TICE Probabilités et statistiques 1 : Mémo de statistiques

Les statistiques descriptives ont pour but de synthétiser un grand nombre de données numériques de même type, de sorte à en tirer un maximum d'informations pertinentes. Il s'agit d'optimiser le nombre de renseignements (qu'on veut minimal) avec la quantité d'informations exploitables à partir de ces renseignements (qu'on souhaite maximale).

1 La série de données.

1.1 Nature des données-données brutes.

On s'intéresse dans cette partie au traitement des informations fournies par n valeurs, notées $(x_1, x_2, .., x_n)$ qui sont le résultat de n mesures qu'on suppose

- **imprédictibles :** chacune de ces valeurs est un résultat variable, non déterminé à l'avance.
- **identiques :** toutes les mesures sont réalisées selon le même protocole.
- **indépendantes :** les différentes mesures n'inter-agissent pas entre elles.

On appelle alors série statistique de taille n la liste des valeurs $(x_1, ., , x_n)$: il s'agit des **données brutes**.

1.2 Les types de données.

Les séries de données se présentent sous différents types :

- **Quantitatif ou qualitatif :** la série est quantitative si les x_i sont des nombres réels. Sinon, on parle de données qualitatives.
- **Discret ou continu :** La série est discrète si les valeurs possibles des mesures sont en nombre fini ou dénombrable (souvent entières). Sinon, on parle de données continues.
- **Exhaustif ou échantillonné :** la série est exhaustive si les données x_i décrivent tous les cas possibles de l'étude. Sinon on parlera de série échantillonnée. Dans cette partie uniquement on supposera que les données sont exhaustives.

1.3 Les fréquences et les classes.

- **Fréquence d'une valeur ponctuelle :** Pour toute valeur k d'une série $(x_1, .., x_n)$, la fréquence de k est la proportion du nombre de valeurs de la série égales à k. Elle est donnée par
$$f_k = \frac{1}{n} Cardi \leq n, x_i = k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{k\}}(x_i), \text{ où } \mathbf{1}_{\{k\}}(x) = 1 \text{ si } x = k \text{ et 0 sinon.}$$
Ces fréquences peuvent être définies pour tous les types de données.
- **Une classe** d'une série de données numériques est un intervalle (semi-ouvert) contenant des valeurs de la série. **L'effectif de la classe** est le nombre de valeurs de la série dans la classe et **la fréquence de la classe** est la proportion du nombre de valeurs de la série dans la classe. **Le centre de la classe** est le milieu de cet intervalle.

1.4 Organisation des données

La série échantillonnée : souvent, les indices n'apportent pas d'information particulière autre qu'un numéro de mesure arbitraire. Dans ce cas, et lorsque les données sont quantitatives, l'étude statistique de la série se ramène à celle de la série triée. La série échantillonnée est la donnée des valeurs distinctes obtenue, affectée de leur effectif ou de leur fréquence. **Attention aux notations !** En effet, pour bien faire il faudrait changer de notation pour les valeurs. Si la série brute s'écrit $(x_1, ..., x_n)$, la série échantillonnée devrait se noter $((v_1, n_1), (v_2, n_2), ..., (v_p, n_p))$. En seconde l'exercice est délicat...

Cas particulier des séries chronologiques : Lorsque les indices réfèrent à une échelle temporelle, on parle de série chronologique. Alors, le tri de la série entraîne une perte d'information. Ce type de données nécessite une étude un peu différente (voir la partie séries statistiques à 2 variables).

2 Synthèse : indicateurs statistiques et représentations graphiques.

Les indicateurs servent à résumer la liste de données. Ils ont plusieurs manières synthétiques de se lire graphiquement. Ils sont plus nombreux pour des données quantitatives : dans ce cas on suppose que la liste est ordonnée.

2.1 Caractéristiques de position.

Ils donnent la tendance "centrale" des valeurs de la série. Seul le mode est pertinent pour des données qualitatives.

- Le(s) **mode(s)** (ou classe modale) donne la ou les valeurs les plus fréquentes (ou l'intervalle de valeurs). Il est facile à calculer, mais apporte peu d'informations.

— La **médiane** est une valeur qui découpe une série numérique en 2 parties de même effectif (en quelque sorte, c'est le"milieu" de la série). Elle est définie par $x_{\frac{1}{2}(n+1)}$ si n est impair, et $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$ sinon. Son avantage est d'être assez stable par rapport aux fluctuations des valeurs de la série (notamment par rapport aux valeurs extrêmes), mais elle est difficile à manipuler. Lorsque n est pair, on parle aussi d'intervalle des médianes $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$).
— La **moyenne** de la série est sa moyenne arithmétique $\bar{x} = \frac{1}{n}(\sum x_i)$. Elle constitue un bon outil mathématique (elle est linéaire), mais elle est sensible aux valeurs extrêmes, ce n'est donc pas toujours un indicateur pertinent (en ce qui concerne les salaires en particulier).

