UNIVERSITE GASTON BERGER
L'excellence au service du développement
UFR des Lettres et Sciences humaines
Section de sociologie

IASS 361
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES
© El Hadj Touré, Ph.D.

LABO SPSS #3
Calcul des variables et analyse par groupes

Y a-t-il une relation entre la scolarité et le revenu ? Pour répondre à cette question de recherche relationnelle, l'analyse bivariée est appropriée. Selon la nature qualitative ou quantitative des deux variables, plusieurs tests d'hypothèses peuvent être utilisés. Le test du chi-carré s'avère approprié pour analyser une relation entre deux variables qualitatives, et les tests de comparaison de moyennes pour analyse une relation entre une VI qualitative et une VD quantitative (test t : 2 groupes et test F d'ANOVA : 3 groupes ou plus). Finalement, la corrélation et la régression linéaires sont indiquées lorsqu'il s'agit d'analyser une relation entre deux variables quantitatives.

Toutefois, l'association statistique n'implique pas une relation de causalité. Une relation entre deux variables ne peut être interprétée comme une relation causale sans contrôler l'effet d'autres variables supplémentaires (Boudon : 2001 ; Lazarsfeld : 1995). Les procédures statistiques qui permettent d'aboutir à une telle conclusion relèvent en grande partie de l'analyse multivariée. Celle-ci consiste à examiner une relation entre une VI et une VD, en neutralisant ou prenant en compte l'effet d'une ou de plusieurs variables-contrôles (VC). *Par exemple, y a-t-il une relation entre la scolarité et le revenu, en contrôlant l'effet du sexe?* Afin d'y répondre, plusieurs techniques statistiques sont disponibles à cet effet : analyse de tableaux multivariés, ANOVA à deux facteurs, régression et corrélation linéaires multiples, corrélation partielle, régression linéaire multiple avec terme d'interaction.

Avant de procéder aux analyses bi-multivariées à l'aide de SPSS, il est nécessaire de préparer les données des variables afin qu'elles se prêtent aux techniques statistiques employées. Tel est l'objet du labo d'aujourd'hui dans lequel nous verrons comment calculer des variables, analyser des variables par groupes. À la fin de ce labo, vous serez en mesure de :

- ✓ Calculer une nouvelle variable à partir d'autres variables (simple, complexe),
- ✓ Sélectionner des observations et procéder à des analyses sur un groupe.
- ✓ Scinder un fichier et procéder à des analyses sur plusieurs groupes.

Des exercices pratiques aideront à atteindre ces objectifs d'apprentissage.

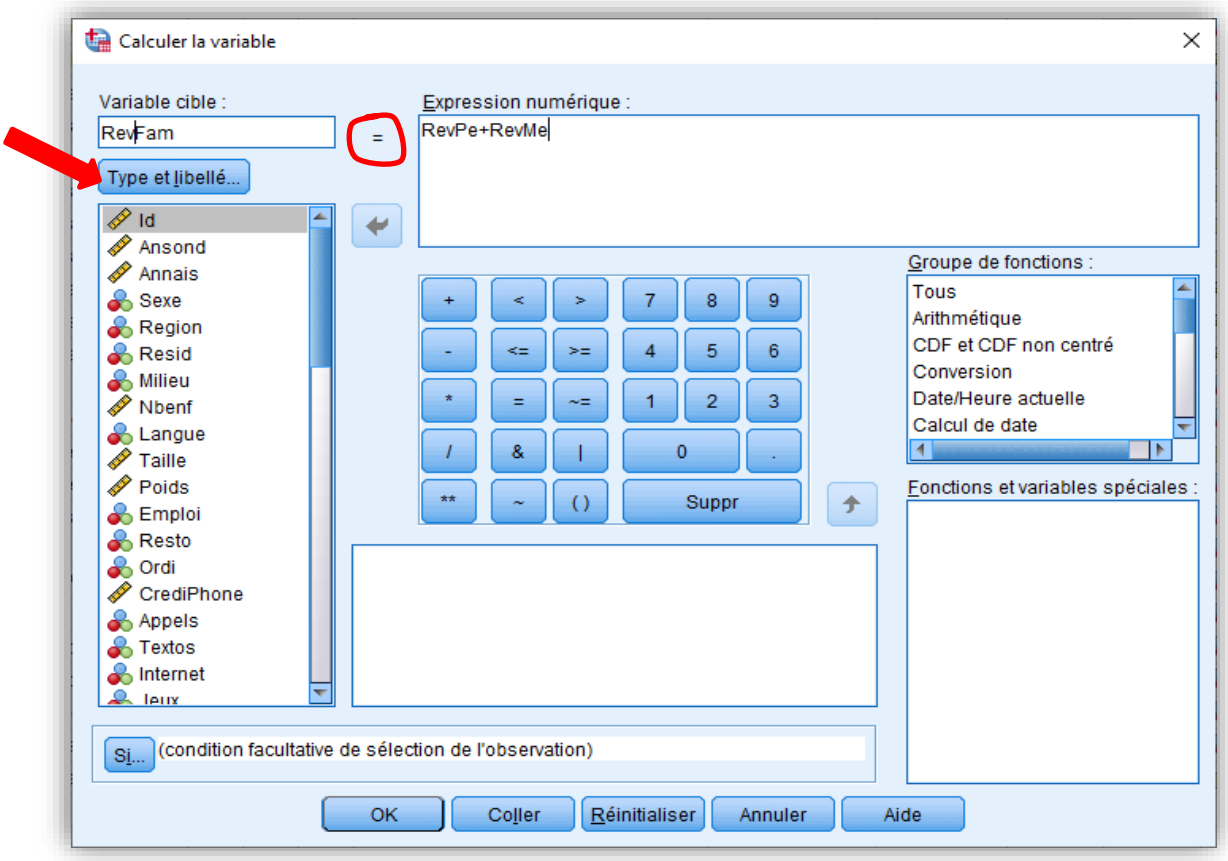
1. Création d'une nouvelle variable avec la fonction « calculer »

La commande « **Calculer la variable** » permet de créer une nouvelle variable, à partir d'une ou de plusieurs variables déjà existantes pour tous les cas.

