

Séance 8. Recodages et construction d'indicateurs

Objectifs de la séance

- Savoir préparer un jeu de données
- Recoder les variables en fonction des contraintes statistiques, méthodologiques, disciplinaires et du type de variable
- Construire des indicateurs

Découvrir, explorer, s'approprier un jeu de données

- En même temps que l'élaboration de la question de recherche, on prépare le jeu de données
 - ➔ supprimer des observations ou des variables, créer de nouvelles variables ou recoder (adapter) des variables existantes
- **Quelques règles**
 - Ne jamais transformer la variable initiale
 - ➔ la transformer en créant une nouvelle variable
 - Toujours recoder les variables en prenant en compte leur type (continue, ordinaire, nominale...)
 - Toujours vérifier les codages et recodages
 - ➔ faire un tri à plat de la nouvelle variable recodée
 - Garder une trace des différentes étapes, des manipulations faites sur le fichier
 - ➔ lignes de code (programme informatique) pour retracer les différentes manipulations ou fichier texte où sont notées chacune des étapes effectuées
 - **Sauvegarder son fichier avec un nouveau nom pour conserver le fichier de données initial**
 - ➔ mettre en place une stratégie de différenciation fichiers mis à jour/anciens fichiers

Exercices d'application sous Jamovi

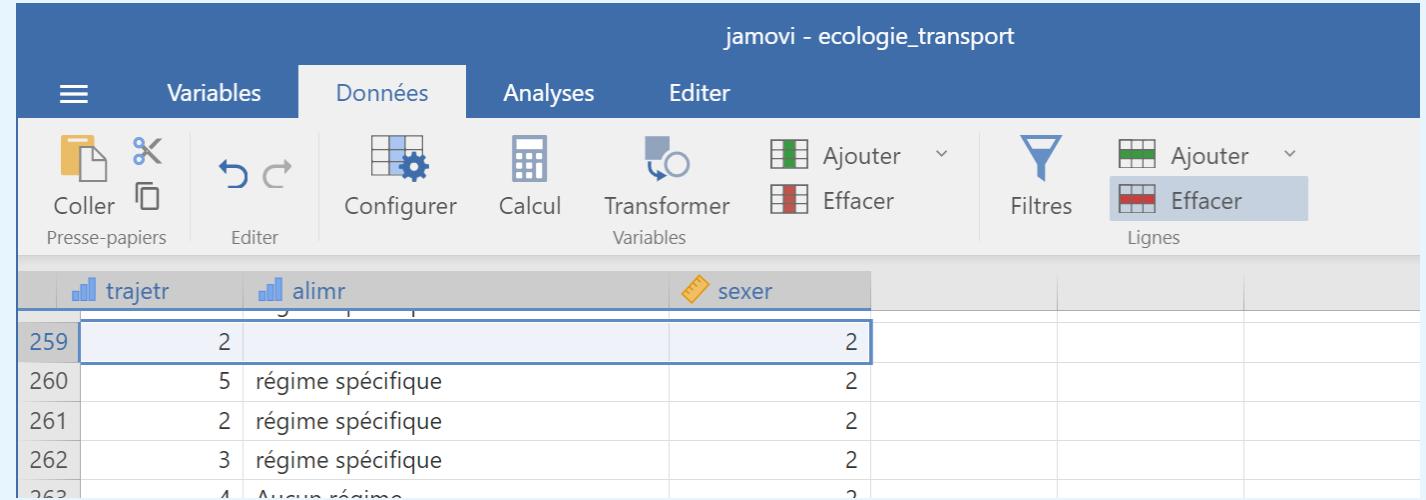
Exercice

Jeu de données SCPOBX-
ETU24, fichier
ecologie_transport

**Supprimer les étudiants qui
n'ont pas répondu à la
variable alimr**

❖ Supprimer des observations

Solution 1 : À partir de l'onglet « Données », dans la section « Lignes », sélectionner la ligne à supprimer puis cliquer sur le bouton « Effacer »



	trajetr	alimr	sexer
259	2		2
260	5	régime spécifique	2
261	2	régime spécifique	2
262	3	régime spécifique	2
263	4	Aucun régime	2

!\\ Avec cette option, il faut supprimer les lignes les unes après les autres, ce qui peut être fastidieux si on a beaucoup de lignes à supprimer !

Exercices d'application sous Jamovi

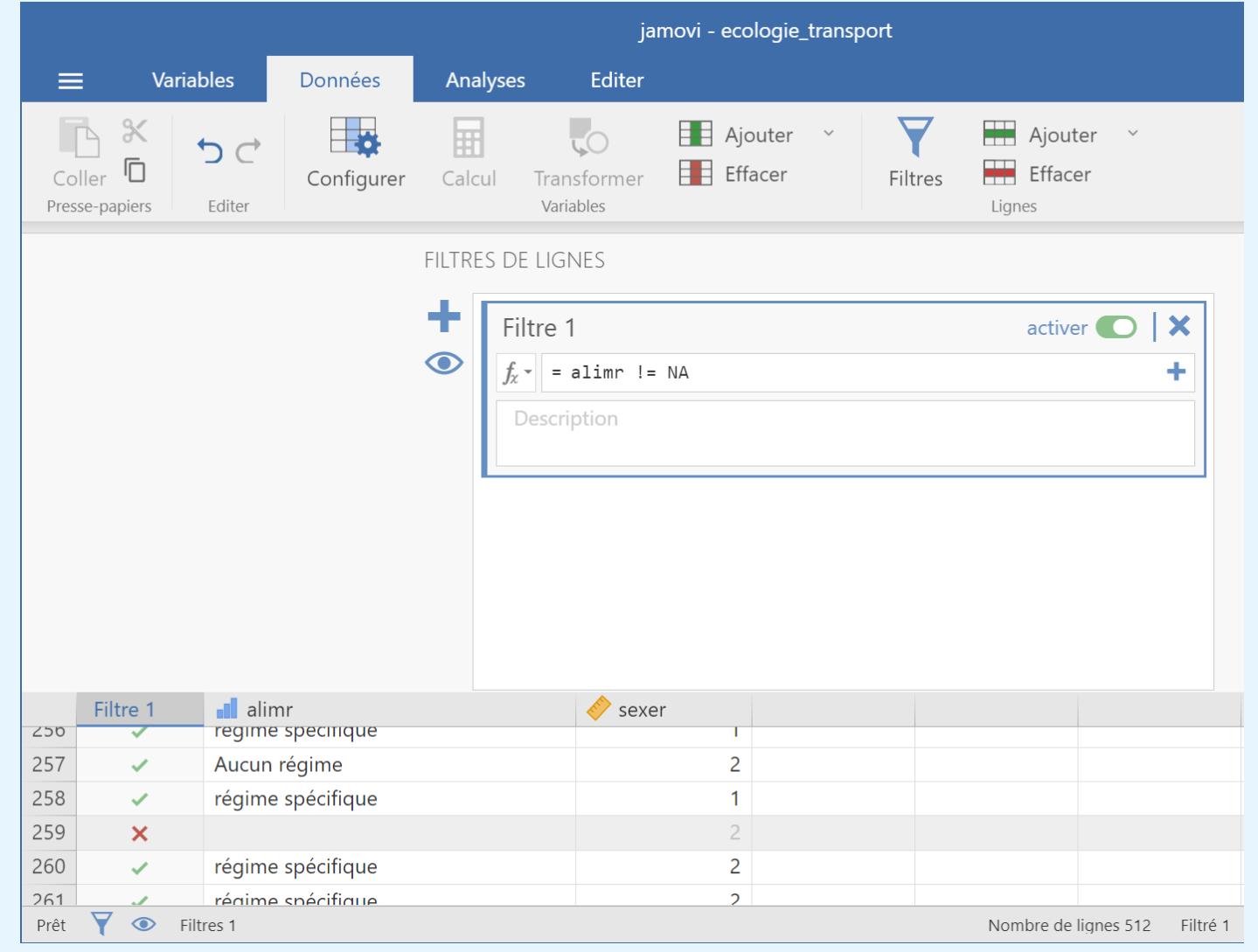
Exercice

Jeu de données SCPOBX-
ETU24, fichier
ecologie_transport

**Supprimer les étudiants qui
n'ont pas répondu à la
variable alimr**

❖ **Supprimer des observations**

Solution 2 : « supprimer » artificiellement les lignes avec un filtre



	Filtre 1	alimr	sexer
256	✓	régime spécifique	1
257	✓	Aucun régime	2
258	✓	régime spécifique	1
259	✗		2
260	✓	régime spécifique	2
261	✓	régime spécifique	2

