

## Module : Statistique pour Machine Learning

### Séance 2 de travail en groupes

#### Objectifs à atteindre :

- Être en mesure de trier, calculer des variables, recoder, sélectionner des données, etc.
- Être en mesure de préparer et transformer des variables.

L'exploration des données est une étape importante dans l'analyse des données car ça permet de comprendre la structure, la distribution de données et d'identifier les cas atypiques.


#### 1- Structuration de données


Dans cette première partie, vous allez découvrir les différentes méthodes de structuration de données.

Ouvrez le fichier de données **Employés.sav**

Allez dans le menu **Données**, puis cliquez sur Trier les observations...

Vous remarquez dans cette boîte de dialogue, que SPSS permet le tri des observations selon plusieurs clés (sans aucune limitation) et selon l'ordre de tri.

Sélectionnez la variable **sexe** et l'ordre de tri **croissant**, puis cliquez sur le bouton .

Sélectionnez ensuite, la variable **Salaire actuel** et l'ordre de tri **décroissant**, puis cliquez sur le bouton .

Puis cliquez sur le bouton **OK**.

Afin de déterminer le **salaire le plus élevé des femmes**, vous générez un tableau récapitulatif à partir du menu Analyse.

Aller au menu **Analyse** puis sélectionnez **Rapports >Récapitulatif des observations**

Sélectionnez les variables **SEXE** et **SALACT**

Vérifier que les options, **Afficher les observations**, **Limitier les observations aux premières 100** et **Montrer uniquement les observations** valides sont cochées.

Cliquez sur le bouton **OK**.

**Vous notez que le salaire maximum d'une femme est de \$58 125.**

#### 2- Transformation des données

Calculer et recoder sont les deux principales composantes de la transformation de données dans SPSS. La première permet de créer de nouvelles variables selon une expression arithmétique, logique ou conditionnelle. Tandis que la deuxième composante offre la possibilité de remplacer une ancienne valeur ou un groupe d'anciennes valeurs par une nouvelle valeur.

Vous allez maintenant créer deux variables avec la fonction Calculer. La première variable calcule une variable année qui spécifie l'année de naissance de l'employé et la seconde variable calcule la progression du salaire de l'employé selon son expérience passée.


Sur la fenêtre principale de SPSS, cliquez sur le menu **Transformer**, puis sélectionnez **Calculer**.

Dans la boîte de dialogue Calculer, tapez **ANNEE\_NAIS** dans le champ : **Variable destination**.

Allez dans la liste de type de fonctions, et cliquez sur la rubrique **Extraction de dates**.

Cliquez sur la rubrique Extraction de dates, une liste de fonctions relatives aux dates est affichée dans la partie **Fonctions et variables spéciales**.

Sélectionnez la fonction **Xdate.Year** et appuyez sur le bouton .

Puis sélectionnez la variable **Date de naissance** de l'employé et cliquez sur le bouton .

Cliquez sur le bouton **OK**.

Dans la fenêtre de données de SPSS, vous remarquez qu'une nouvelle variable **ANNEE\_NAIS** est créée. Maintenant suivez les étapes suivantes pour créer la variable de la progression du salaire des employés.

Allez dans le menu Transformer et cliquez sur **Calculer**.

Calculer une variable nommée **PROGRESSION** qui est dérivée de l'expression numérique suivante :

## Module : Statistique pour Machine Learning


**PROGRESSION = (Salaire actuel – Salaire d'embauche)/ Expérience passée**

Exécutez cette expression en cliquant sur le bouton OK.

### Recodage de variables pour une discrétisation

La deuxième technique la plus utilisée dans la transformation de données concerne le recodage des variables. Deux options de recodages sont disponibles dans SPSS, le recodage manuel (avec ou sans création d'une nouvelle variable) ou automatique. Vous allez découvrir dans les exercices suivants ces deux techniques.

Dans le menu Transformer, cliquez sur **Recoder > Création de variables...**

Sélectionnez la variable **salaire actuel** (SALACT), et cliquez sur le bouton .

Dans le champ sur Nom de la **variable destination**, tapez **Tranche\_Salaire**, puis cliquez sur le bouton **Valider**.

Cliquez sur le bouton **Anciennes et nouvelles valeurs...**

Dans cette boîte de dialogue de recodage, vous diviserez les anciennes valeurs de salaire les 3 tranches suivantes :

**Tranche 1** : Salaire strictement inférieur à 30 000

**Tranche 2** : Salaire entre 30 000 et 60 000

**Tranche 3** : Salaire supérieur ou égale à 60 000

Dans la boîte de dialogue Création de variables, sélectionnez l'option d'ancienne valeur, **Manquante par défaut ou spécifié**.

Dans la partie de Nouvelle valeur, Tapez la valeur **9**, cliquez sur **Ajouter**.