1.1. Calcul simple du revenu familial

Créons la variable « revenu familial » qui serait la somme du revenu du père (**revpe**) et de celui de la mère (**revme**). Voici la procédure :

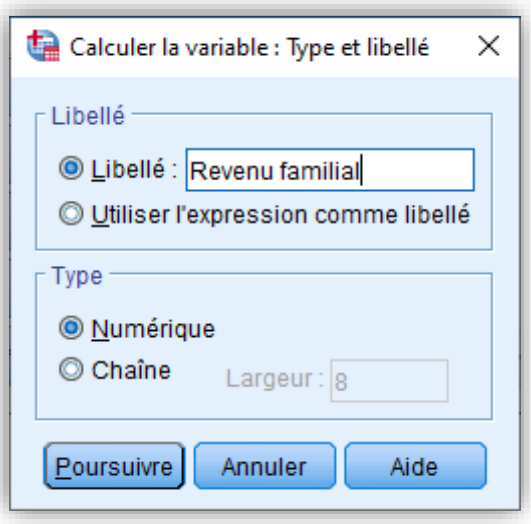
Transformer
Calculer



Sur l'écran de dialogue qui apparaît, inscrire, à l'aide du clavier d'ordinateur, le nom de la nouvelle variable dans le rectangle « Variable cible » situé en haut à gauche de l'écran de dialogue (**RevFam**). Écrire dans le rectangle, à droite, l'expression numérique permettant la création de cette nouvelle variable (vous pouvez soit l'écrire directement grâce au clavier, soit utiliser la zone de dialogue en sélectionnant les variables à sa gauche, les opérateurs et les fonctions en dessous et en cliquant successivement sur les touches correspondant aux données nécessaires à la création de la nouvelle variable). Dans l'exemple du revenu des deux parents, vous sélectionnez la variable « revenu actuel du père » que vous transférez dans le rectangle à l'aide de la flèche, à l'aide de la souris ou du clavier, vous ajoutez le signe « + » et vous sélectionnez la variable « revenu actuel de la mère ». Vous aurez ceci!

RevFam = revpe+revme

Cliquer sur « Type et libellé » (ou Type et étiquette) pour apposer une étiquette à cette nouvelle variable (Revenu familial).



Poursuivez et validez le tout, la nouvelle variable « revenu familial » sera créée!

```
COMPUTE RevFam=RevPe+RevMe.
VARIABLE LABELS  RevFam 'Revenu familial'.
EXECUTE.
```

Nous pouvons calculer la moyenne, la médiane, l'écart-type, les valeurs max et min de la variable **RevFam** en suivant les étapes décrites ci-dessous.

Analyse
Statistiques descriptives
Fréquences
Statistiques
- ✓ **Moyenne**
- ✓ **Médiane**
- ✓ **Écart-type**
- ✓ **Max et Min**

Cliquez dans le carré « **Afficher les tables de fréquences** » afin d'enlever le X qui s'y trouve (s'il s'y trouve). Vous indiquez à SPSS que vous voulez obtenir des statistiques tout en lui précisant que vous ne désirez pas voir apparaître le tableau de fréquences, lequel s'étendrait probablement sur plusieurs pages.

La page des résultats s'affiche!

Statistiques

Revenu familial

N	Valide	66
	Manquant	37
Moyenne		288100.4545
Médiane		125000.0000
Ecart type		444451.4297
Minimum		.00
Maximum		2800000.00

Dans cet exemple, la nouvelle variable comporte beaucoup de valeurs manquantes (37 contre 66 valides). Cela est dû en partie au fait que les valeurs manquantes des variables que l'on utilise pour créer une nouvelle variable s'additionnent et que leur somme est considérée comme manquante.

Interprétation statistique : La moyenne est de 89,47 (en milliers de dollars). *Les familles des étudiants ont un* ***revenu moyen*** *de 89 475 dollars*. La médiane, quant à elle, est de 80 (en milliers de dollars). *Autrement dit,* ***au moins 50%*** *des familles ont un revenu inférieur à 80 000 dollars* ***ou moins***.

La moyenne est de 12% (89,47-80/80) supérieure à la médiane. Ce qui suggère que la distribution est asymétrique positive. La moyenne étant sensiblement gonflée par les revenus élevés des cas déviants, la médiane s'avère plus appropriée.

Interprétation théorique (sociologique): Le revenu médian des parents des étudiants (80 000$) s'avère élevé comparativement au revenu médian des Canadiens, estimé à 69 860$ en 2010 par Statistique Canada. Il semble donc que l'accès aux études universitaires soit relié au statut socioéconomique des parents, comme l'ont déjà montré les sociologues de l'éducation, notamment Bourdieu…

Remarque : La commande « **Calculer** » peut être utilisée à l'aide d'une grande variété d'opérateurs et de fonctions mathématiques, qui peuvent être combinés à volonté pour créer des variables complexes, dont les plus couramment mobilisés sont :

Opérateurs:	**+ ; -**	Addition; soustraction
	*** ; /**	Multiplication; division
	< ; >	Plus petit; plus grand;
	<= ; >=	Plus petit ou égal; plus grand ou égal.
	&	Et
	|	Ou
	**	Exponentiel
Quelques fonctions:	**Abs:**	valeur absolue
	Mean:	moyenne
	Squt:	racine carrée
	Sum:	faire la somme

