**Data Analyse**

# Cycle de vie de la donnée

Définition d’objectifs 🡪 Collecter 🡪 Nettoyer / Transformer / Contrôler / Harmoniser 🡪 Analyser / Modéliser 🡪 Interpréter / Visualiser / Diffuser

Méthode CRISP : compréhension du problème métier 🡪 compréhension des données 🡪 construction du data hub (nettoyage, recodage, stockage des données) 🡪 modélisation (descriptif, prédictif, prescriptif) 🡪 évaluation (vérification des modèles) 🡪 déploiement (mise en production dans une forme adaptée et intégrée au processus de décision, exemple rmarkdown ou rshiny)

# Types d’analyses

* **Analyse descriptive** : décrire et résumer un ensemble de données. L’analyse exploratoire permet de repérer les corrélations entre variables et potentiellement formuler une hypothèse sur ces relations
* **Analyse inférentielle** : permet de généraliser à l’échelle de la population, des conclusions tirées à partir des données d’un échantillon. L’échantillon doit être représentatif de la population, car ce n’est que sous cette hypothèse que l’on va pouvoir généraliser les conclusions.
* **Analyse prédictive** : utiliser des données historiques et en temps réel, prédiction de tendances pour prévoir les résultats futurs. Elle est basée sur de la modélisation. L’analyse causale permet de déterminer s’il existe une relation entre plusieurs variables et tente d’expliquer cette relation.

# Types de variables

* **Variables quantitatives** : variables numériques **discrètes** (nombre fini de valeurs) ou **continues** (nombre infini de valeurs)
* **Variables qualitatives** : variables par modalité ou facteur (**nominale** ou **ordinale**)

# Statistiques descriptives

## Variables quantitatives

* Moyenne et moyenne pondérée *(R : mean(x) ;weighted.mean(x,w))*
* Médiane : ne dépend pas des valeurs extrêmes de la distribution *(R : median(x))*
* Min et Max
* 1er et 3ème quartiles
* Mesure de dispersion : variance, écart type, intervalle interquartile *(R : var(x) ; sd(x))*

## Variables qualitatives

* Effectifs *(R : table(x))*
* Mode : modalité la plus représentée dans la variable *(R : library(modeest) ; mlv(x, method = « mfv »))*
* Fréquence *(R, library(questionr) ;freq(x))*
* Mesure de dispersion : variance, écart type, intervalle interquartile *(R : var(x) ; sd(x))*

# Statistiques inférentielles

## Données normalisées ou non

Définir sir les données suivent une distribution normale va déterminer la classe de tests réalisables. Pour tester la **normalité des données**, on peut utiliser le **test de shapiro-wilk** (R : rstatix ::shapiro\_test(x$y). Pour les échantillons de très grande taille, le test est trop sensible aux écarts infimes, il vaut mieux alors tester la normalité des données via une visualisation graphique **Densité Plot et QQ plot** (R : library(ggpubr) ; ggdensity(x$y) ; ggqqplot(x$y)).

Si les données sont normales, il faut également tester **l’égalité des variances**, à savoir l’homoscédasticité des résidus. Pour cela, on peut utiliser le **test de Bartlett** (R : bartlett.test(weight ~ group, data = PlantGrowth)) ou le test de Fisher

Si on a vraiment besoin d’avoir des données normales, on peut transformer. Plusieurs types de transformations existent comme la **transformation logarithmique**, la **transformation racine carrée**, ou la **transformation inverse**.

Si les données suivent une loi normale, on peut utiliser des **tests paramétriques**. Dans le cas contraire, on doit utiliser des **tests non paramétriques**.

## Echantillons appariés ou indépendants

Les **échantillons dépendants** sont des mesures appariées pour un ensemble d’éléments. Les **échantillons indépendants** sont des mesures réalisées sur 2 ensembles d’éléments différents.

## Détection de valeurs extrêmes

Tests pour détecter des valeurs extrêmes : Test de Dixon, test de Grubbs (paramétrique) / boxplot / (voir somlit)

## Tests statistiques

|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Hypothèse nulle** | **Tests paramétriques** | **Tests non paramétriques** |
| Comparaison d’une moyenne observée avec une tendance théorique | Moyenne observée = moyenne théorique | **t-test pour un échantillon** | **Test de rang signé de wilcoxon** |
| Comparaison de 2 positions observées (échantillons dépendants) | Les positions sont identiques | **t-test pour échantillons appariés** | **Test de wilcoxon** |
| Comparaison de 2 positions observées (échantillons indépendants) | Les positions sont identiques | **t-test pour échantillons indépendants** | **Test de Mann-Whitney** |
| Comparaison de plusieurs positions observées (échantillons indépendants) | Les positions sont identiques | **ANOVA** | **Test de Kruskal-Wallis** |
| Comparaison de plusieurs positions observées (échantillons dépendants) | Les positions sont identiques | **ANOVA à mesures répétées ; modèles mixtes** | **Test de Friedman** |
| Comparaison de plusieurs mesures binaires (échantillons dépendants) | Les positions sont identiques |  | **Test Q de Cochran** |
| Comparaison d’une proportion observée avec une proportion théorique | Proportion observée = proportion théorique | **Test pour une proportion (khi²)** |  |
| Comparaison de plusieurs proportions observées | Egalité des proportions | **Khi²** |  |
| Comparaison de proportions observées à des proportions théoriques | Proportions observées = proportions théoriques | **Test d’ajustement multinomial** |  |
| Test d’association entre 2 variables qualitatives | Les variables sont indépendantes | **Khi²** | **Test exact de Fisher ; méthode de Monte Carlo** |
| Test d’association entre 2 variables quantitatives | Les variables sont indépendantes | **Corrélation de Pearson** | **Corrélation de Spearman** |
| Comparaison d’une distribution observée à une distribution théorique | les distributions observées et théoriques sont les mêmes |  | **Test de Kolmogorov-Smirnov** |
| Comparaison de 2 distributions observées | Les échantillons suivent la même distribution |  | **Test de Kolmogorov-Smirnov** |

## Analyses descriptives multivariées

**ACP - analyse en composantes principales** : permet d’analyser et de visualiser un jeu de données contenant des individus décrits par plusieurs **variables quantitatives**. Cette analyse est utilisée pour extraire et visualiser les informations importantes contenues dans une table de données multivariées.