Nombre de lignes 512 Filtré 1

Exercices d'application sous Jamovi

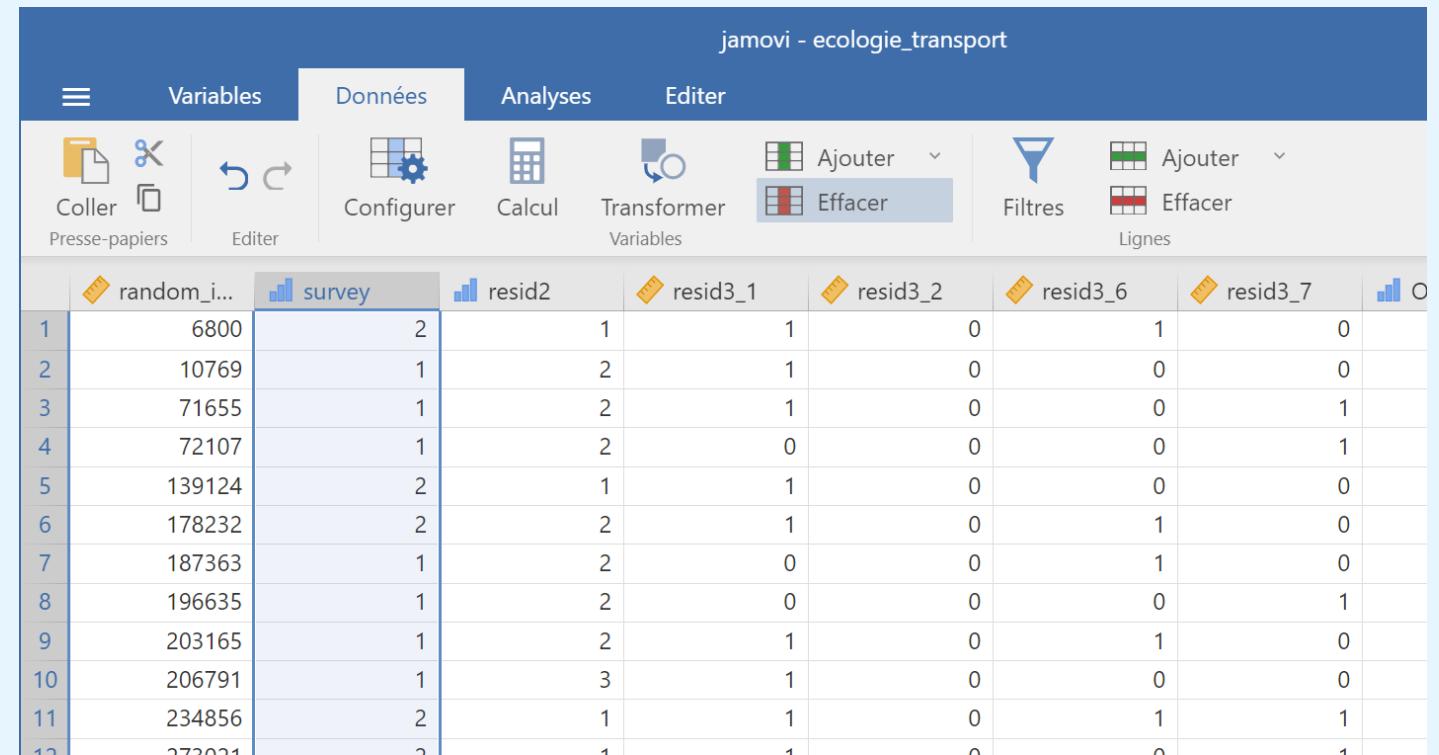
Exercice

Jeu de données SCPOBX-
ETU24, fichier
ecologie_transport

Supprimer la variable survey

❖ Supprimer des variables

À partir de l'onglet « Données », dans la section « Variables », sélectionner la colonne à supprimer puis cliquer sur le bouton « Effacer »



random_i...	survey	resid2	resid3_1	resid3_2	resid3_6	resid3_7	O
1	6800	2	1	1	0	1	0
2	10769	1	2	1	0	0	0
3	71655	1	2	1	0	0	1
4	72107	1	2	0	0	0	1
5	139124	2	1	1	0	0	0
6	178232	2	2	1	0	1	0
7	187363	1	2	0	0	1	0
8	196635	1	2	0	0	0	1
9	203165	1	2	1	0	1	0
10	206791	1	3	1	0	0	0
11	234856	2	1	1	0	1	1
12	272021	2	1	1	0	0	1

Pré-codage, codage, recodage

= Une opération de recherche

- Un difficile arbitrage, jamais neutre entre :
 - simplifier les données = réduire le nombre de modalités, qui doivent rester mutuellement
 - exclusives : ne doivent pas inclure des individus qui sont couverts par d'autres modalités
 - exhaustives : doivent inclure tous les individus concernés par la variable
 - conserver la richesse d'analyse et la complexité d'un phénomène à travers plusieurs modalités
 - adapter les variables et l'information qu'elles nous donnent à nos hypothèses de recherche (format et modalités) et aux analyses ultérieures
 - gérer les valeurs manquantes et non-réponses
 - résoudre les problèmes d'effectifs (fiabilité statistique)
- ➔ des contraintes statistiques, méthodologiques et propres à la discipline pour transformer les variables initiales et en créer de nouvelles

Exercices d'application sous Jamovi

Exercice

Jeu de données SCPOBX-
ETU24, fichier *homogamie*

**Pour la variable `mere_dipr`,
remplacer NSP en valeur
manquante**

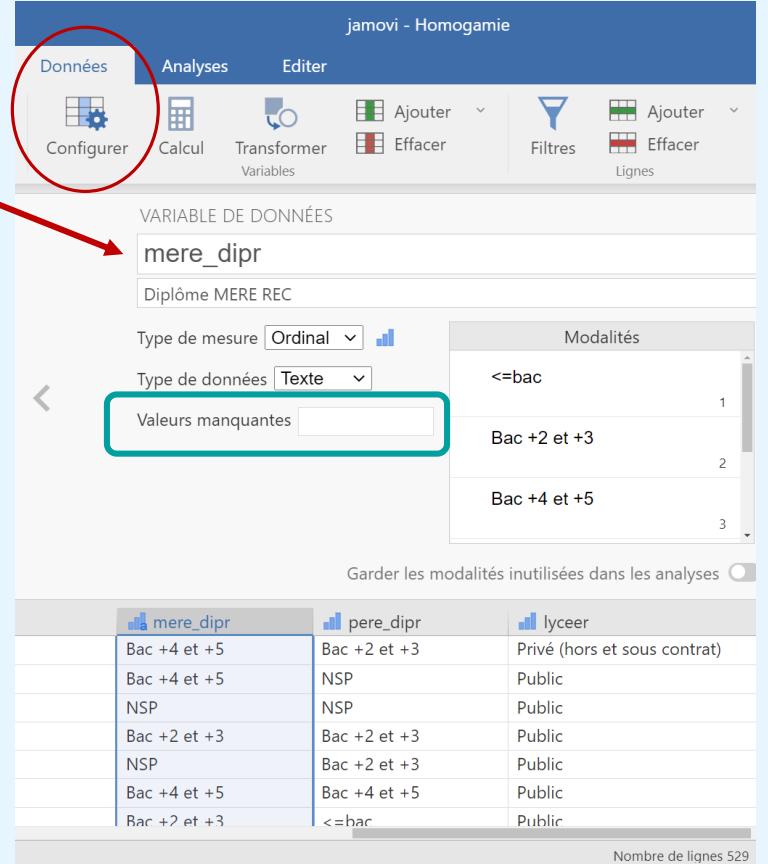
❖ Codage des valeurs manquantes

Si l'on étudie la distribution de la variable `mere_dipr`, les pourcentages tiennent compte de la modalité NSP. Il est difficile de se rendre compte de la distribution de cette variable en conservant cette modalité NSP. On va donc la considérer comme une valeur manquante.

