Analyses bivariées et multivariées

Du tri à plat au tri croisé

Le tri à plat d'un jeu de données (*dataset*) quelle que soit sa source (expérience, récupération de données dans une analyse de data science, sondes, satellites, …), constituant l'analyse univariée (l'examen des variables prises unes à unes) est une première lecture nécessaire des résultats mais elle ne présente pas de véritable intérêt en termes d'analyse.

Les descriptions faites sur les variables soulèvent toute une série de questions sur leurs relations, qui devront être mises en lumière en les rapprochant deux à deux dans des analyses bivariées et l'on pourra, bien entendu, s'intéresser à une analyse multivariée plus poussée, liant une variable à plus d'une autre variable.

Ainsi, les tris croisés (analyses bivariées) permettent d'examiner les relations entre deux ou plusieurs variables. Ces relations peuvent être symétriques et, dans ce cas, l'analyse cherche à mesurer la liaison entre les deux variables et à en tester la signification), ou dissymétriques, dès lors, l'analyse cherche à expliquer les variations d'une variable dépendante (dite aussi variable à expliquer) par les variations d'une variable indépendante (dite également variable explicative). Ce dernier cas fait appel à des méthodes explicatives telles le test d'indépendance du Chi2, l'ANOVA, la régression linéaire simple ou multiple ou encore la régression logistique, pour ne citer que les méthodes les plus classiques. Ces méthodes ont toutes des contraintes, pour le chi-deux : effectif des classes, pour les méthodes quantitatives : normalité des données à interpréter (testée par plusieurs tests à l'aide de logiciels : Kolmogorov-Smirnov, Shapiro-Wilk, Q-Q plot ou Quantile-Quantile plot, droite de Henry,…) ou encore homoscédasticité* (test de Fisher, test de Levene, …). Il conviendra de respecter ces prérequis pour procéder à une analyse sérieuse et pertinente des données au risque, sinon, de s'égarer et de conclure à une relation ou bien à une non relation à tort.

Il ne faudrait surtout pas oublier l'analyse des séries temporelles, n'ayant plus pour objectif d'expliquer les variations d'une variable quantitative par une autre variable mais d'exploiter au mieux l'information temporelle pour trouver le meilleur modèle prédictif avec une question essentielle : quelle sera la valeur de la grandeur étudiée à un temps donnée (prédiction) ou y a-t-il une différence significative entre deux séries temporelles (courbes de survie).

(*) = "qui a une dispersion identique", vient du grec ancien ὁμός (*homós*) qui signifie « semblable », « pareil » et σκεδαση (skedasê) qui signifie « dissipation », « dispersement ». On parle d'homoscédasticité lorsque la variance de la variable dépendante est la même pour toutes les valeurs de la variable indépendante. Ce qui, en régression linéaire, correspond à une uniformité de la variance de l'erreur dans l'ensemble des valeurs observées.

Contexte, buts et outils mis en œuvre en Statistique Bivariée

Entrons un peu plus dans le contexte de la statistique bivariée en nous intéressant à une potentielle relation entre 2 variables (jeu de données typiquement expérimental).

Quatre cas peuvent être identifiés, décrits ci-après.

Dans chacun de ces cas, outre la question de la significativité de la relation à mettre en évidence, il conviendra d'évaluer la part de la variance expliquée par le modèle, qui mesurera la force de cette association.

Il sera important pour établir, outre une liaison statistique, une probable liaison de cause à effet, de distinguer une situation d'observation (*jugement de signification*) d'une situation d'expérience (*jugement de causalité ou d'interprétation*).

- A/ <u>Etude des relations entre deux variables qualitatives</u> ;

L'une d'entre elle peut être la variable dépendante (à expliquer par l'autre : variable indépendante) mais ce n'est pas une nécessité.

Tests mis en œuvre : test du Chi-deux d'indépendance, test exact de Fisher, RR (Risk Ratio) , OR (Odd Ratio) et intervalles de confiance du RR et du OR.

- B/ <u>Etude des relations entre une variable quantitative (dépendante) et une variable qualitative</u>

(La variable quantitative est la variable dépendante et la variable qualitative est la variable indépendante)

> Fondée sur la comparaison de moyennes : Test de Student (paramétrique), Mann-Whitney (non paramétrique), ANOVA (analyse de variance), Kruskal-Wallis (non paramétrique)

> Fondée sur la comparaison de variance : Test de Fisher, test de Hartley

- C/ <u>Etude des relations entre deux variables quantitatives</u>

Si les 2 variables sont toutes deux des variables aléatoires : Tests de Corrélation de Pearson (linéaire, avec contrainte) ou test de Corrélation de Spearman (non nécessairement linéaire, non paramétrique)

Si l'une des variables est indépendante, fixée, (variable explicative, encore appelée régresseur) : Régression linéaire (modèle linéaire simple ou multiple). La variable dépendante doit suivre une distribution normale et la variance être homogène sur tout le nuage de points (homoscédasticité).

- D/ <u>Etude des relations entre une variable qualitative (dépendante) et une variable quantitative</u>

(La variable qualitative est la variable dépendante et la variable quantitative est la variable indépendante)

Classiquement fondé sur le modèle du *logit* : Régression logistique. La variable dépendante est la variable qualitative

Nous allons (re)découvrir et approfondir cela à travers différents exemples choisis dans l'univers de la biologie. Nous terminerons, si le temps nous en donne le loisir, sur l'analyse factorielle ; principalement l'analyse en composante principale (ACP) et l'analyse factorielle des correspondances (AFC).

Bon parcours !