**ACM - analyse des correspondances multiples** : permet d’analyser des jeux de données multivariées contenant des **variables qualitatives**, telles que des données d’enquête et de sondage. Cette analyse permet de résumer et visualiser un tableau de données contenant plus de 2 variables catégorielles. C’est une généralisation de l’ACP lorsque les variables à analyser sont catégorielles plutôt que quantitatives. L’ACM est souvent utiliser pour analyser des données d’enquête ou de sondage. L’objectif est d’identifier un groupe de personnes ayant un profil similaire dans leurs réponses aux questions, et d’identifier les associations entre les catégories des variables.

**AFC - analyse factorielle des correspondances** : permet d’analyser l’association entre **2 variables qualitatives**. L’AFC permet de résumer et de visualiser l’information contenue dans le tableau de contingence formé par les 2 variables catégorielles. Le tableau de contingence contient les fréquences formées par les 2 variables.

Il existe d’autres analyses telles que **HCPC** (classification hiérarchique sur composantes principales), **AFM** (analyse factorielle multiple), **GPA** (analyse procustéenne généralisée)

# Statistiques prédictives

Les statistiques prédictives permettent de prédire des données. Selon le type de données à prédire, les algorithmes utilisés sont différents et font appel à des méthodes de régression ou de classification. Pour réaliser des analyses prédictives, il est nécessaire de disposer **de variables prédictives** et d’une **variable dite réponse** (c’est la variable que l’on cherche à prédire). Le principe est de découper ce jeu de données en 2 parties :

* une **partie training** constituée d’environ 70% des données. Le jeu de données training est utilisé pour identifier les variables nécessaires à la prédiction, pour construire le modèle de prédiction et entrainer ces modèles
* une **partie testing** constituée de 30% des données. Le jeu de données testing permet de vérifier les modèles entrainés et de classer ces modèles du plus performant (plus faible erreur de prédiction) au moins performant (plus grande erreur).

## Méthode de classification

Les méthodes de classifications permettent de grouper des objets (observations ou individus) dans des classes (clusters) de manière à ce que les objets appartenant à la même classe soient plus similaires entre eux qu’aux objets appartenant aux autres classes. Exemple : dans le marketing, ces méthodes permettent de détecter différents profils de consommateurs à partir de sondages.

**Classification k-means clustering** (classification par les nuées dynamiques) : méthode de classification non supervisée pour des **variables quantitatives**. Assignement des classes au centroïde le plus proche. Attention, le modèle peut ne pas être robuste si le nombre de variables dans le modèle est important.

**Classification ascendante hiérarchique** (CAH) : sert à définir des classes d’individus à partir d’une ou plusieurs **variables quantitatives**. Le principe s’exprime sous la forme d’une matrice de distances. La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendogramme ou arbre de classification. La classification est ascendante car elle part des observations individuelles. Elle est hiérarchique car elle produit des classes ou groupe de plus en plus vastes, incluant des sous-groupes en leur sein.

Il est possible d’effectuer des classifications des objets caractérises par des variables qualitatives en utilisant les scores des observations des premiers axes sur une ACM, ACP ou AFC.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Méthode** | **Type de variable** | **Nombre de variables** | **Résultat** | **Particularité** |
| **CAH** | Quantitative continue | 1 au moins | Déterministe | Dendogramme, graphique du profil des classes |
| **k-means** | Quantitative continue | 1 au moins | Déterministe | Graphique du profil des classes |

Intervalle de confiance : plage de valeurs qui a une probabilité de 95% de contenir la valeur calculée.

Différence entre corrélation et régression : régression donne une idée d’association entre 2 variables mais ne permet pas de déterminer le degré de cette association. L’intensité de cette liaison s’exprime au moyen du coefficient de corrélation (Pearson).

La corrélation mesure l’intensité de la liaison entre des variables, tandis que la régression analyse la relation d’une variable par rapport à une ou plusieurs autres.

Régression multiple

Regression logistique

Random forest

Définition variance, ecart type, covariance, intervalle de confiance intervalle de prédiction etc

Neural network

Analyse des mots

Tests arbres décisionnels

Quelle condition pour faire une régression

Analyse de la variance ANOVA bivariée, et généralisation multivariée

Analyse discriminante

Analyse canonique des corrélations

Regréssion logistique

Machine learning :

Evaluer des données nouvelles et réaliser une prédiction basée sur les connaissances tirées de données anciennes

Régression linéaire simple : consiste à trouver la meilleure fonction permettant de définir une variable de sortie (l’élément à prédire) à partir d’une seule variable explicative en entrée (le prédicteur). Après création du modèle lm, plus le r.squared est proche de 1, plus le modèle linéaire est bon. Il suffit ensuite d’utiliser la fonction predict() pour prédire les données.

Classification : classer l’information nouvelle dans l’une des catégories des données anciennes