Comment faire ?

On sélectionne la variable à modifier, ici `mere_dipr`

Puis, à partir de l'onglet « Données », dans la section « Variables », on clique sur le bouton « Configurer ». Il est alors possible de définir les valeurs manquantes en cliquant dans la boîte blanche « Valeurs manquantes »



The screenshot shows the Jamovi interface with the 'Homogamie' dataset open. The top navigation bar includes 'Analyses', 'Editor', and various data management tools like 'Ajouter' (Add), 'Effacer' (Delete), and 'Filtres' (Filters). The main window is titled 'VARIABLE DE DONNÉES' and displays the variable 'mere_dipr'. The 'Type de mesure' is set to 'Ordinal' and the 'Type de données' is set to 'Texte'. A red box highlights the 'Valeurs manquantes' input field, which is currently empty. To the right, a table lists the categories for 'mere_dipr': '=>bac', 'Bac +2 et +3', and 'Bac +4 et +5'. Below this, a table shows the data for 'mere_dipr', 'pere_dipr', and 'lyceer' variables across 529 rows. The 'mere_dipr' column includes values like 'Bac +4 et +5', 'NSP', and 'Bac +2 et +3'. The 'pere_dipr' column has values 'NSP' and 'Bac +2 et +3'. The 'lyceer' column has values 'Privé (hors et sous contrat)', 'Public', and 'Public'.

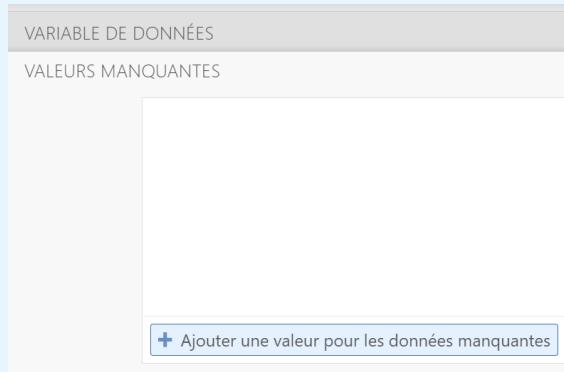
Exercices d'application sous Jamovi

Exercice

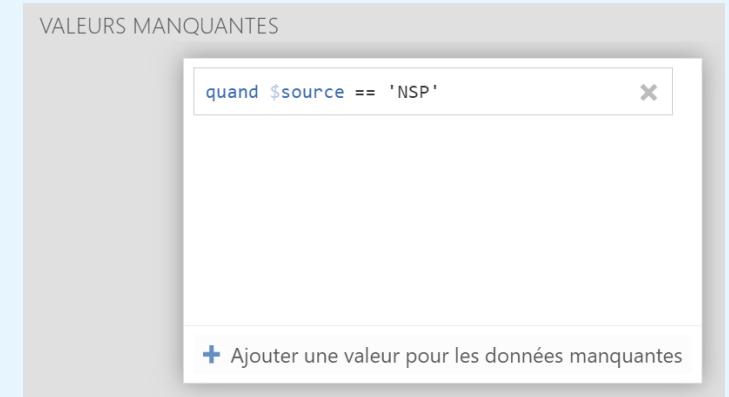
Jeu de données SCPOBX-
ETU24, fichier *homogamie*

**Pour la variable `mere_dipr`,
remplacer NSP en valeur
manquante**

Une nouvelle fenêtre s'ouvre pour configurer les valeurs manquantes
On clique sur « Ajouter une valeur pour les données manquantes »



On paramètre les valeurs manquantes



Résultat :
les valeurs NSP
apparaissent en
grisé

	VALEURS MANQUANTES
<code>quand \$source == 'NSP'</code>	
+ Ajouter une valeur pour les données manquantes	
<code>mere_dipr</code>	
Bac +4 et +5	Bac +2 et +3
Bac +4 et +5	NSP
NSP	NSP
Bac +2 et +3	Bac +2 et +3
NSP	Bac +2 et +3
Bac +4 et +5	Bac +4 et +5

NB : pour tester une
condition, on utilise une
double égalité (`==` ou `!=`)

Exercices d'application sous Jamovi

Exercice

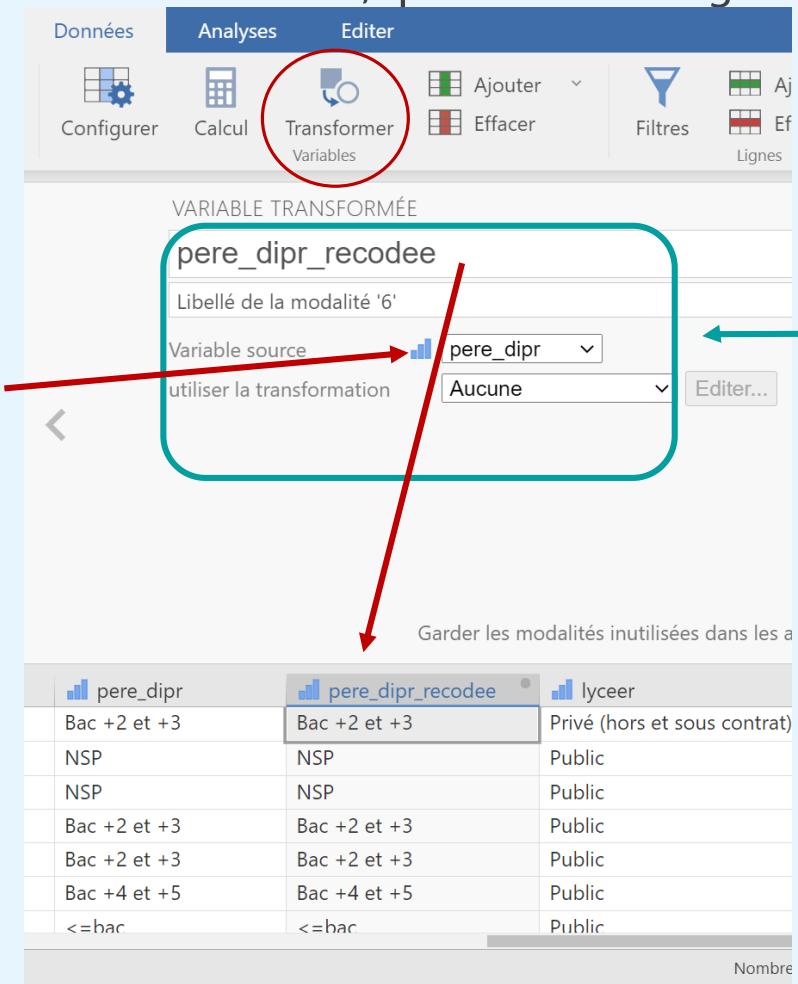
Jeu de données SCPOBX-
ETU24, fichier *homogamie*