1.2. Calcul complexe de l'indice de masse corporelle

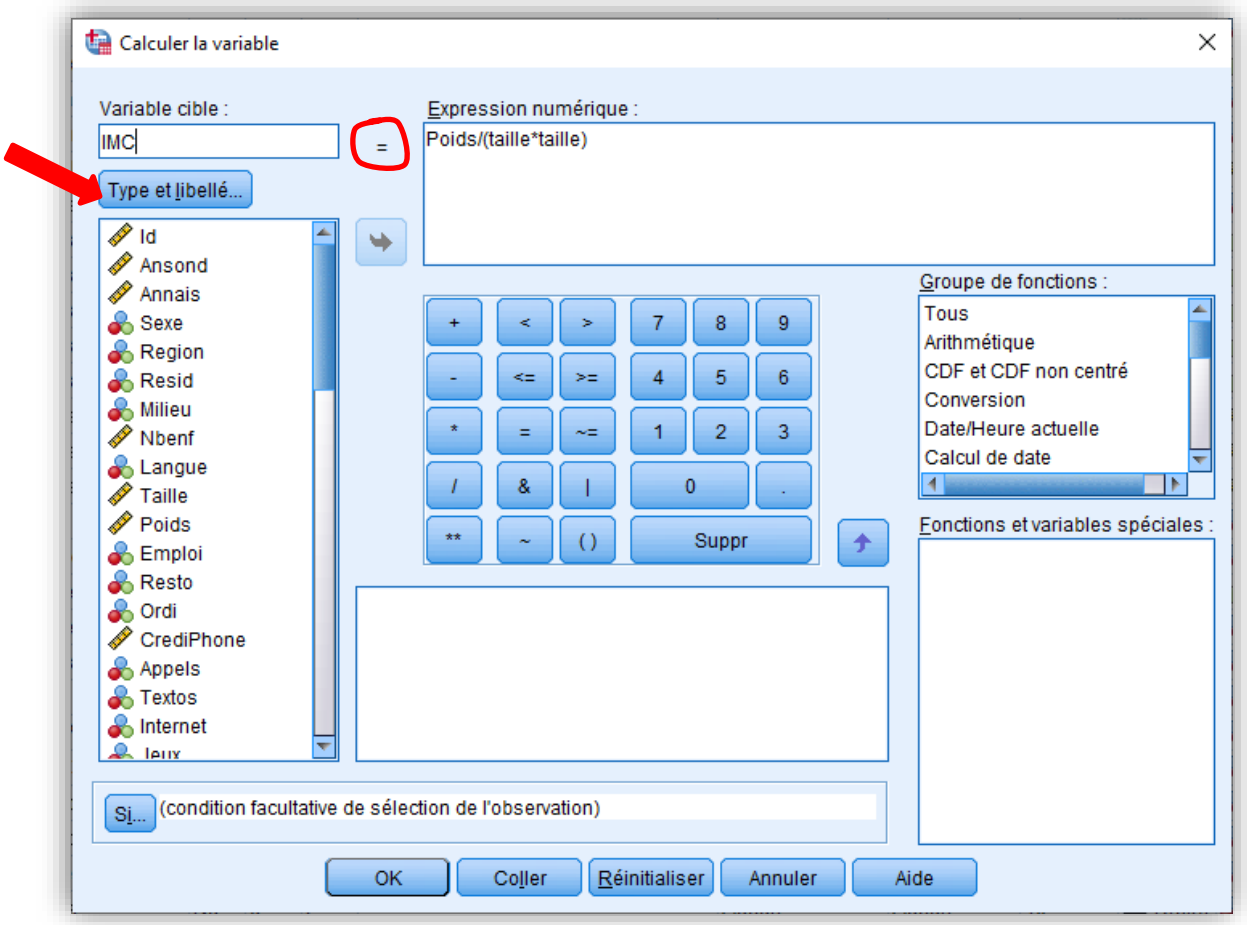
Créons la variable « Indice de masse corporelle » (IMC) à partir des variables Poids et Taille. Cet indice a été créé par Quételet (1835) pour mesurer le phénomène de l'obésité et de la dénutrition afin d'évaluer les risques pour un individu de développer des problèmes de santé pouvant conduire à la mortalité. Il se calcule ainsi [1]:

IMC= Poids(kg) / Taille (m^2)

Avant de procéder au calcul sortez les statistiques descriptives des variables Poids et Taille pour vous assurer que tout est correct… Voici la procédure pour créer l'IMC:

Transformer
Calculer

1 Cette formule découle du constat suivant : le poids évolue proportionnellement au poids au carré.



Sur l'écran de dialogue qui apparaît, inscrire le nom de la nouvelle variable dans le rectangle « Variable cible ». Écrire dans le rectangle, à droite, l'expression numérique permettant la création de cette nouvelle variable. Vous aurez ceci!

IMC = Poids/(Taille*Taille)

Cliquer sur « Type et libellé » (ou Type et étiquette) pour apposer une étiquette à cette nouvelle variable (Indice de masse corporelle). Poursuivez et validez le tout!

```
COMPUTE IMC = Poids/(Taille*Taille).
VARIABLE LABELS  IMC 'Indice de masse corporelle'.
EXECUTE.
```

À des fins de vérification et de validation, sortez les statistiques descriptives de la variable IMC : moyenne, médiane, écart-type, max et min, percentiles.

Statistiques

Indice de masse corporelle

N	Valide	87
	Manquant	16
Moyenne		19.6197
Médiane		19.3906
Ecart type		4.10298
Minimum		10.03
Maximum		32.65
Percentiles	25	16.8450
	50	19.3906
	75	22.0386

Plus l'indice est élevé, plus l'étudiant est obèse. Sur la base de la relation constatée entre l'IMC et le taux de mortalité, l'indice de Quételet s'interprète ainsi :

- Moins de 17=Dénutrition
- 17-18,4= Maigreur
- 18,5-24,9= Corpulence normal
- 25-29,9= Surpoids
- 30-39,9= Obésité modérée à obésité sévère
- 40 et plus =Obésité morbide

Interprétation statistique : L'IMC moyen des étudiants tourne autour de 19.62±4.10. Selon la règle d'interprétation ci-dessus, au moins 25% des étudiants sont en dénutrition (Q1=16.8). Il semble que la plupart des étudiants sont maigres et normaux, le pourcentage d'étudiants en surpoids ou obèses étant négligeable.

Interprétation théorique/sociologique : Comment expliquer le faible score des étudiants sur l'échelle de l'indice de masse corporelle ? Cela peut s'expliquer, d'une part par l'âge très jeune des étudiants, et d'autre part par l'appartenance des Sénégalais au type sahélien, donc souvent mince.

Par ailleurs, même si l'IMC peut aider à identifier la population à risque de dénutrition et d'obésité morbides, il doit être interprété avec prudence. Premièrement, le poids relativisé par la taille ne suffit pas pour mesurer de façon exhaustive l'obésité ou la dénutrition. La mesure du tour de la taille peut conférer une information supplémentaire concernant l'importance de la graisse. Deuxièmement, une personne qui se retrouve dans la catégorie des « surpoids » pourrait être un athlète à la masse musculaire importante. Troisièmement, l'indice de masse corporelle varie selon le sexe, l'âge, l'ethnie. Surtout, il ne peut être utilisé sans prendre en compte les habitudes de vie. Quoique son calcul soit stable, l'IMC n'est pas interprété ou ne peut pas être interprété de la même manière dans tous les pays. On comprend pourquoi aux États-Unis le seuil critique a été relevé alors que chez certains peuples (sénégalais ou asiatiques par exemple) le seuil critique doit être abaissé. Des problèmes de comparabilité internationale résultent nécessairement de l'interprétation de l'IMC.

2. Analyse par groupes d'une base de données

Au lieu d'analyser tous les cas d'une base de données, il est parfois utile de mener des analyses sur un seul groupe d'intérêt sélectionné ou de les scinder selon différents groupes à des fins comparatives. On peut procéder ainsi à l'aide de SPSS.

