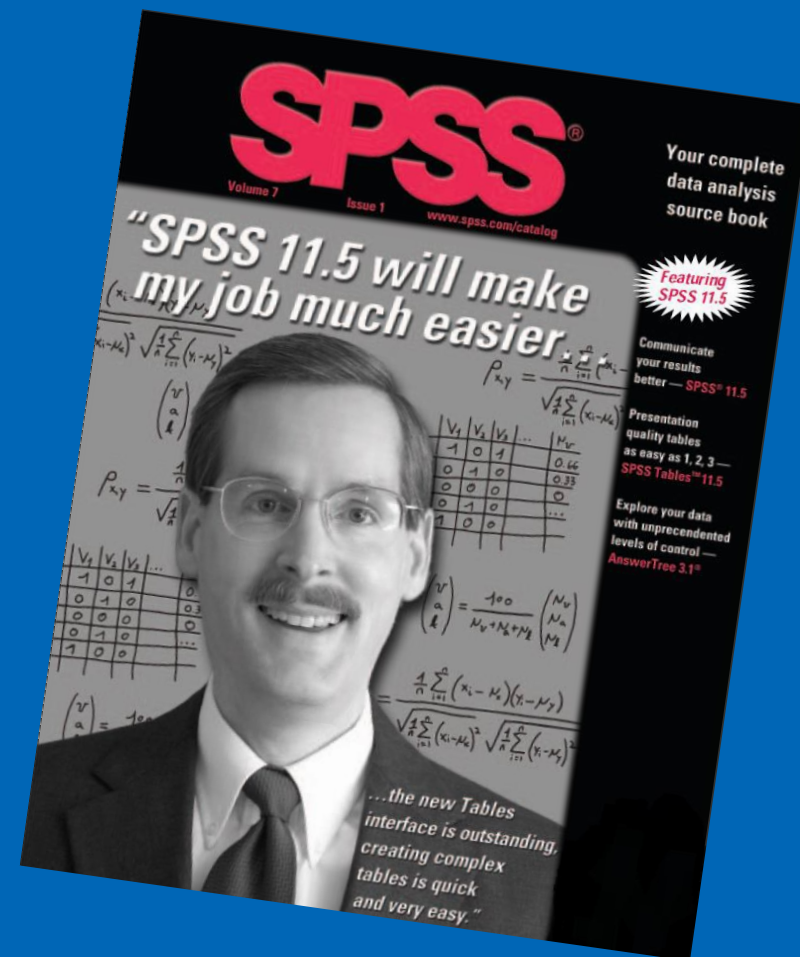


Initiation à SPSS à l'aide des microdonnées du recensement 2016



Bibliothécaire – Données | Publications gouvernementales & internationales
Bibliothèque des lettres et sciences humaines
Juin 2022

<https://github.com/CRLNP>

<https://bib.umontreal.ca/guides/donnees-statistiques-geospatiales/donnees-statistiques>

les bibliothèques

Université 
de Montréal

Objectifs généraux

- Connaître les principales caractéristiques de l'interface.
- Se familiariser avec les composantes d'un fichier de microdonnées.
- Réviser de petites notions statistiques.
- Maîtriser les fonctions de base grâce à des exercices pratiques.



Accent mis sur la **maîtrise du logiciel** et des **procédures descriptives** et non sur les statistiques inférentielles et l'interprétation de tests d'hypothèses.



*Explorer, résumer,
organiser et simplifier
les données*



Statistiques descriptives

Fréquences

Moyenne

Écart type

Médiane

Étendue ...

Statistiques inférentielles (estimations sur la population et tests d'hypothèse)

Khi deux

T-tests

ANOVA (stat F)

Corrélation

Régression ...



*Étudier les échantillons et faire
des estimations (inférences) au
sujet de la population de laquelle
les échantillons sont tirés*

Plan

1. Logiciel et fichiers de microdonnées
2. SPSS et cie...
3. Fichier de microdonnées utilisé
4. Ouvrir un fichier .sav et explorer l'interface
 - 4.1. Ouverture d'un fichier de microdonnées
 - 4.2. Tour d'horizon de SPSS -> Trois fenêtres
 - 4.3. Impression et sauvegarde
5. Principales fonctions
 - 5.1. Menu Édition
 - 5.2. Menu Données
 - 5.3. Menu Transformer
 - 5.4. Menu Analyse
6. Création d'une base de données
 - 6.1. Les valeurs manquantes
 - 6.2. Échelles de mesure
7. Nettoyer et préparer les données
8. Quelques mots sur les postulats
9. Importer un fichier de données en format .csv