**Pour la variable pere_dipr,
remplacer 6 par ' Bac +5'**

❖ Recodage de libellé de modalités

On sélectionne la variable à modifier, puis, à partir de l'onglet « Données », dans la section « Variables », on clique sur le bouton « Transformer »

→ cela crée une nouvelle variable, que l'on va configurer



Variable source que l'on a sélectionnée, ici *pere_dipr*

pere_dipr	pere_dipr_recodee	lyceer
Bac +2 et +3	Bac +2 et +3	Privé (hors et sous contrat)
NSP	NSP	Public
NSP	NSP	Public
Bac +2 et +3	Bac +2 et +3	Public
Bac +2 et +3	Bac +2 et +3	Public
Bac +4 et +5	Bac +4 et +5	Public
<=bac	<=bac	Public

Exercices d'application sous Jamovi

Exercice

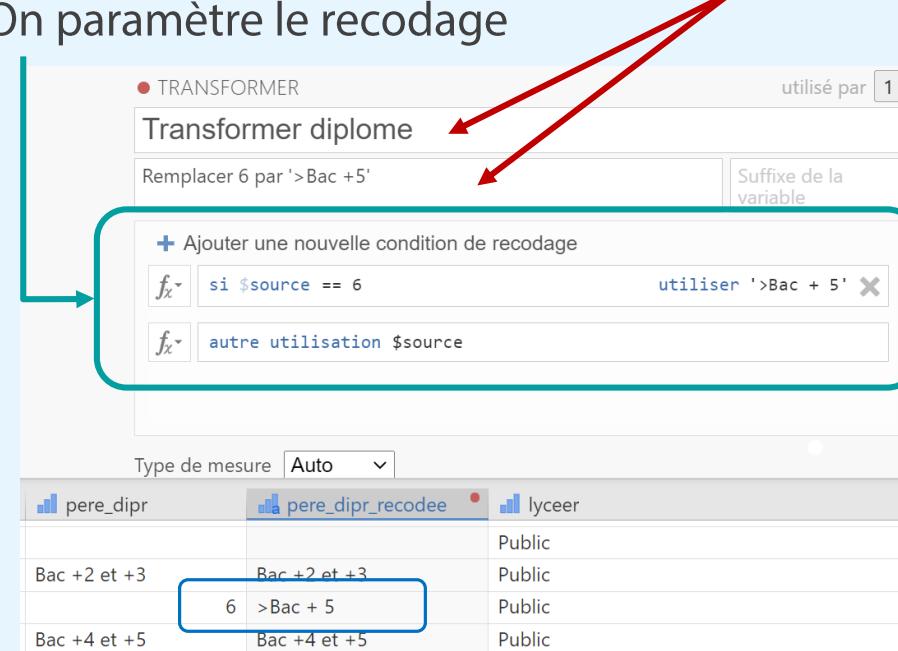
Jeu de données SCPOBX-
ETU24, fichier *homogamie*

**Pour la variable pere_dipr,
remplacer 6 par > Bac +5**

Le menu « Utiliser la transformation » permet de programmer la modification de la variable, en choisissant l'option « Créer une nouvelle variable transformée »

On décrit la transformation (le recodage) que l'on effectue

On paramètre le recodage



● TRANSFORMER

Transformer diplome

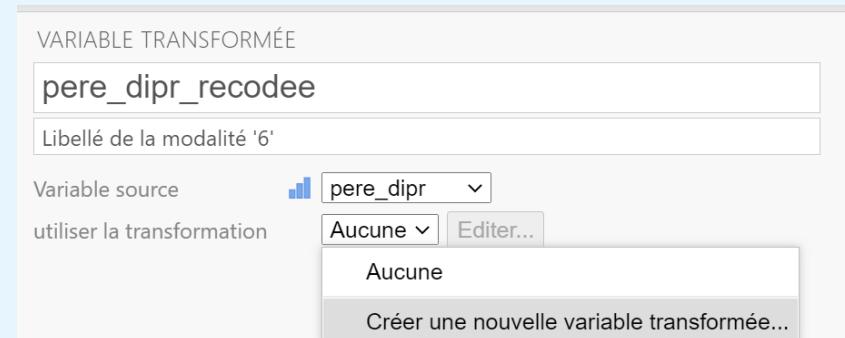
Remplacer 6 par '>Bac +5'

Ajouter une nouvelle condition de recodage

si \$source == 6 utiliser '>Bac + 5'

autre utilisation \$source

Type de mesure	Auto
pere_dipr	pere_dipr_recodee
Bac +2 et +3	Bac +2 et +3
6	>Bac + 5
Bac +4 et +5	Bac +4 et +5



Cette technique permet de conserver

- la variable originale
- le paramétrage pour le réutiliser pour une autre variable

Penser à utiliser la nouvelle variable pour faire les calculs !

NB : on peut utiliser cette manip pour traiter les valeurs manquantes. Pour mentionner « vide », il suffit de laisser un espace après **utiliser**

Démarche hypothético-déductive ou inductive ?

- **Différentes logiques de recodage**

- **inductive** = le recodage se fonde sur la distribution des données (quartiles, déciles, écarts-types par exemple)
- **hypothético-déductive** = le recodage ne tient pas compte de la distribution des données mais se fonde sur des hypothèses et un cadre théorique

/!\ ce recodage doit être acceptable d'un point de vue statistique (attention aux faibles effectifs)

Dans la pratique, on mêle souvent un peu les deux afin obtenir un compromis qui soit acceptable à la fois du point de vue statistique mais aussi du point de vue de la problématique

- On ne recode pas toutes les variables, mais celles qui sont utiles pour l'analyse !

Un exemple : comment recoder le vote des élections ? (1/2)

Frequencies of V2

Levels	Counts	% of Total	Cumulative %
Arthaud	8	0.5 %	0.5 %
Poutou	21	1.4 %	1.9 %
Melenchon	337	21.7 %	23.5 %
Hamon	97	6.2 %	29.8 %
Macron	353	22.7 %	52.5 %
Lassale	24	1.5 %	54.0 %
Fillon	210	13.5 %	67.5 %
Dupont-Aignan	56	3.6 %	71.1 %
Asselineau	16	1.0 %	72.2 %
Le Pen	291	18.7 %	90.9 %
Cheminade	1	0.1 %	90.9 %
Blank or spoiled	68	4.4 %	95.3 %
Refusal	63	4.1 %	99.4 %
DK	10	0.6 %	100.0 %

Plusieurs problématiques à prendre en compte

- Combien de modalités/valeurs présentes : y'en a-t-il trop ?
- Vérifier l'effectif pour chaque modalité : est-il suffisant ?
- Que faire des non-réponses ? Cassent-elles la hiérarchie des modalités présentes (variable ordinaire) ou font-elles artificiellement augmenter la moyenne des valeurs (variable discrète) ?
- Comment puis-je simplifier la variété des modalités SANS perdre en richesse des données ?
- Quelles modalités m'intéressent pour mesurer le lien entre deux variables que je veux étudier ?

Un exemple : comment recoder le vote des élections ? (2/2)

Recoder, c'est aussi réfléchir théoriquement sur des catégories d'analyse pertinentes

= le recodage est basé sur des hypothèses de recherche

Exemple

- *I'électorat Mélenchon et Hamon sont proches socialement (faut-il donc les recoder ensemble ?)*
- *les catégories « Extrême-gauche » ou « Extrême-droite » sont questionnables, MAIS sont théorisées dans la littérature*
- *Mélenchon doit-il être rassemblé avec d'autres candidats d'extrême-gauche ?*
- *Marine Le Pen (RN) doit-elle être classée comme un parti populiste de « droite radicale » ou « d'extrême-droite » ?*
- *Dupont-Aignan doit-il être classé à « l'extrême-droite » (scission de l'UMP) ?*

→ Allier les hypothèses avec les effectifs présents

Quels recodages ?