2.2 Caractéristiques de dispersion.

Ceux-ci rendent compte des fluctuations des valeurs de la série autour d'un indicateur de position.
— L'**étendue** de la série est associée au mode, c'est $x_n - x_1$. (On rappelle que la série est triée).
— L'**intervalle interquartile (IQ)** est associé à la médiane, c'est l'intervalle autour de la médiane qui contient la moitié des valeurs (il exclut les quarts extrêmes). Il peut être disymétrique (et indique dans ce cas une différence de comportement entre les petites et les grandes valeurs de la série). Les quartiles $Q1$ et $Q3$ délimitent les extrêmités de la boite à moustache (voir paragraphe suivant). Ils délimitent le premier et le dernier quart de la série.
— L'**écart-type**, σ_x, est associé à la moyenne. C'est un terme d'erreur (voir la variance).
— La **variance** représente la moyenne des carrés des fluctuations de la série par rapport à sa moyenne. Elle mesure l'erreur quadratique moyenne. C'est aussi le carré de l'écart-type :

$$V_x = \sigma_x^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2.$$

— **L'écart-type de l'échantillon** est un indicateur de dispersion qui intervient lors de l'interprétation des données comme celle d'un échantillon (voir paragraphe statistiques inférentielles). Sa valeur est donnée par $s_x^2 = \frac{1}{n-1}(\sum (x_i - \bar{x})^2)$.

2.3 Représentations graphiques

— Le **diagramme en barres** ou bâtons est utilisé dans le cas des valeurs discrètes ou qualitatives. Il représente en abscisse les valeurs de la série, en ordonnée les effectifs ou les fréquences de chaque valeur (dans ce dernier cas, on dit qu'il est relatif). Ce diagramme est pertinent pour un petit nombre de valeurs dans la série.
— Le **diagramme en boîte** ou boîte à moustaches représente visuellement la médiane, l'intervalle interquartiles, ainsi que le premier et dernier décile et/ou l'étendue. Son avantage est d'être extrêmement concis, mais il est peu précis.
— L'**histogramme** des fréquences par classe (obtenu en regroupant les valeurs par intervalles de même longueur ou non) est utilisé dans le cas de données continues, ou lorsque les valeurs sont nombreuses. Il est aussi précis qu'on veut, en fonction du choix de la taille des classes. Lorsqu'il est normalisé, c'est à dire d'aire 1, il a une interprétation probabiliste.

2.4 Interprétation probabiliste, cas d'une série de données exhaustives.

Si les données sont le résultat d'une étude exhaustive sur l'ensemble de la population (comme les données sur les salaires ou les âges par exemple), du point de vue probabiliste la série $(x_1, .., x_n)$ décrit **l'ensemble des valeurs possibles** de la quantité étudiée. Alors on peut associer à la série une variable aléatoire X dont la répartition (ou la loi) est donnée par les fréquences de la série (comme la pyramide des âges représente la loi de l'âge d'une personne choisie au hasard). La moyenne et l'écart-type de la série correspondent exactement à l'espérance et à l'écart-type de la variable aléatoire X.

3 Cas des données numériques couplées. Régression linéaire.

On parle de données couplées lorsqu'on réalise 2 mesures associées à un même individu choisi dans un échantillon de taille donnée. Par exemple des données *(taille, poids); (taille du père, taille du fils), (année, pluviométrie)*, etc.. En géneral dans ce cas, on cherche à étudier l'exitence d'une relation ou d'un lien entre les 2 mesures, le but étant de permettre de "prédire" la seconde lorsque la première est prévisible ou connue. On se donne une série de 2 mesures numériques $(x_1, y_1), (x_2, y_2)..(x_n, y_n)$.

3.1 Représentation graphique.

— Le **nuage de points** est la représentation graphique de la série qui s'obtient en interprêtant les mesures comme des coordonnées. On peut remarquer que les indices n'ont pas d'importance.
— Le **point moyen** $(\bar{x}, \bar{y})$ est le centre de gravité du nuage de points, il correspond à la valeur moyenne de la série couplée.
— **Le choix de l'abscisse est important :** il donne une position préfentielle à la première série. Ce choix correspond à la série la mieux connue. Par exemple, on choisira en abscisse la série temporelle si elle est présente, ou la mesure définie antérieurement à la seconde lorsque c'est possible. Parfois le choix est arbitraire comme le couple *(taille,poids)*.

3.2 Covariance de la série et coefficient de corrélation :

— La **covariance** de la série couplée mesure l'écart moyen à la position moyenne du nuage et est définie par

$$Cov_{x,y} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Elle est un paramètre qui rend compte du "lien" entre les 2 séries numériques.
La covariance de la série est nulle dès que le nuage est symétrique par rapport à l'un des axes passant par le point moyen. Attention cependant : les variables peuvent Ãªtres liées sans Ãªtres corrélées !

