- Statistiques -

Principe

Si je lance un dé à 6 faces parfaitement équilibré, j'ai théoriquement 1 chance sur 6 de tomber sur une face choisie à l'avance. Cela paraît évident, mais dans les faits, rien ne le prouve : amusez vous à lancer ce même dé 6000 fois, la chance que vous tombiez 1000 fois exactement sur la face que vous aviez choisie est infime. Il paraît donc fort hasardeux, voire dangereux de bâtir des théories (ici celle des jeux) sur des 'ça paraît évident'.

Tout le fondement des statistiques réside en ce point : répéter un certain nombre de fois une même expérience, sonder une quantité plus ou moins importante d'individus d'une même population afin de tenter de valider une théorie par la pratique.

L'essentiel du cours

VOCABULAIRE

Une série statistique est un échantillon de valeurs observées dans une population donnée. Par exemple, on pourra observer les notes à un devoir dans une classe de seconde, les âges dans une entreprise, les tailles ou les poids dans une famille…

Une population représente l'intégralité des membres concernés par l'étude (par exemple les élèves d'une classe de seconde). Une sous population représente les membres concernés par un caractère donné (par exemple les élèves de cette classe ayant eu la moyenne à un contrôle).

Un effectif est la taille d'une population, un sous effectif la taille d'une sous population.

Une fréquence est le taux d'apparition d'un caractère donné dans une population. Elle est donnée par $fréquence = \frac{nombre\ d'occurences\ du\ carctère}{effectif\ total}$

Un pourcentage est une fréquence ramenée en base 100. Il est donné par $pourcentage = fréquence \times 100$. Son unité est le $\%$

L'étendue d'une série statistique est l'écart entre ses deux valeurs extrêmes (maximum et minimum)

Un caractère statistique est la dénomination d'une sous population. Il peut s'agir d'un nombre (l'âge, la note à un contrôle…) ou d'une quantité non chiffrée (la couleur des yeux, la marque des vêtements…)

Un caractère statistique est dit discret lorsque les observations successives aboutissent à une mesure chiffrée finie. Par exemple : la taille de cet individu est de 1.75m

Un caractère statistique est dit continu lorsque les observations successives aboutissent à une mesure chiffrée approximative, c'est-à-dire comprise dans un intervalle. Par exemple, la taille de cet individu est comprise entre 1.70m et 1.80m.

Une variable statistique est dite quantitative lorsque la mesure est effectuée sur une quantité chiffrée. Par exemple, la taille, l'âge, le nombre de cigarettes fumées quotidiennement…

Une variable statistique est dite qualitative lorsque la mesure est effectuée sur une quantité informative : par exemple la couleur de cheveux, la nationalité, le prénom…

MOYENNE ARITHMETIQUE

- Moyenne d'une série statistique

Soit une série statistique composée de n données $x_1, x_2, \dots, x_n$

La moyenne de cette série vaudra alors :

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$

Remarque : on pourra noter $\sum_{i=1}^{n} x_i$ *la somme* $x_1 + x_2 + \dots + x_n$

- Moyenne pondérée

Soit une série statistique composée de n données $x_1, x_2, \dots, x_n$ chacune pondérée par une valeur $n_1, n_2, \dots, n_n$

La moyenne de cette série vaudra alors :

$$\frac{n_1 \times x_1 + n_2 \times x_2 + \dots + n_n \times x_n}{n_1 + n_2 + \cdots + n_n}$$

- Moyenne élaguée

Il est possible que pour une raison ou une autre (erreur de mesure, valeur aberrante…) l'on ne souhaite pas prendre en compte une ou plusieurs valeurs dans le calcul de notre moyenne : on appellera cette nouvelle moyenne la moyenne élaguée. Pour la déterminer, il suffit d'enlever les valeurs problématiques du calcul ci-dessus.