Variable d'origine	Objectif du recodage	Variable recodée
Continue	Regrouper des valeurs pour faire des catégories (= discréétisation)	Ordinal
Nominale	Regrouper ou fusionner des modalités entre elles, pour simplifier ou assurer des effectifs corrects	Nominale
Ordinal	Regrouper des valeurs pour faire des catégories	Ordinal
Nominale / Ordinale	Avoir une variable par modalité initiale	Dichotomique (<i>dummy</i>) / binaire [0/1]

Exercices d'application sous Jamovi

Exercice

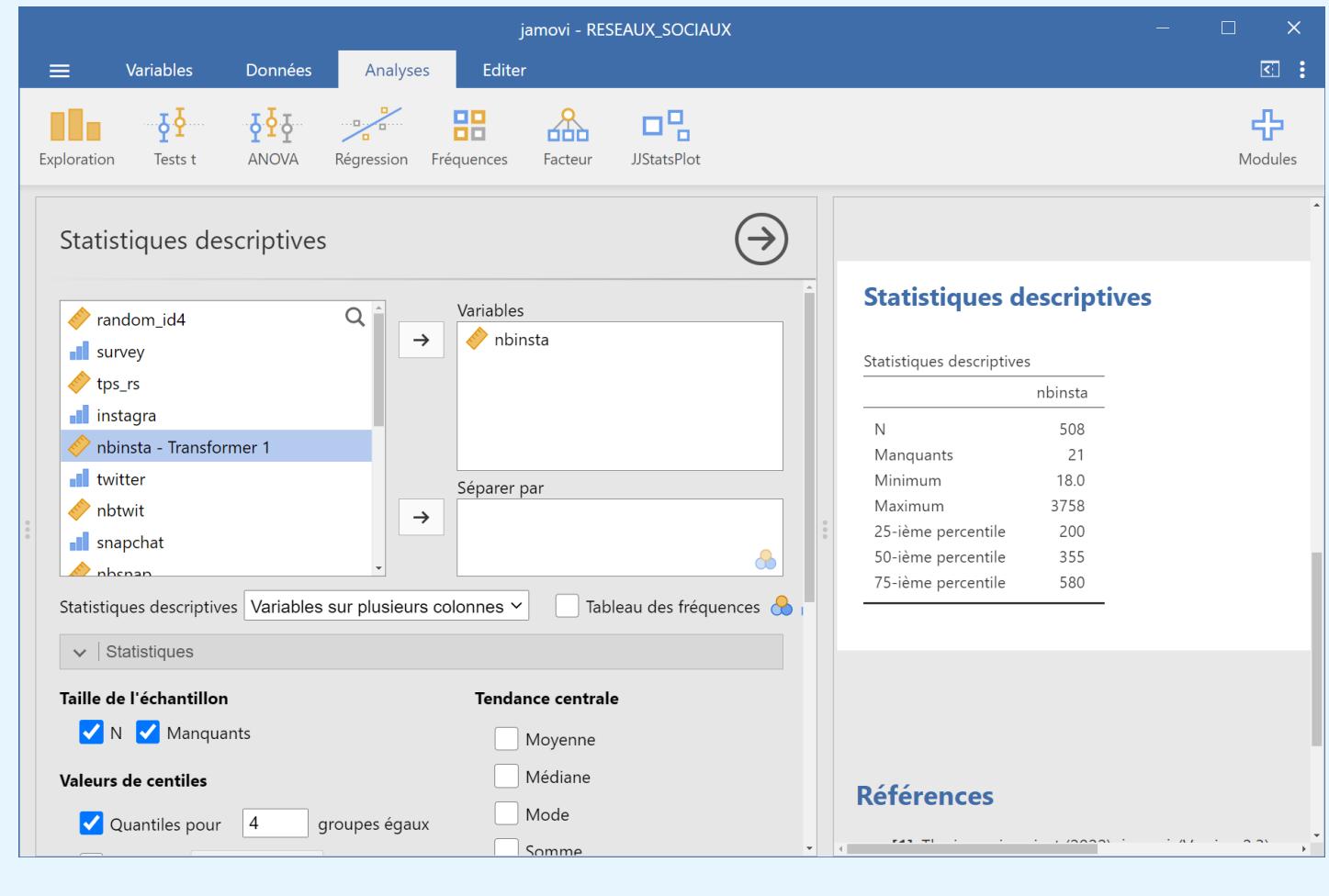
Jeu de données SCPOBX-
ETU24, fichier *reseaux_sociaux*

**En utilisant les quartiles de la
variable nb_insta, créer une
nouvelle variable
nb_insta_4cl**

❖ Discréter une variable continue

On veut discréter la variable `nb_insta`. On étudie sa distribution, en particulier la valeur des quantiles, ainsi que le minimum et maximum.

On va utiliser ces seuils pour faire la discréétisation de la variable



The screenshot shows the Jamovi interface with the title bar "jamovi - RESEAUX_SOCIAUX". The top menu bar includes "Variables", "Données", "Analyses" (selected), "Editor", and "Modules". The "Analyses" tab has several icons: Exploration, Tests t, ANOVA, Régression, Fréquences, Facteur, and JJStatsPlot. A "+" icon is also present in the top right corner.

The main window displays the "Statistiques descriptives" (Descriptive Statistics) module. On the left, a list of variables includes "random_id4", "survey", "tps_rs", "instagra", "nbinsta - Transformer 1" (highlighted in blue), "twitter", "nbtwit", "snapchat", and "phenan". An arrow points from this list to a "Variables" section containing "nbinsta". Below this is a "Séparer par" (Separate by) section with an empty box and an arrow pointing to it.

At the bottom of the left panel, there are buttons for "Statistiques descriptives" and "Variables sur plusieurs colonnes", and a checkbox for "Tableau des fréquences".

The right panel shows the results of the descriptive statistics for "nbinsta". It includes:

	nbinsta
N	508
Manquants	21
Minimum	18.0
Maximum	3758
25-ième percentile	200
50-ième percentile	355
75-ième percentile	580

Below the results, there are sections for "Références" (References) and other Jamovi settings.

Exercices d'application sous Jamovi

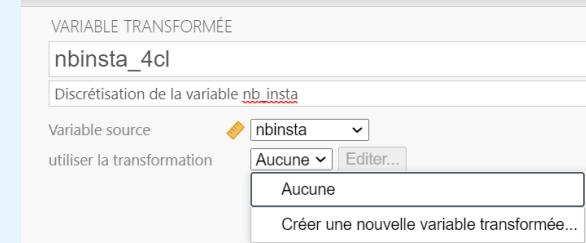
Exercice

Jeu de données SCPOBX-
ETU24, fichier *reseaux_sociaux*

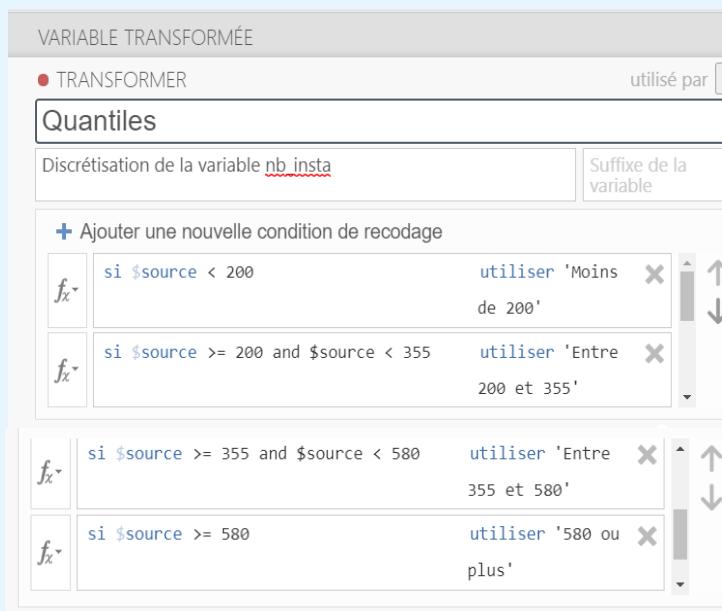
En utilisant les quartiles de la variable nb_insta, créer une nouvelle variable nb_insta_4cl

❖ Discréter une variable continue

- On effectue une transformation de la variable nb_insta, comme on l'a vu précédemment



- Pour créer la variable nb_insta_4cl, il faut paramétriser tous les cas possibles. On les ajoute en cliquant sur « Ajouter une nouvelle condition de recodage »



NB : on peut combiner les conditions, en utilisant **and** ou **or**

Ne pas oublier de bien spécifier le type de cette nouvelle variable !