10. Exercices

- 10.1. Tableau de fréquence (+ Pondération)
- 10.2. Sélectionner des sous-groupes.
- 10.3. Recoder une variable catégorielle
- 10.4. Tableau croisé
- 10.5. Calculer une variable
- 10.6. Tableau de variables d'échelle
- 10.7. Comparer des moyennes de groupes
- 10.8. Corrélation
- 10.9. Graphiques à barres

1. Logiciel et fichiers de microdonnées

- Où trouver SPSS ? [Logithèque](#) > [Procédures d'installation SPSS](#)
 - SPSS AMOS: logiciel de modélisation par équation structurelle (analyses multivariées, relations complexes, ...)
- Option [PSP](#) – Logiciel libre.
- Où trouver des fichiers de microdonnées?
 - Statistique Canada
 - [Odesi](#)
 - [ICPSR](#)
 - [Banque Mondiale](#)
 - [Baromètres](#)
 - Voir [Guide Données statistiques](#)
- Pourquoi SPSS? (Stata, SAS, R,...)

2. SPSS et cie...

If statistics programs/languages were cars...



https://twitter.com/kai_arzheimer/status/974280365446717441/photo/1



Issues With Using Microsoft Excel For Statistical Analysis and Graphics
Problems with using Microsoft Excel for Statistical Analysis & Graphics

3. Fichier de microdonnées utilisé

Recensement de la population, 2016 [Canada] Fichier de micro-données à grande diffusion (FMGD): Fichier des particuliers

« Le fichier du recensement de 2016 fournit des données sur les caractéristiques de la population canadienne. Il contient un échantillon de 2,7 % de réponses anonymes tirées du questionnaire du Recensement de 2016, soit **930 421 individus** ».

« Les fichiers de micro-données sont les seuls produits donnant aux utilisateurs l'accès à des données non agrégées. L'utilisateur des FMGD peut grouper et manipuler ces variables en fonction de ses besoins et de l'objet de ses recherches ».

Le FMGD de 2016 comporte **123 variables** (pour l'atelier: 20).

➡ Télécharger le fichier du recensement via [Odesi](#)

[Dictionnaire du recensement](#)

[Guide de l'utilisateur](#)