— Le **coefficient de corrélation** est la renormalisation de la covariance. C'est un nombre sans dimension compris entre -1 et 1. Il est défini par $r = \frac{Cov_{x,y}}{\sigma_x \sigma_y}$.

3.3 Forme du nuage, corrélation et ajustements affines :

— **Les données ne sont pas corrélées** lorsque le coefficient de corrélation est nul.
C'est le cas lorsqu'il n'y a pas de relation entre les 2 mesures. Les points se répartissent alors de façon homogène dans un rectangle limité par les valeurs extêmes de chacune des mesures.

— **La forme du nuage est lié au coefficient de corrélation**. Il donne donc des informations sur les liens entres les 2 mesures :
• Plus la corrélation est proche de 1, plus le nuage est allongé le long d'une droite à pente positive.
• Plus la corrélation est proche de -1, plus le nuage s'accumule sur une droite à pente négative.
Dans ces conditions, cela signifie que la mesure y est proche d'une fonction affine de x.

— Un **ajustement affine du nuage** (ou de y en fonction de x) consiste à assimiler celui-ci à une droite. Cela revient à remplacer les ordonnées (y_i) par des valeurs du type $(ax_i + b)$, où a et b peuvent être choisis selon différentes méthodes bien précises.
Il y a de nombreux ajustements affines possibles pour un nuage de point donné.
L'ajustement affine est envisagé lorsque la corrélation est forte (disons supérieure à 80%). L'étude de la pertinence de ces ajustements est difficile et utilise des modèles mathématiques complexes.

3.4 Droite des moindres carrés.

Il y a beaucoup de méthodes pour approcher un nuage de points par une droite. Parmi celles-ci, l'une d'entre elles consiste à choisir la droite qui minimise les carrés des écarts entre les points du nuage et leur projetés verticaux sur la droite :

Théorème-Définition :
Si $(x_i, y_i)_{i\leq n}$ est une série numérique couplée, alors il existe une unique droite d'équation $y = ax + b$ telle que $\sum_1^n (y_i - (ax_i + b))^2$ soit minimale.
*Cette droite est appelée **droite des moindres carrés** de y par rapport à x.*

— Cette droite passe par **le point moyen** $(\overline{x}, \overline{y})$,

— Son **coefficient directeur** vaut $a = \frac{Cov_{x,y}}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$.

— Les **résidus** représentent les écarts entre les points du nuage et leur projection sur la droite d'ajustement. C'est la série des erreurs $(y_i - (ax_i + b))_{i\leq n}$.

— Le **résidu quadratique** désigne la somme des carrés des écarts s'appelle aussi parfois le résidu quadratique. C'est un terme d'erreur (au carré). On peut montrer qu'il vaut $nV_y(1 - r^2)$.

— Cette méthode s'appelle **régression linéaire de y par rapport à x par la méthode des moindres carrés**.

3.5 D'autres types d'ajustement.

Parfois, la forme du nuage suit la forme d'une courbe qui n'est pas une droite : dans ces conditions, on peut être amené à choisir un autre type d'ajustement que l'ajustement affine. Cela revient à remplacer les points du nuage par des points de la forme $(x_i, f(x_i))$, où la fonction f sera choisie du mieux possible. Les techniques décrivant le type de fonction choisie et ses paramètres sont nombreuses. L'une d'entre elle consiste à se ramener par changement de variable en x à une regression linéaire par la méthode des moindres carrés avec la nouvelle variable.

3.6 Corrélation, lien et causalité. Interprétation et point de vue.

Lorsque l'étude des données montre une corrélation forte entre les données, on peut raisonnablement affirmer (dans un sens à définir précisément) que les 2 séries sont liées. **Cela ne signifie pas toujours qu'il existe un lien de causalité entre ces 2 mesures.** Par exemple, l'étude a pu omettre une troisième donnée qui influence les 2 autres (on appelle cela une variable cachée), ou bien il peut être délicat de juger quelle donnée découle de l'autre, surtout lorsque celles-ci ne sont pas liées temporellement : ces problèmes relèvent

alors de la discipline à l'origine de l'étude (sociologie par exemple). Les mathématiques et l'étude des données seules ne permettent pas de conclure sur ce point.
On dit parfois que lorsque des variables ne sont pas corrélées, c'est qu'elles n'ont pas de lien. Il s'agit d'une interprétation de données numériques qui doit toujours être prise avec beaucoup de précautions. En effet, du point de vue probabiliste, **indépendance et corrélation nulle ne sont pas équivalentes** : l'indépendance entraine l'absence de corrélation mais le contraire est en général faux. De plus, le problème de l'échantillonnage des mesures dans ce contexte ne permet pas toujours de décider si l'apparition d'une corrélation non nulle mais faible signifie quand même qu'on ne peut pas conclure sur le lien entre les données. Ce problème de représentativité des données devra donc être traité de manière plus élaborée : c'est la motivation principale des statistiques inférentielles.