Exercices d'application sous Jamovi

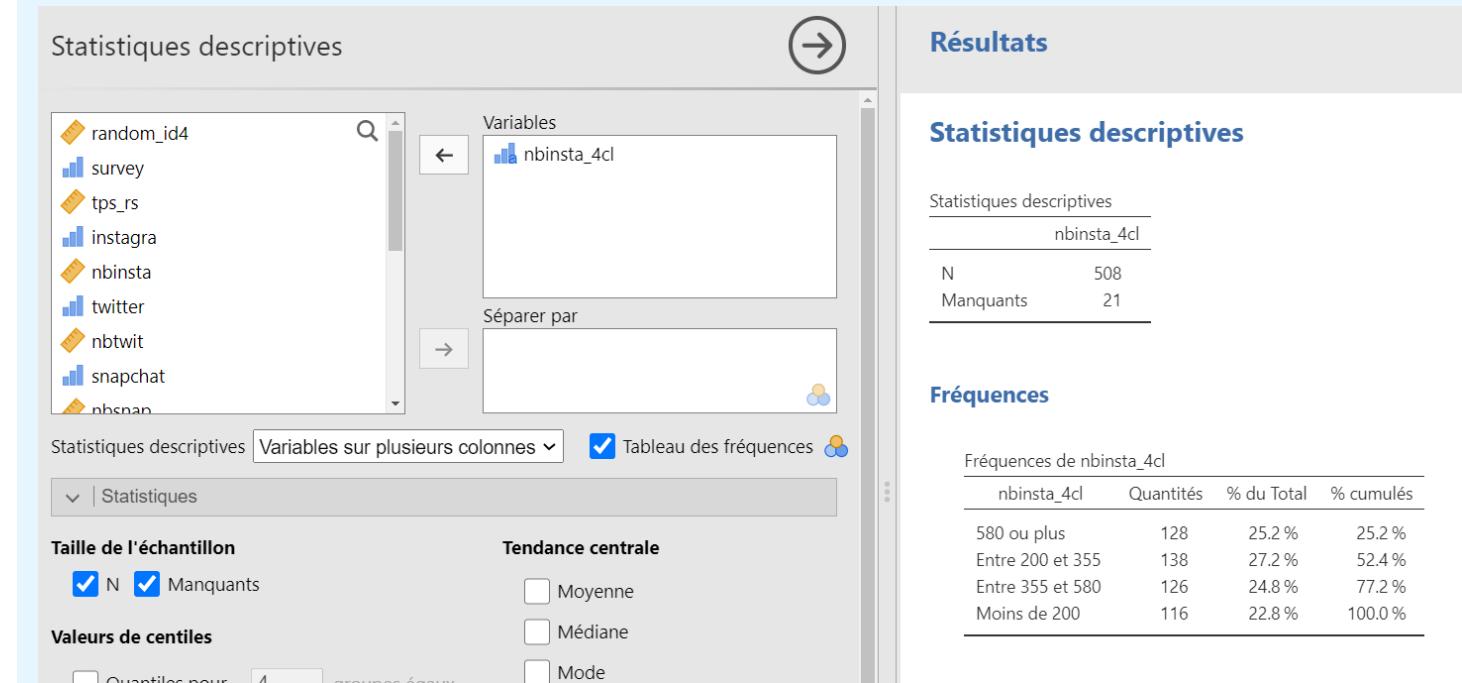
Exercice

Jeu de données SCPOBX-
ETU24, fichier *reseaux_sociaux*

**En utilisant les quartiles de la
variable nb_insta, créer une
nouvelle variable
nb_insta_4cl**

❖ Discréter une variable continue

On vérifie le recodage par une analyse de la nouvelle variable



Exercices d'application sous Jamovi

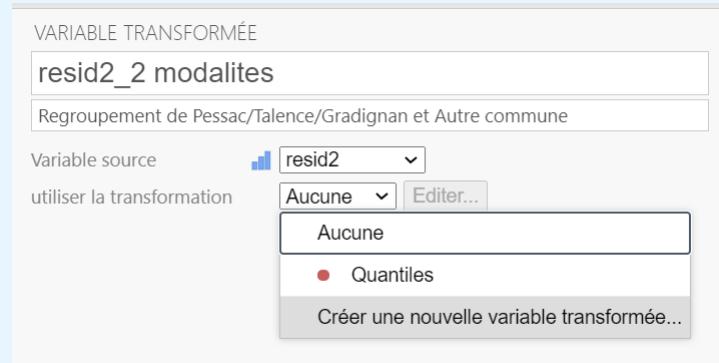
Exercice

Jeu de données SCPOBX-
ETU24, fichier *reseaux_sociaux*

**Regrouper les modalités
'Pessac/Talence/Gradignan'
et 'Autre commune' de la
variable resid2**

❖ Regrouper des modalités d'une même variable

1. On effectue une transformation de la variable resid2



2. On crée la variable resid2_2modalites, en « programmant » ce que l'on souhaite faire



Exercices d'application sous Jamovi

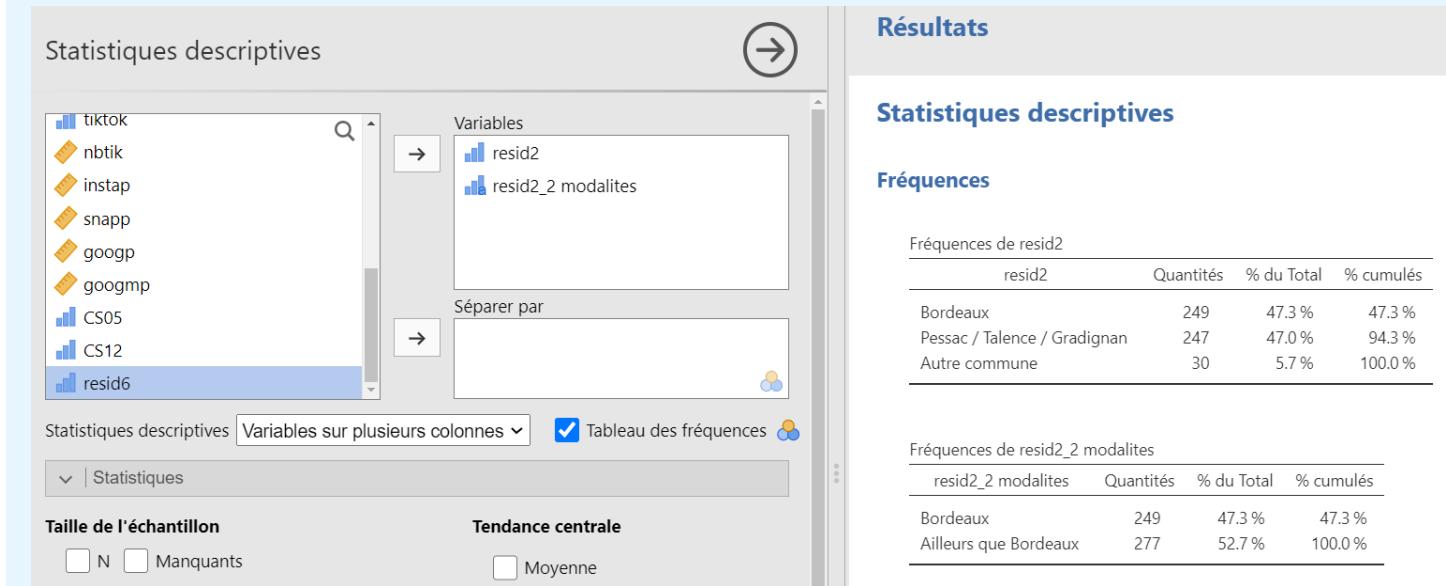
Exercice

Jeu de données SCPOBX-
ETU24, fichier *reseaux_sociaux*

**Regrouper les modalités
'Pessac/Talence/Gradignan'
et 'Autre commune' de la
variable resid2**

❖ Discréter une variable continue

On vérifie le recodage par une analyse de la nouvelle variable



Création d'indicateurs

- = résumer de façon synthétique l'information de plusieurs variables simultanément
- Pour mesurer un phénomène complexe, aux dimensions multiples : le sexisme, le racisme, l'europhilie, la compétence politique, etc...
- Un gain dans l'analyse : créer une échelle d'attitude facilement interprétable (0-100)
- Dépend des hypothèses initiales et des données à disposition (valable seulement pour les variables quantitatives et ordinaires)
- Un long travail préalable de recodage et de reconnaissance du lien entre plusieurs variables (doivent être corrélées entre elles)
- **3 options**
 - Croiser les modalités d'au moins 2 variables (patron (*pattern*) de réponses)
 - Compter les caractéristiques au sein de différentes variables (compteurs d'occurrences)
 - Additionner les codes des réponses à différentes questions échelles (échelles d'attitude)

Les patrons (*patterns*) de réponse

- Correspondent à la création d'une variable croisant les modalités d'au moins 2 variables
- Cas le plus simple = regroupement de modalités
 - lorsqu'il y a un grand nombre de modalités
 - lorsque certains croisements comportent peu d'observations
- Cette nouvelle variable peut être utilisée comme n'importe quelle autre variable

Exemple

2 variables avec 2 modalités de réponse chacune

- V1 : modalités 'oui' et 'non'
- V2 : modalités 'oui' et 'non'



Patron de réponse = 1 variable avec 4 modalités

1. 'oui' à V1 et 'oui' à V2
2. 'oui' à V1 et 'non' à V2
3. 'non' à V1 et 'oui' à V2
4. 'non' à V1 et 'non' à V2

Les compteurs

- Consistent à compter une ou plusieurs caractéristiques au sein de plusieurs variables
- Hypothèse sous-jacente = il y a un effet spécifique du caractère cumulatif de certains phénomènes sociaux

Exemple

5 variables à 2 modalités ('oui', 'non') chacune. On cherche à combien de ces variables les individus ont répondu 'oui'

- On compte les réponses 'oui'
 - ➔ Le compteur sera compris entre 0 (l'individu a répondu 'non' aux 5 variables) et 5 (l'individu a répondu 'oui' aux 5 variables)
- Pour les cas plus complexes, en fonction de la distribution du compteur, il sera nécessaire de regrouper les modalités