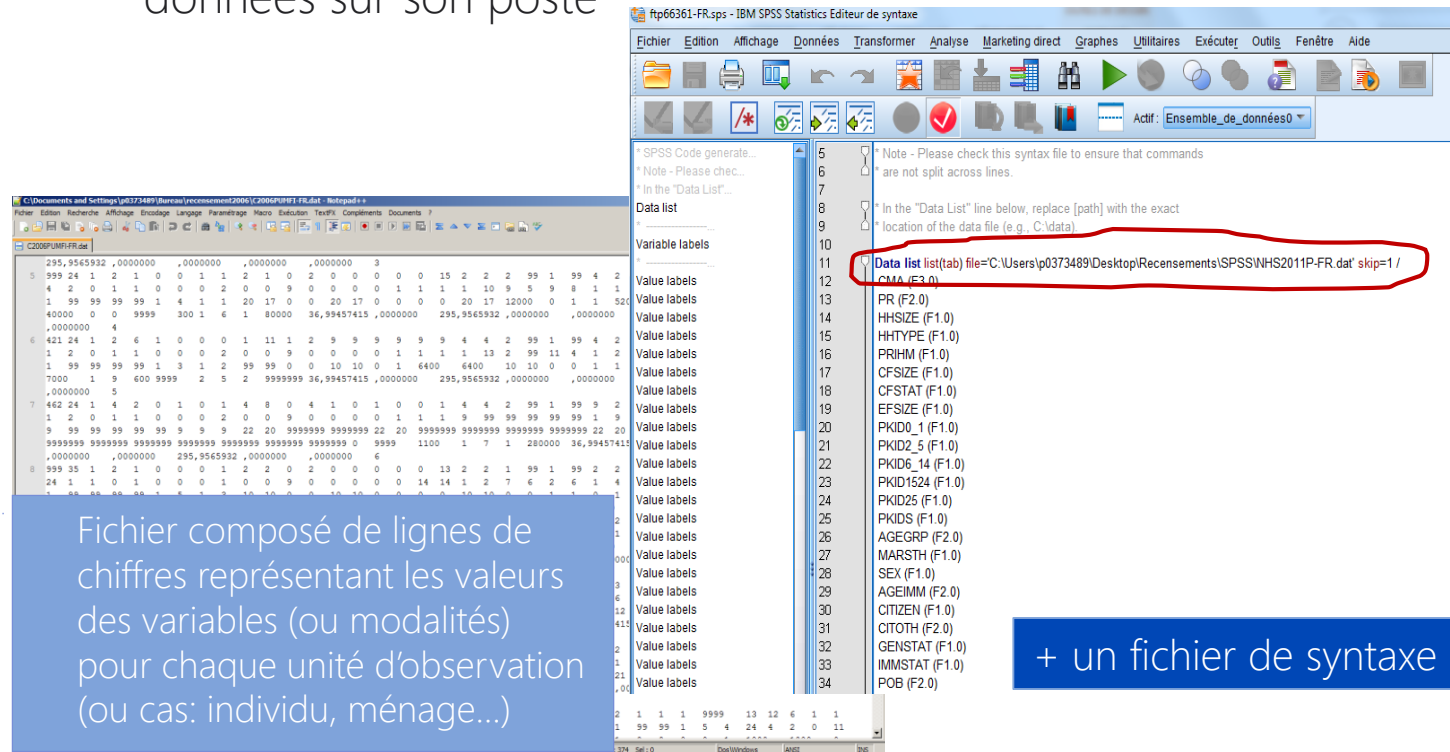


Importance de la documentation -> normes de diffusion, mesures de précision des estimations (erreur-type, coefficient de variation...)
EX: [EDTR](#)

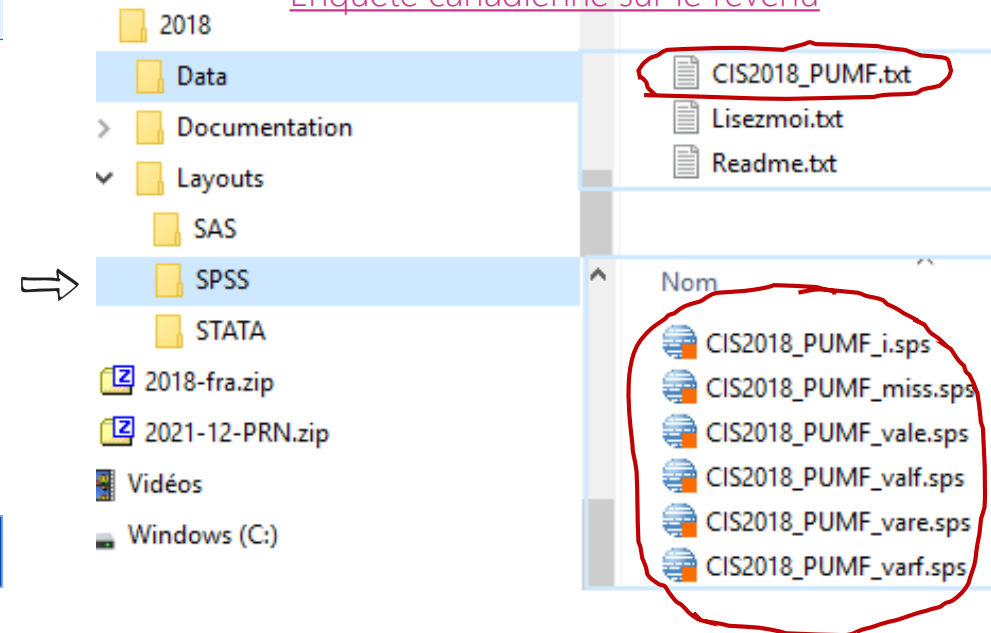
4. Ouvrir un fichier .sav et explorer l'interface

4.1. Ouverture d'un fichier de microdonnées

- Un seul fichier **.sav**
OU
- Fichier de données brutes (.txt) + fichier de syntaxe (.sps) -> changer le 'Chemin' vers le fichier de données sur son poste



Exemple des fichiers sur le site de StatCan:
[Enquête canadienne sur le revenu](#)



Ouverture d'un fichier de données d'un autre format (excel, txt, csv, stata, sas,...):

- Fichier > Ouvrir (OU Importer des données) > Données (suivre les indications, **attention à la mise en page du fichier original**).
- On peut également **exporter** les données en **différents formats**.