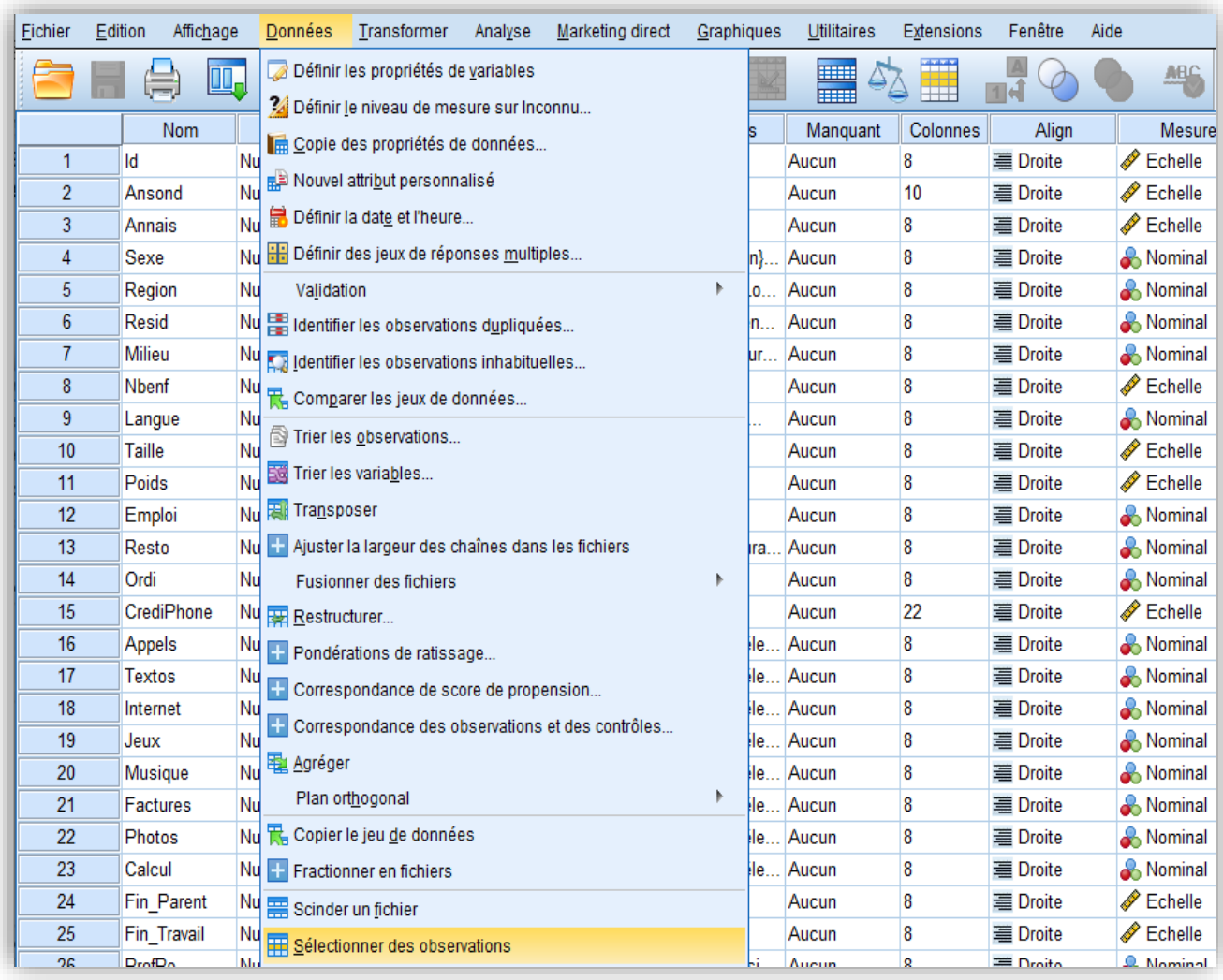
2.1. Sélectionner des observations (cas)

Il est possible d'aller plus loin dans la description des données en sélectionnant des cas (observations) sur lesquels nous souhaitons faire porter l'analyse. Nous pourrions vouloir obtenir la tendance centrale et la variation d'une variable donnée pour seulement une catégorie d'une autre variable. À titre d'exemple, nous pouvons demander l'indice de masse corporelle (IMC) chez les étudiantes seulement (1=femme). Autrement dit, quel est l'IMC typique chez les étudiantes ? Encore, comment varie l'IMC chez les étudiantes ?

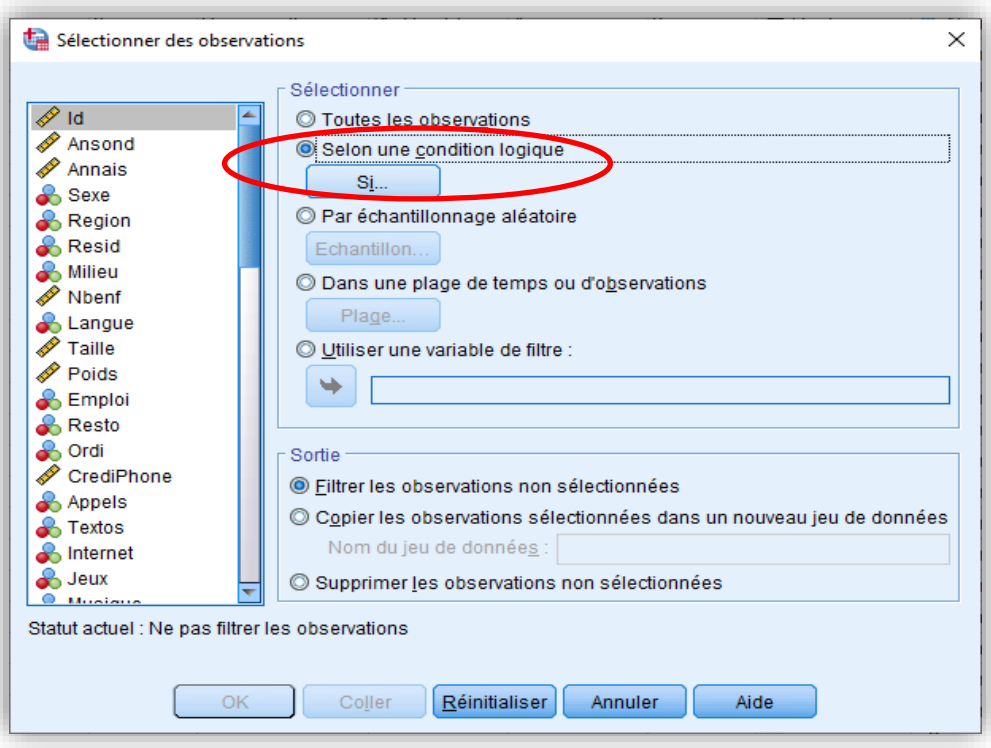
Dans ce genre de situation, la commande « Sélectionner des observations » est indiquée.