Les échelles d'attitudes

- Reposent sur un principe de cumul également
- Qu'est-ce qu'une **attitude** ?

« une disposition relativement **persistante** à présenter une réaction organisée d'une certaine façon, c'est-à-dire à manifester un certain type de comportement motivé, vis-à-vis d'un objet (ou d'une situation donnée) quand cet objet est en cause. L'attitude ainsi définie est un concept purement opératoire qui rend compte de l'organisation des comportements qui sont seuls observables. Il faut donc la construire à partir des régularités observées dans les comportements, l'inférer ou l'induire de ces comportements.»

Alain Lancelot, « l'orientation du comportement politique », dans Madeleine Grawitz, Jean Leca (dir.), *Traité de Science Politique*, Paris, PUF, 1985, tome 3, p.368

Les échelles d'attitudes.

Comment mesurer une attitude ?

- L'idée principale est de traduire le concept en question(s) accessibles à tous, pour mesurer une plus ou moins forte inclinaison à une attitude sur un continuum
 - = questions qui traduisent des opinions pré-existantes chez les enquêtés (= les activer)
 - ≠ une imposition d'une problématique
- Applicable seulement avec des variables quantitatives et/ou ordinales
- Nécessité de multiplier les mesures (ou tests) pour connaître le niveau de l'attitude (dans toutes ses dimensions)

Les échelles d'attitudes.

Les étapes pour créer une échelle d'attitude

- **Étape 1. Vérification de la cohérence d'ensemble des variables : les variables font-elles échelle ?**
 - = analyse de fiabilité : calcul de l'*alpha de Cronbach*
 - indice qui varie entre 0 (absence totale d'homogénéité = l'échelle ne fait pas sens) et 1 (homogénéité totale = l'échelle fait sens, les variables sont liées entre elles)
 - $\alpha = 0,55$: échelle considérée comme suffisante (acceptable)
 - $\alpha = 0,7$: échelle considérée comme importante
- **Étape 2. Vérification de l'orientation commune des variables**
 - = veiller à ce que toutes les variables soient codées dans le sens de l'attitude à tester (1 : attitude --, 2 : attitude -, 3 : attitude +, 4 : attitude++) ou toutes sur la même échelle (0/10, par ex.)
- **Étape 3. Gestion des réponses manquantes**
 - les supprimer ou les imputer par une valeur de remplacement
- **Étape 4. Créeer l'échelle**
 - = additionner toutes les valeurs des variables

Les échelles d'attitudes.

Un exemple d'échelle d'attitude avec des variables ordinaires

O25 (« assimilation ») : « Les minorités devraient s'adapter aux coutumes et aux traditions françaises »

O26 (droits des minorités) : « La volonté de la majorité doit toujours l'emporter, même aux dépends des droits des minorités »

O27 (éco) : « Les immigrés sont une bonne chose pour l'économie française »

O28 (*culturel*) : « En général, la culture française est menacée par les immigrés »

O29 (*sécurité*) : « Les immigrés font augmenter le taux de criminalité »

O30 (*social chauvinism, méfiance envers la venue des immigrés*) : « De nombreux immigrés viennent en France uniquement pour profiter de la Sécurité sociale »

Items = 0. Très d'accord 1. Plutôt d'accord 2. Ni d'accord, ni pas d'accord 3. Pas d'accord 4. Pas du tout d'accord 98. Ne sait pas 99. Ne répond pas

