)$.

Remarque :en pratique le statisticien ignore ensuite quels sont les individus exacts qui présentent la modalité x_i, il se contente de l'information sur l'effectif.

On a bien sur $\sum_{k=1}^{p} n_i = N$.

Définition : La fréquence f_i de la modalité x_i dans la population E est définie par $f_i = \frac{n_i}{N}$.

Remarque :Une fréquence est donc un nombre réel compris entre 0 et 1.

la somme des fréquences est égale à 1.

Les effectifs ne permettent pas la comparaison entre deux séries statistiques tandis que les fréquences le permettent.

Définition : On suppose que les modalités de C peuvent être ordonnées et que $x_1 \leq ... \leq x_p$. La fréquence cumulée croissante de la modalité x_i, , notée F_i, est défini par $F_i = f_1 + ... + f_i$.

Remarque :Elle représente la proportion d'individus qui ont une modalité inférieure ou égale à x_i.

2.2 Comment résumer une série statistique

Définition : Les paramètres de position (mode, moyenne, médiane) permettent de savoir autour de quelles valeurs se distribuent les effectifs des différentes modalités.

Définition : Le mode d'une série est la modalité qui à la plus grande fréquence (ou le plus grand effectif de manière équivalente).

Remarque :Cet indicateur est pertinent pour les variables quantitative et qualitative.

Un même caractère peut avoir plusieurs modes.

Le mode est insensible aux extrêmes mais il ne donne pas d'information sur les autres modalités du caractère.

Définition : La médiane est la modalité qui a une fréquence cumulé croissante de 0.5.

Remarque :Cet indicateur n'a de sens que pour les caractères ordonnées.

Pour des variables discrètes, on prendra la moyenne des deux valeurs encadrant la fréquence cumulée de 0.5 si elle n'existe pas.

La médiane n'est pas influencée par des valeurs extrêmes des modalités.

Définition : La moyenne (arithmétique), notée $\overline{x}$, est définie par $\overline{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$.

Remarque :On peut écrire une formule de la moyenne avec les fréquences : $\overline{x} = \sum_{i=1}^{p} f_i x_i$.

La moyenne est sensible aux valeurs extrêmes.

Définition : Les indicateurs de dispersion (étendue, intervalle interquartile, variance et écart-type) ne sont valables que pour les variables statistiques quantitatives. Ils donnent des informations sur la manière dont les effectifs des modalités se répartissent autour des indicateurs de position.

Définition : On appelle quartiles les modalités Q_1, Q_2, Q_3 pour lesquelles les fréquences cumulées croissantes sont respectivement 0.25, 0.5, 0.75.

Remarque :Q_2 correspond à la médiane.
Pour les variables discrètes, on applique la même méthode que pour la médiane.
L'intervalle interquartile $[Q_1, Q_3]$ contient la moitié de l'effectif total.
On peut généraliser avec les déciles et les percentiles.

Définition : La variance statistique d'une série X, notée $s^2(X)$, est définie par :
$s^2(X) = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$.

Remarque :La variance est donc la moyenne des carrés des écarts à la moyenne.
Il existe un écart-type de population dans lequel on divise par $(N-1)$ au lieu de diviser par N.
Plus les valeurs sont loin de la moyenne, plus la variance augmente. Plus les valeurs sont dispersées, plus la variance augmente. C'est pourquoi la variance est un indicateur de dispersion.

Propriété :

1. $s^2(X) \geq 0$
2. $s^2(X) = 0$ si et seulement si X est une série statistique constante.
3. $s^2(X) = \frac{1}{N}\sum_{i=1}^{p} n_i {x_i}^2 - (\overline{x})^2$ (formule de la variance ou de Koenig-Huygens)
4. $s^2(aX + b) = a^2 s^2(X)$ pour tout (a, b) réels.

Remarque :On note parfois la formule de la variance : $s^2(X) = \overline{X^2} - \overline{X}^2$

Définition :L'écart-type de la série X, noté $s(X)$ est définie par $s(X) = \sqrt{s^2(X)}$.

Remarque :L'écart type est bien défini.
L'écart-type est une grandeur homogène à une distance, d'où une interprétation plus facile en terme de dispersion des valeurs de la série autour de la moyenne.

2.3 représentation graphique Les *diagrammes circulaires* sont utiles pour visualiser rapidement les rapport entres les différentes valeurs des variables (par ex : proportion d'individus récupérant en moins de 30 secondes, entre 30 sec et 60 sec, entre 60 sec et 90 sec, entre 90 sec et 120 sec après un test à l'effort).

Les *courbes* sont utiles pour représenter une évolution continue d'une variable en fonction d'une autre (par exemple : pH en fonction du volume d'une solution titrante).

les *histogrammes* sont utiles pour représenter une variables en fonction d'une variable regroupée par classe d'intervalles (par ex : quantité d'oxygène absorbé en fonction de la taille de la plante).

les *diagramme en bâton* (ou en barre pour être plus visibles) sont utiles pour représenter une variable en fonction de valeur discrètes ou nominales (par exemple : temps de récupération à l'effort en fonction du sexe de l'individu).

Les *diagrammes en boite à moustaches* sont utiles pour représenter les données statistiques de positions liées à un échantillon. Dans ce cas, la boîte représente l'intervalle inter-quartile, la barre dans la boîte représente la médiane, les moustaches représentent le minimum et le maximum (parfois le premier et le neuvième décile, parfois le 5ème et 95-ème centile).