Sélectionnez l'option Intervalle **à la plus grande**, dans la rubrique **Ancienne valeurs**

Insérer la valeur **60 000** dans le champ devant à la plus grande.

Dans la partie **Nouvelle valeur**, insérer la valeur **3**, puis cliquez sur le bouton Ajouter

Pour deuxième tranche, sélectionnez l'option Intervalle dans la partie Ancienne valeur

Insérer les valeurs **30 000 et 60 000** respectivement dans le premier et deuxième champ

Cette tranche aura la nouvelle valeur **2**, puis cliquez sur Ajouter.

Complétez le recodage de la tranche 1, puis cliquez sur le bouton **Poursuivre** et **OK**.


Mettez à jour le dictionnaire de données, en ajoutant les étiquettes de valeurs de la variable **Tranche\_Salaire**. Ne pas oublier de spécifier la valeur 9 comme valeur manquante dans le dictionnaire de données.

Dans le menu **Analyse**, sélectionnez **Rapports > Récapitulatif des observations...**

Dans la partie Variable, mettez la variable Salaire actuel, et dans la partie Variable(s) de ventilation mettez la variable Tranche\_Salaire.

Désélectionnez l'options Afficher les observations

Cliquez sur le bouton Statistiques...

Sélectionnez les Statistiques Minimum et Maximum et cliquez sur le bouton .


Cliquez sur le bouton Poursuivre puis OK.

Vérifier dans le tableau récapitulatif généré que les bornes de salaires sont respectées.

### Recodage automatique de variables

Le recodage automatique est généralement utilisé pour convertir les valeurs numériques et alphanumériques en entiers consécutifs. Cette transformation conserve les étiquettes des anciennes valeurs. L'exercice suivant recode les valeurs f et m de la variable sexe en nouvelles valeurs 1 et 2.

Allez dans le menu **Transformer** et cliquez sur **Recoder automatiquement...**

Sélectionnez la variable **Sexe de l'employé**, et cliquez sur le bouton .

Entrez dans le champ **Nouveau nom** : **Sexe\_recode**, puis cliquez sur Valider

Cliquez sur le bouton **OK**.

### 3- Manipulation des sélections


## Module : Statistique pour Machine Learning



SPSS offre un large éventail de choix pour sélectionner les observations qui seront utilisées pour les analyses. Les différents types de sélections sont :

- sélection selon une condition logique
- sélection par échantillonnage aléatoire
- sélection dans un intervalle de temps ou d'observations
- sélection selon une variable filtre

Dans le menu **Données**, cliquez sur **Sélectionner les observations...**

Sélectionner l'option **Selon une condition logique**, puis cliquez sur le bouton **Si...**

Dans la liste des variables, sélectionnez la variable **Sexe**, puis cliquez sur le bouton .

Cliquez sur le bouton **EGALE** , saisissez **'f'** et cliquez sur le bouton **ET** .

Sélectionnez ensuite la variable **Salaire Actuel** et cliquez sur le bouton .

Cliquez sur le bouton **EGALE** , saisissez **30 000**

Cliquez sur le bouton **Poursuivre** puis sur le bouton **OK**.

Afin de vérifier que la sélection est effectuée, générez un tableau de fréquence sur la variable sexe, et un tableau descriptif sur la variable salaire actuel.

La sélection d'un échantillon aléatoire s'effectue de la manière suivante :

Dans le menu **Données**, cliquez sur **Sélectionner les observations...**

Cliquez sur l'option **Sélectionner par échantillonnage aléatoire**, puis cliquez sur le bouton **Aperçu...**

Saisissez **25%** de toutes les observations et cliquez sur le bouton **Poursuivre**

Cliquez sur le bouton **OK**.

Décrivez l'échantillon obtenu.

### 4- Quelques analyses descriptives multi-variées

Dans le menu **Analyse**, cliquez sur **Statistiques descriptives>Tableaux croisés...**

Mettez dans la partie **Ligne(s)**, la variable **sexe de l'employé**, et dans la partie **Colonne(s)** la variable **catégorie d'employé**.

Cliquez sur le bouton **OK**.

### 5- Exercices

- Déterminer le sexe le plus fréquent dans cette banque mais sans produire la table détaillée des fréquences. (faites un choix judicieux de la bonne statistique).
- A l'aide d'un histogramme, illustrez la distribution de fréquences de l'âge des employés.
- Déterminer l'âge du plus jeune employé, du plus vieux ainsi que l'âge moyen.
- Calculer le 80<sup>ème</sup> percentile de l'âge et sa médiane.
- Recoder l'âge des employées selon le tableau ci-dessous :

Code-Valeur	Tranche d'âge
1	Moins de 40 ans
2	Entre 40 et 50 ans
3	Entre 50 et 60 ans
4	Plus de 60 ans

- Calculer le pourcentage des employées qui ont moins de 40 ans et qui sont cadre.

Répose :

**Module : Statistique pour Machine Learning**