Données

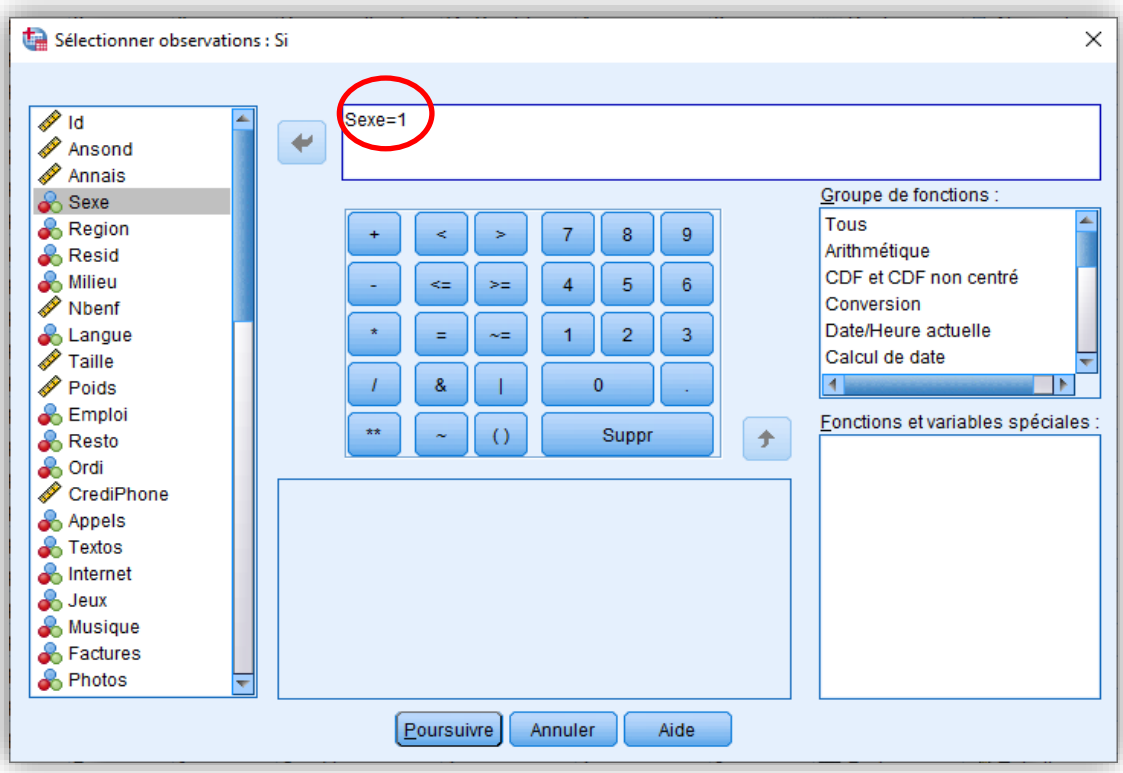
Sélectionner des observations



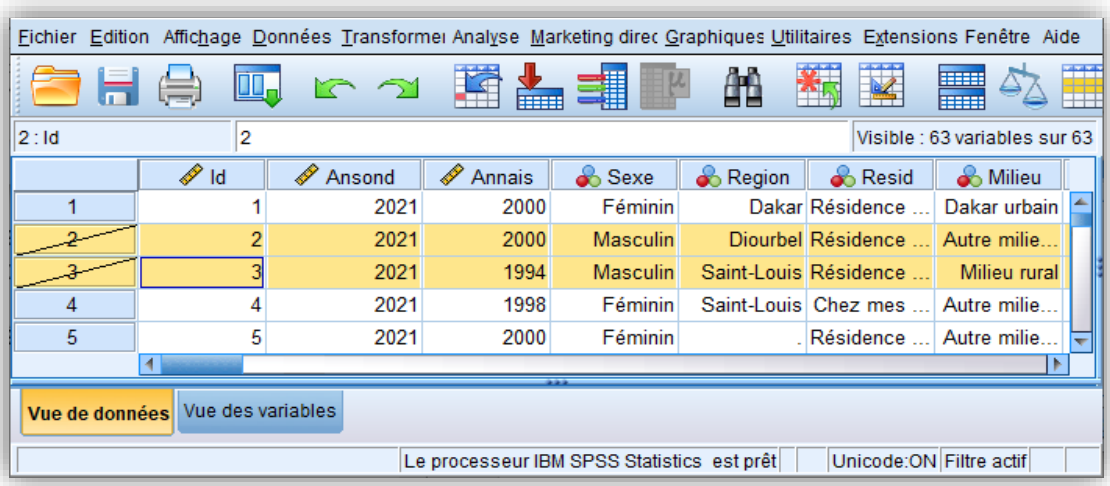
Cliquez sur « Sélectionner des observations » pour obtenir l'écran ci-dessous !



Cliquez sur le bouton radio « Selon une condition logique », puis cliquez sur l'icône « Si ». Un écran apparaît !



Cliquez sur la variable « sexe » pour la faire passer dans le rectangle, puis saisir =1, pour demander au logiciel de sélectionner seulement les étudiantes (1=femme). Poursuivez et validez ! SPSS nous élimine temporairement les étudiants en mettant une ligne oblique sur les cas hommes.



Il ne reste plus qu'à demander les statistiques descriptives concernant les étudiantes (femmes) pour la variable **IMC**. Suivez la procédure :

Analyse

Statistiques descriptives

Fréquences

Statistiques

- ✓ **Moyenne**
- ✓ **Médiane**
- ✓ **Écart-type**
- ✓ **Min et Max**
- ✓ **Quartiles**
  - **IMC**