Exercices d'application sous Jamovi

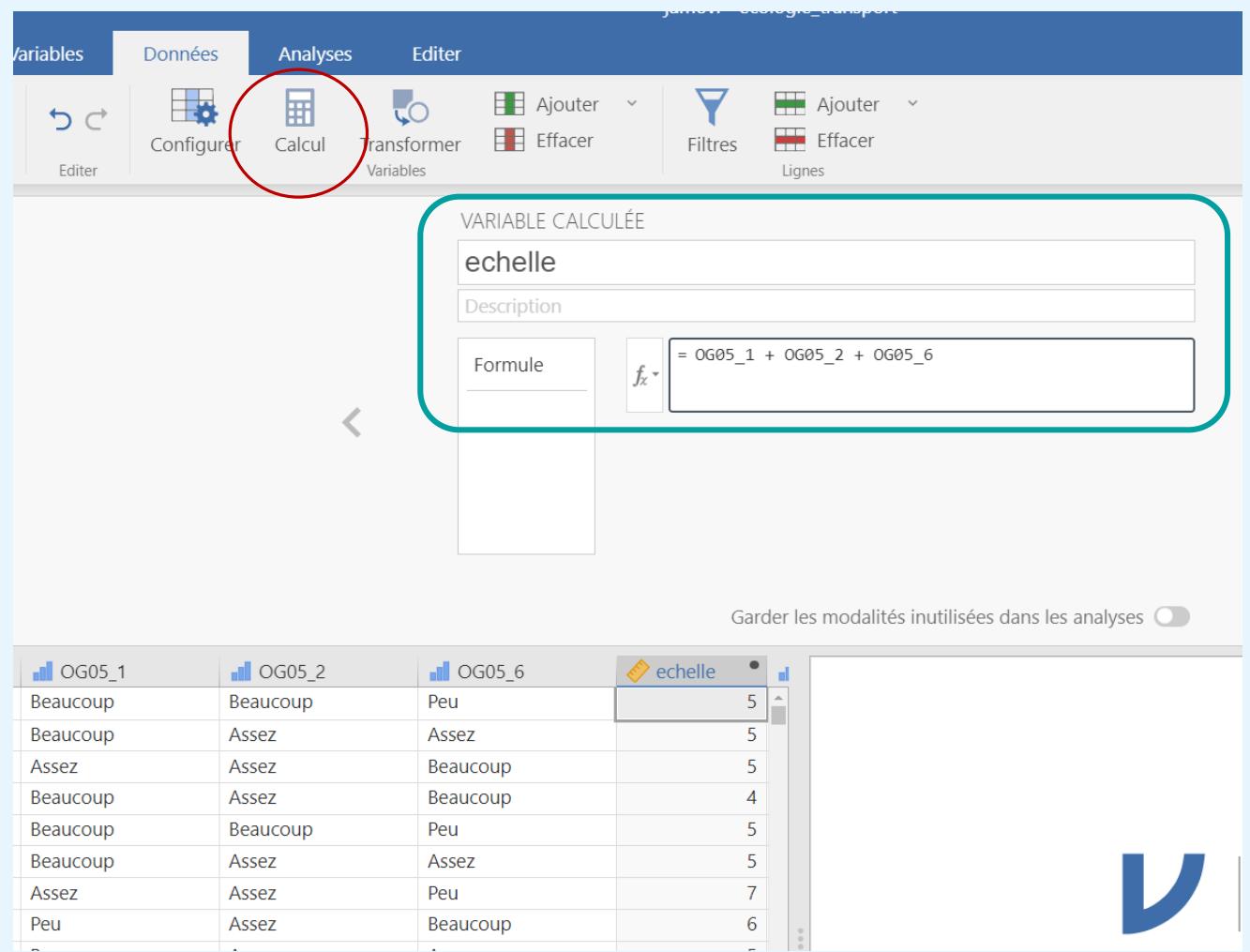
Exercice

Jeu de données SCPOBX-
ETU24, fichier
ecologie_transports

**Créer une échelle d'attitude
à partir des variables OG05_1,
OG05_2 et OG05_6 et calculer
l'alpha de Cronbach**

❖ Crée une échelle d'attitude et calculer l'alpha de Cronbach

À partir de l'onglet « Données », dans la section « Variables », on clique sur le bouton « Calcul » et on écrit la formule de calcul de l'échelle (= ajouter les valeurs des 3 variables)



The screenshot shows the Jamovi software interface. The top navigation bar has tabs for 'Variables', 'Données' (which is selected), 'Analyses', and 'Editer'. Below the tabs are several icons: 'Configurer' (Configure), 'Calcul' (Calculate, which is highlighted with a red circle), 'Transformer Variables', 'Ajouter' (Add), 'Effacer' (Delete), 'Filtres' (Filters), and 'Ajouter' (Add) for rows and columns. A large central window is titled 'VARIABLE CALCULÉE' and contains a field labeled 'echelle'. Below it is a 'Description' field and a 'Formule' field containing the formula '= OG05_1 + OG05_2 + OG05_6'. A green rounded rectangle highlights this entire calculation setup. At the bottom of the window, there is a checkbox 'Garder les modalités inutilisées dans les analyses' (Keep unused categories in analyses) with a checked status. To the right of the main window, there is a small logo of a blue 'V'.

OG05_1	OG05_2	OG05_6	echelle
Beaucoup	Beaucoup	Peu	5
Beaucoup	Assez	Assez	5
Assez	Assez	Beaucoup	5
Beaucoup	Assez	Beaucoup	4
Beaucoup	Beaucoup	Peu	5
Beaucoup	Assez	Assez	5
Assez	Assez	Peu	7
Peu	Assez	Beaucoup	6

Exercices d'application sous Jamovi

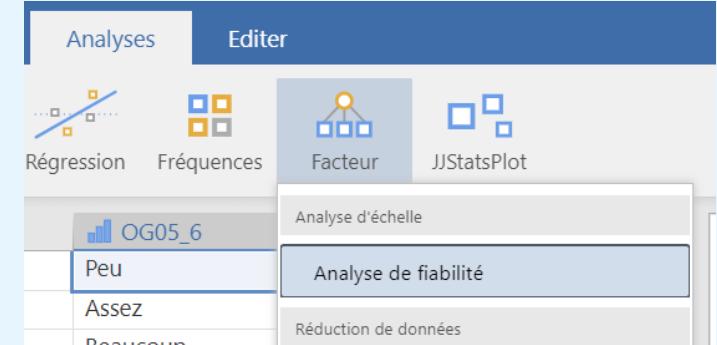
Exercice

Jeu de données SCPOBX-
ETU24, fichier
ecologie_transports

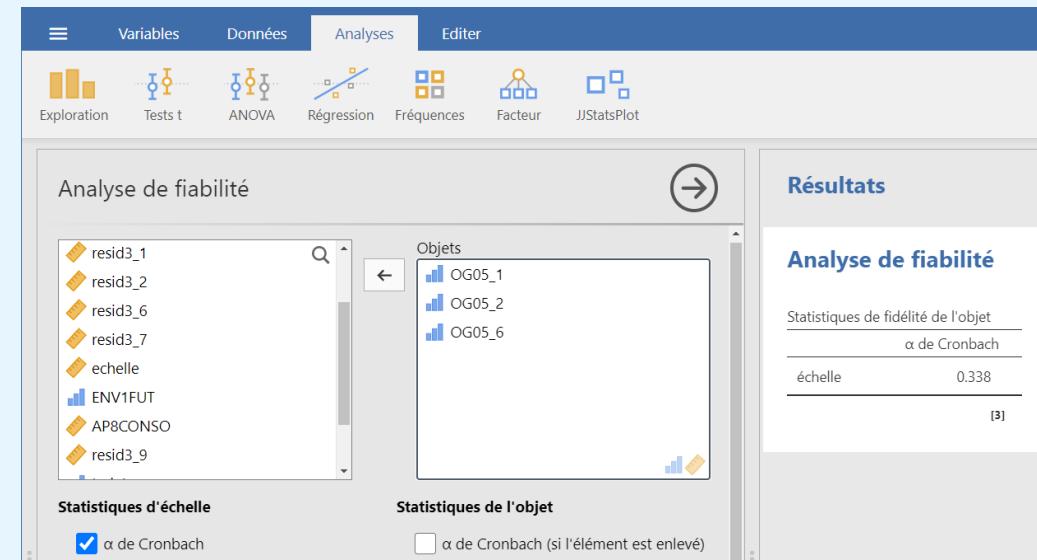
**Créer une échelle d'attitude
à partir des variables OG05_1,
OG05_2 et OG05_6 et calculer
l'alpha de Cronbach**

❖ Créer une échelle d'attitude et **calculer l'alpha de Cronbach**

À partir de l'onglet « Analyses », cliquer sur le bouton « Facteur » et sélectionner « Analyse de fiabilité »



On sélectionne les 3 variables de l'échelle pour avoir la valeur de α



Bibliographie selective

Chanvril-Ligneel F and Le Hay V (2014) Méthodes statistiques pour les sciences sociales. Paris: Ellipses.
[Chapitre 3]