Propriétés :

- Si l'on ajoute une même quantité q à chacune des valeurs de la série statistique, sa moyenne augmente également de q. La moyenne de $x_1 + q, x_2 + q, \dots, x_n + q$ est $$\frac{x_1 + x_2 + \dots + x_n}{n} + q$$
- Si l'on multiplie par une même quantité q chacune des valeurs de la série statistique, sa moyenne est multipliée par q. La moyenne de $x_1 \times q, x_2 \times q, \dots, x_n \times q$ est $$\frac{x_1 + x_2 + \dots + x_n}{n} \times q$$

Remarques :

- *La moyenne arithmétique d'une série statistique est un indicateur de position. Elle sera notée* $\bar{x}$
- *La moyenne est une variable nécessairement quantitative. Lorsque la variable sera ordinale (ou qualitative), on lui préfèrera la médiane.*
- *Il existe d'autres types de moyennes, comme la moyenne harmonique ou la moyenne géométrique*

MEDIANE ET QUARTILES

- Médiane

La médiane d'une série statistique est la valeur du caractère qui sépare la population étudiée en deux groupes de même effectif, c'est-à-dire de tailles identiques.

- Quartiles

Le premier quartile d'une série statistique est la valeur du caractère qui sépare la population étudiée en un groupe représentant 25% de l'effectif et un autre en représentant 75%

Le troisième quartile d'une série statistique est la valeur du caractère qui sépare la population étudiée en un groupe représentant 75% de l'effectif et un autre en représentant 25%

L'écart (ou intervalle) interquartile est l'intervalle d'extrémités le premier et le troisième quartile.

Propriété :

- Si l'on ajoute une même quantité q à chacune des valeurs de la série statistique, sa médiane, son premier et son troisième quartile augmentent également de q

Remarques :

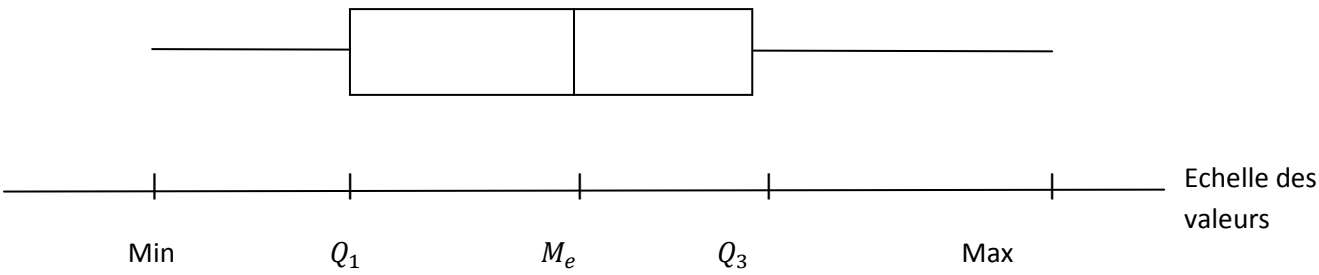
- *On notera que la médiane, le premier quartile et le troisième quartile séparent la population en 4 groupes de même effectif*
- *Il existe d'autres quantités permettant de scinder un effectif en un certain nombre de parts égales, telles les déciles (10 parts) ou bien les centiles (100 parts), amenant chacune un peu plus de précision encore à l'étude.*

DIAGRAMME EN BOITE

Un diagramme en boîte permet de visualiser instantanément les informations suivantes :

- médiane
- quartiles
- minimum et maximum.

Il sera toujours présenté sous cette forme :



Tout l'intérêt de la réalisation de ce genre de diagramme est de pouvoir les comparer entre eux, c'est-à-dire comparer les distributions de caractères dans deux populations distinctes.

ECHANTILLONNAGE ET INTERVALLE DE CONFIANCE

Soit p la proportion d'un caractère dans une population donnée (dite fréquence théorique)

On prélève un échantillon de taille n et on note f la fréquence du même caractère dans l'échantillon (dite fréquence observée).

Alors, dans au moins 95% des cas, f appartient à l'intervalle

$$I = [\, p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \,]$$

I est appelé intervalle de fluctuation (ou de confiance) au seuil 95%

Remarque : En règle générale, on réservera l'utilisation de cet intervalle à des cas tels que $n \geq 25$ *et* $0.2 \leq p \leq 0.8$