Statistiques

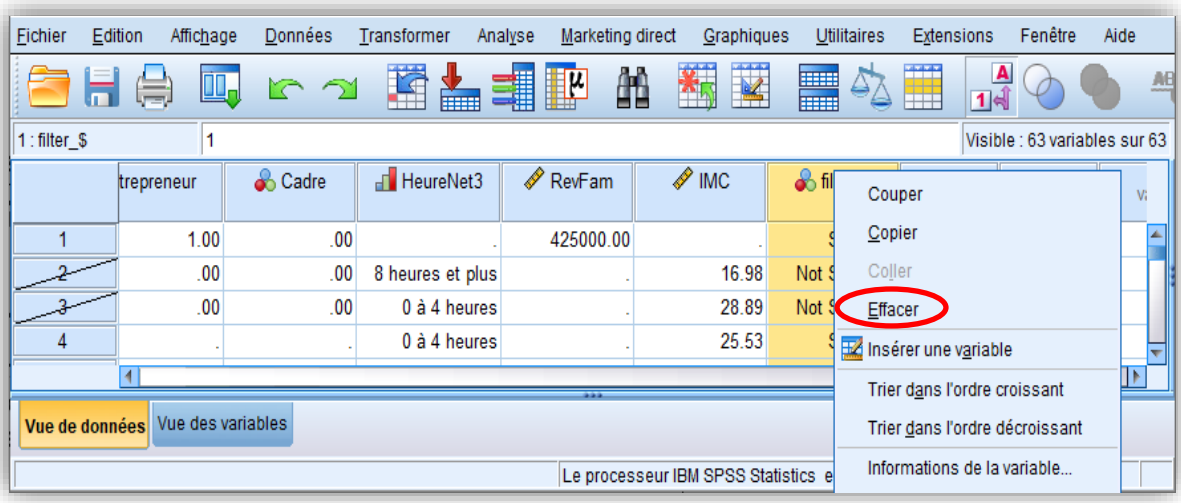
Indice de masse corporelle

N	Valide	51
	Manquant	5
Moyenne		19.2094
Médiane		18.7305
Ecart type		4.29458
Minimum		10.03
Maximum		29.08
Percentiles	25	16.3265
	50	18.7305
	75	22.0386

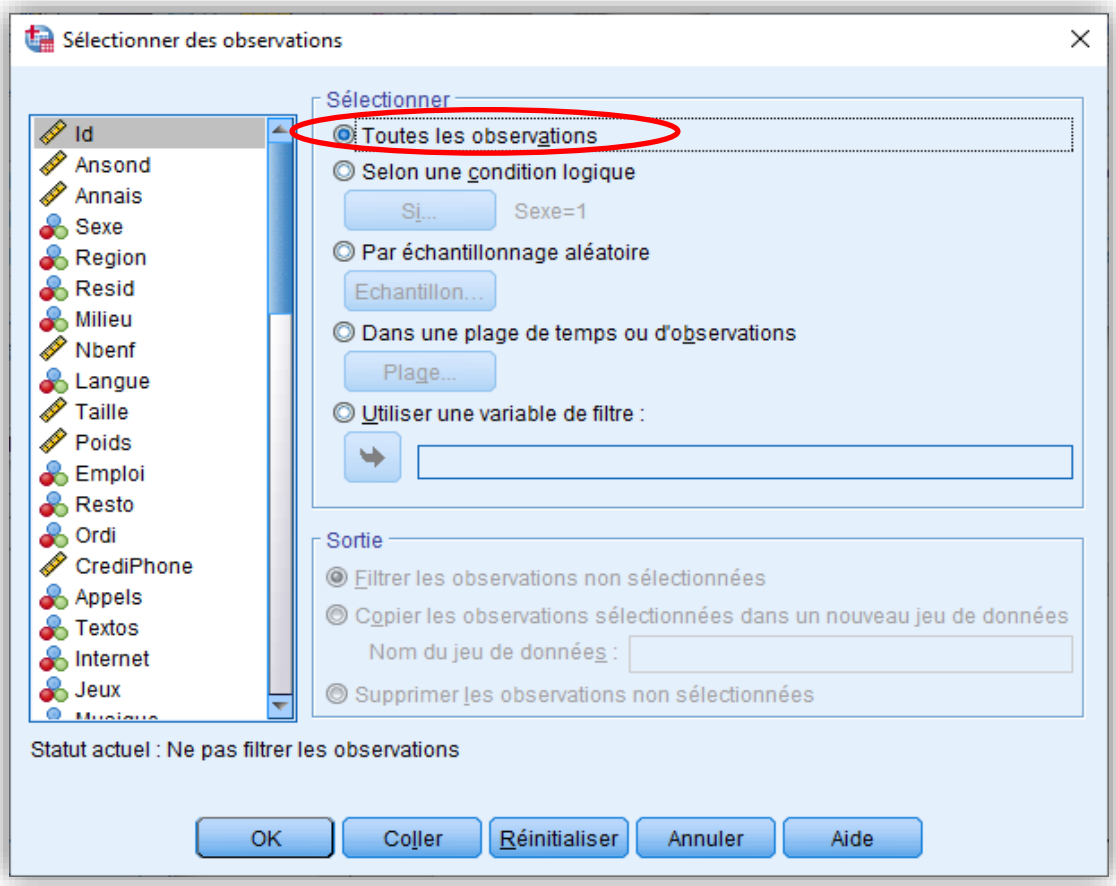
<u>Interprétation statistique (analyse):</u> L'analyse du tableau montre qu'environ 2/3 des femmes ont en moyenne un IMC de 19,21±4,29. La distribution est hétérogène, puisque le coefficient de variation est de 22% (4,29/19,21). L'échelle de l'IMC varie de 10,03 à 29,08.

Attention : Après avoir terminé l'analyse, n'oubliez pas de demander à SPSS de dé-sélectionner les observations. Tant que vous ne l'aurez pas fait, toutes les analyses seront effectuées sur les cas sélectionnés, soit les femmes seulement.

Pour désélectionner ou supprimer le filtre, repérer la variable « filter_$ » à la toute fin sur l'affichage des variables, cliquez dessus, puis faites clic droit et effacer.



Une autre possibilité consiste à désélectionner « Selon une condition logique » et sélectionner « Toutes les observations », tel qu'illustré ci-dessous.

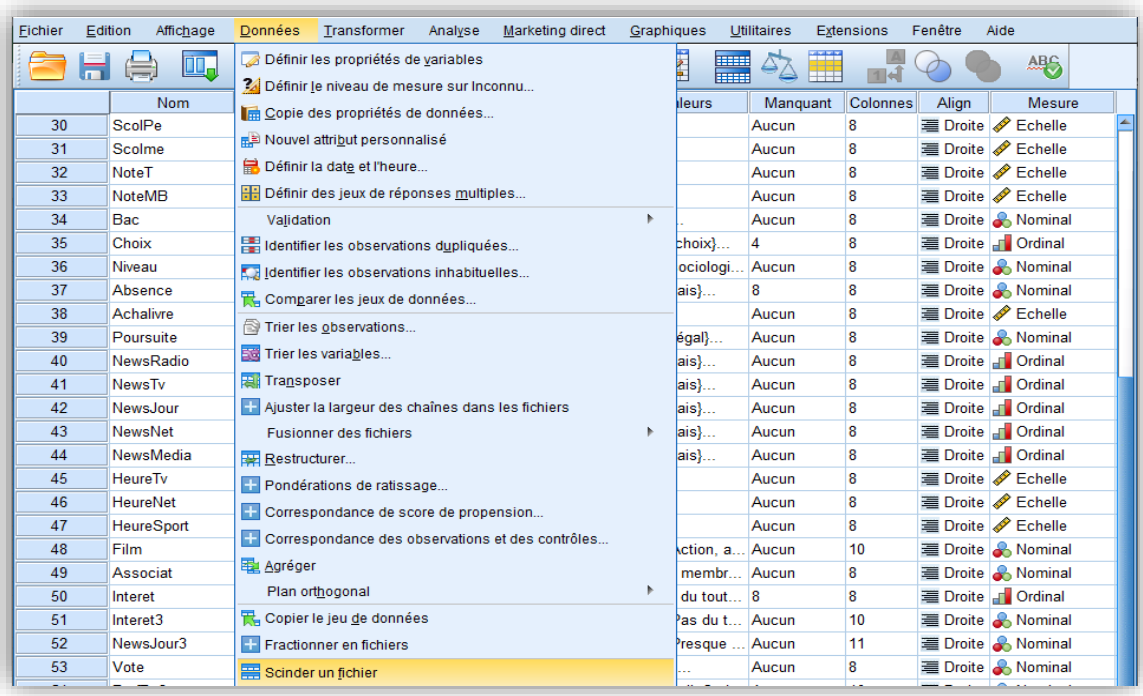


2.2. Scinder un fichier de données

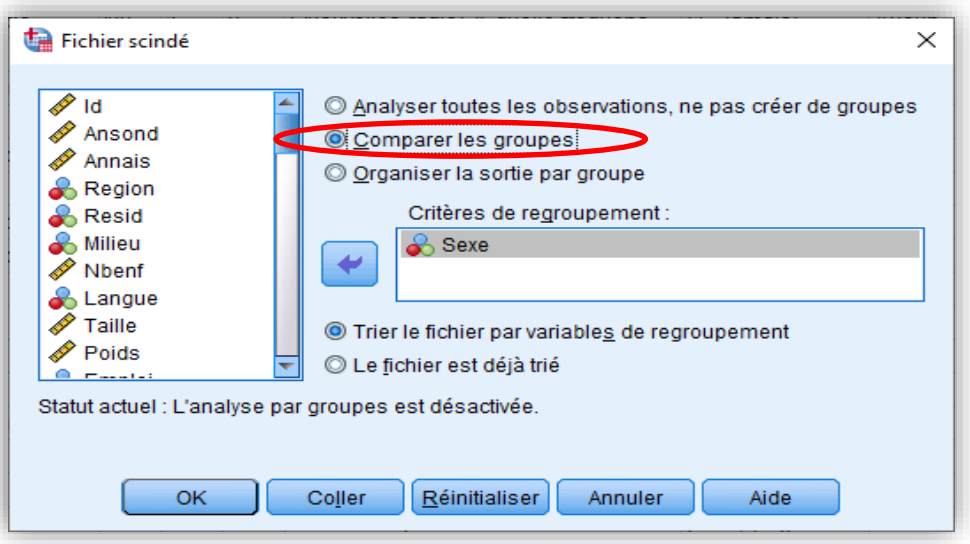
Dans le labo 4, nous avions appris à sélectionner des cas (observations) sur lesquels nous souhaitons faire porter l'analyse. Nous pouvons toujours aller plus loin en comparant des groupes eu égard à une variable d'intérêt. À titre d'exemple, nous pouvons obtenir les mesures de tendance centrale et de variation de l'indice de masse corporelle (**IMC**) selon le sexe (**sexe**) afin de comparer les femmes et les hommes. *Autrement dit, la tendance centrale et variation de l'IMC est-elle plus élevée chez les femmes que chez les hommes ?* Dans ce genre de situation, la commande « **scinder un fichier** » est indiquée afin d'obtenir des résultats comparatifs.

Données

Scinder un fichier



Cliquez sur « Scinder un fichier » pour obtenir l'écran ci-dessous !



Faites passer « sexe » dans le rectangle pour le considérer comme **critère de re-**

groupement. Cliquez sur « Comparer les groupes » pour que les résultats soient présentés dans un tableau compact comparatif. Vous aurez pu également cliquer sur « Organiser la sortie par groupe » (ou Séparer résultats par groupes) pour obtenir séparément des tableaux.

Validez le tout ! SPSS nous indique que le fichier a été scindé en deux.

```
SORT CASES  BY sexe.
SPLIT FILE LAYERED BY sexe.
```

Maintenant, on peut faire calculer la tendance centrale et la variation de l'IMC chez les hommes et les femmes.

Analyse
Statistiques descriptives,
Fréquences
Statistiques
- ✓ **Moyenne**
- ✓ **Médiane**
- ✓ **Écart-type**
- ✓ **Min et Max**
- ✓ **Quartiles**
  - **IMC**

Statistiques

Indice de masse corporelle

Féminin	N	Valide	51
		Manquant	5
	Moyenne		19.2094
	Médiane		18.7305
	Ecart type		4.29458
	Minimum		10.03
	Maximum		29.08
	Percentiles	25	16.3265
		50	18.7305
		75	22.0386
Masculin	N	Valide	36
		Manquant	11
	Moyenne		20.2009
	Médiane		19.6579
	Ecart type		3.79791
	Minimum		14.69
	Maximum		32.65
	Percentiles	25	17.1470
		50	19.6579
		75	21.9877

Le tableau comparatif gagnerait à être reconfiguré de sorte que seules les informations jugées utiles soient retenues.

Tableau 1. Variation de l'IMC selon le sexe chez les étudiants

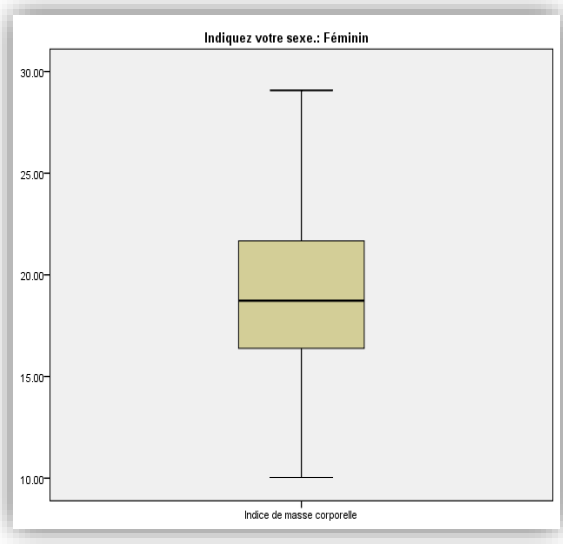
IMC	Sexe	
	Femmes	Hommes
Moyenne	19,2	20,2
Écart-type	4,3	3,8
Nombre de cas (n)	51	36

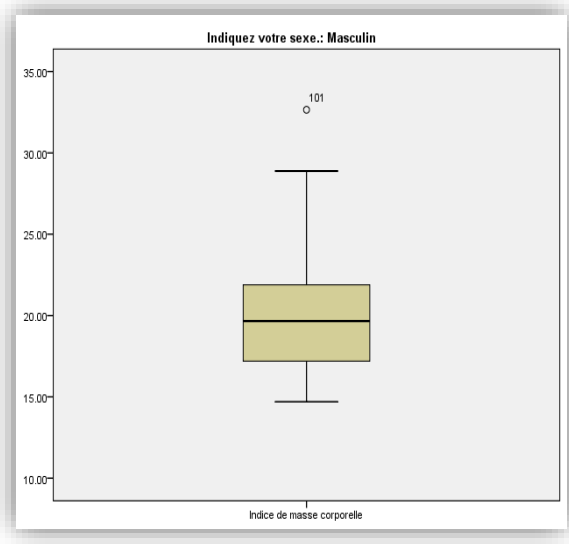
Note : (n=87; non-réponses=16).
Source : Sondage_ÉtudiantsL2_Socio_2021.

Interprétation statistique (analyse): L'analyse du tableau des moyennes montre qu'environ 2/3 des femmes ont en moyenne un IMC de 19,2±4,3, tandis qu'on 20,2±3,8 chez les hommes. La variation de l'IMC semble plus élevée chez les femmes alors que la moyenne des hommes est légèrement plus élevée que celle des femmes. Ces tendances sont observables à travers les boîtes à moustaches ci-dessous.

Pour sortir les boites à moustaches suivez la procédure déjà indiquée :

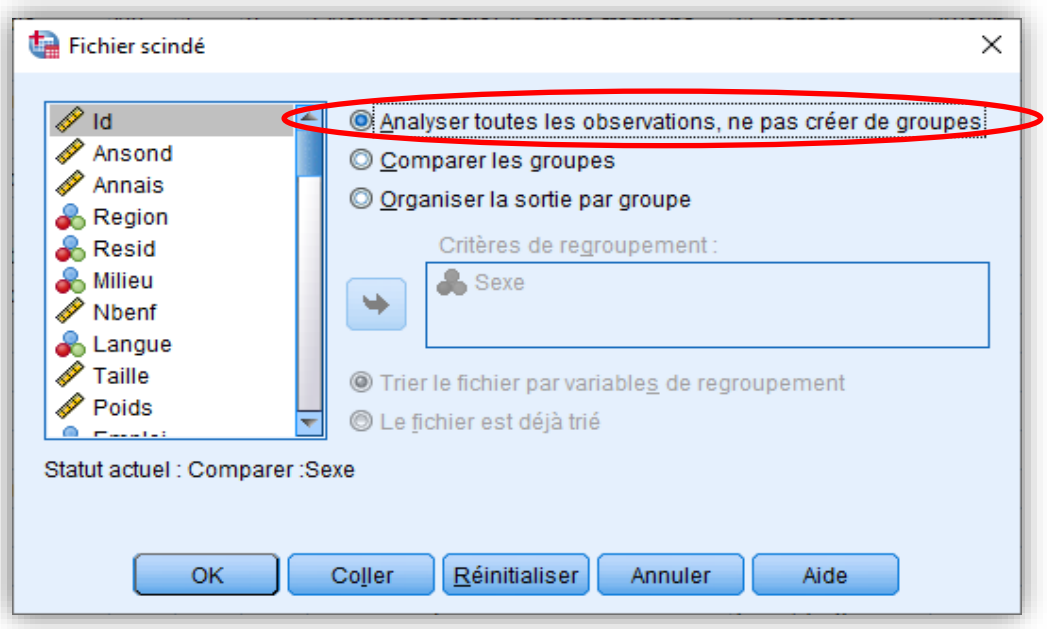
Graphiques
Boîte de dialogue
Boîte à moustaches (Diagramme à surfaces)





Interprétation théorique/sociologique (discussion): Les hommes sont plus sujets au surpoids en raison probablement de leur caractéristique physiologique, qui les pousse à manger plus. À l'opposé, les femmes sont davantage soucieuses de leur poids. Quoi qu'il en soit, les données analysées ici ne semblent pas révéler des différences significatives. Des données plus générales sont nécessaires pour approfondir l'analyse comparative.

Une fois l'analyse comparative terminée, n'oubliez pas de retourner sur **« Données »**, puis « **Scinder un fichier »**, pour reconsidérer toutes les observations et ne pas créer de groupes, tel qu'illustré ci-dessous, sinon toutes les analyses seront groupées.



3. Exercices pratiques

Exercice 1. Transformation d'une variable quantitative (IMC)

À l'aide de SPSS, demandez une distribution de fréquences et de % de la variable **IMC** et répondez aux questions suivantes :

a) Quelle est la nature de la variable ? (qualitative nominale, qualitative ordinale, quantitative discrète ou quantitative continue). Justifiez ?

b) Combien y a-t-il de valeurs manquantes ?

c) Pourquoi est-il nécessaire de transformer cette variable?

d) Transformez la variable en créant une nouvelle variable (**IMC5**) comportant les cinq catégories ci-dessous :

- Moins de 17=Dénutrition
- 17-18,4= Maigreur
- 18,5-24,9= Corpulence normal
- 25-29,9= Surpoids
- 30 et plus =Obésité

NB : Commencez par renseigner que tout ce qui est manquant dans l'ancienne variable devient manquant dans la nouvelle variable.

e) Sortez la distribution de fréquences et de % de la variable **IMC5** et représentez-la à l'aide d'un graphique approprié.

f) Interprétez statistiquement les résultats.

Exercice 2. Calcul d'une variable (age)

Calculez l'âge (**Age**) des étudiants à partir des variables Ansond (Année de sondage) et Annais (Année de naissance).

Puis, scindez le fichier selon le sexe, et sortez les stats descriptives de la var. **Age**.