4.2. Tour d'horizon de SPSS -> Fenêtres

1. Base de données
Éditeur de données
.sav

2 onglets:
Vue des données
Vue des variables



2. Résultats
Statistics Viewer
.spv

Sortie
Tableaux et
graphiques



3. Éditeur de Syntaxe
.sps

Commandes à
effectuer



	ABOD	MODE	MARSTR	WEIGHT	AGEGRP	PPSORT	AGEMM	MMSTAT	PKIDS	KOL	FOL	POB
1	6	8	5	32.39361116	15	1	10	2	9	3	1	16
2	6	2	4	32.39361116	11	2	99	1	9	1	1	1
3	6	7	2	42.94465734	9	3	99	3	9	1	1	23
4	6	2	4	32.39361116	13	4	99	1	1	1	1	1
5	1	9	1	32.39361116	13	5	99	1	1	4	4	1
6	6	2	3	32.39361116	13	6	99	1	1	3	2	1
7	6	9	2	32.39361116	10	7	99	1	0	1	1	89
8	6	9	2	32.39361116	14	8	99	1	0	3	2	1
9	6	7	2	32.39361116	12	9	99	3	1	3	3	5
10	6	2	1	32.39361116	9	10	99	3	9	1	1	19
11	6	9	1	32.39361116	3	11	99	1	1	2	2	1
12	6	2	3	32.39361116	9	12	99	1	0	2	2	1
13	6	9	1	32.39361116	12	13	99	1	9	1	1	1
14	6	2	2	32.39361116	10	14	99	1	1	1	1	1
15	6	8	4	32.39361116	9	15	5	2	1	1	1	22
16	6	9	1	116.7244234	1	16	99	1	1	2	2	1
17	6	2	2	32.39361116	9	17	99	1	1	2	2	1
18	6	2	2	197.1039676	10	18	99	3	0	1	1	2
19	6	2	1	32.39361116	12	19	99	1	1	1	1	1
20	6	7	6	65.08974802	13	20	7	2	1	1	1	24
21	6	9	2	32.39361116	15	21	99	1	0	1	1	1
22	6	9	3	32.39361116	16	22	99	1	0	2	2	1
23	6	9	2	32.39361116	12	23	7	2	1	1	1	21
24	6	9	1	32.39361116	16	24	6	2	9	4	4	15
25	6	9	2	32.39361116	19	25	99	1	0	1	1	1
26	6	9	1	32.39361116	3	26	99	1	1	1	1	1
27	6	2	2	32.39361116	11	27	99	1	1	1	1	1
28	6	2	3	32.39361116	13	28	99	1	0	1	1	1
29	6	8	1	32.39361116	11	29	99	1	9	1	1	1

	N	Minimum	Maximum	Moyenne	Écart-type	Variance
IMMSTAT Immigration - Statut d'immigrant	887012	1	3	1.24	.449	.202
N valide (Total)	887012					

	SEV Sexe	N	Moyenne	Écart-type	Moyenne erreur standard
IMMSTAT Immigration - Statut d'immigrant	1 Femme	451596	1.24	.452	.001
	2 Homme	435416	1.23	.446	.001
MOVINC Revenu - Revenu du ménage	1 Femme	377094	27363.66	37814.615	61.578
	2 Homme	356965	45165.95	75801.208	126.888

```
1 DATASET ACTIVATE Jeu_de_donnees1.  
2 DESCRIPTIVES VARIABLES=IMMSTAT  
3 /STATISTICS=MEAN STDEV VARIANCE MIN MAX.  
4  
5  
6 T-TEST GROUPS=SEX(1 2)  
7 /MISSING=ANALYSIS  
8 /VARIABLES=IMMSTAT MOVINC  
9 /CRITERIA=CI(.95).  
10  
11 CORRELATIONS  
12 /VARIABLES=MOVINC IMMSTAT  
13 /PRINT=BOTH TWOTAL NOSIG  
14 /MISSING=PAIRWISE.  
15  
16 NONPAR CORR  
17 /VARIABLES=MOVINC IMMSTAT  
18 /PRINT=BOTH TWOTAL NOSIG  
19 /MISSING=PAIRWISE.  
20  
21 EXAMINE VARIABLES=MOVINC BY SEX  
22 /PLOT=BOXPLOT STEMLEAF  
23 /COMPARE=GROUPS  
24 /PERCENTILES(5 10 25 50 75 90 95) HAVERAGE  
25 /STATISTICS=DESCRIPTIVES EXTREME  
26 /CRITERIAL=95  
27 /MISSING=LISTWISE  
28 /NOTOTAL.  
29  
30 CROSSTABS  
31 /TABLES=AGEMM BY KOL  
32 /FORMAT=AVALUE TABLES  
33 /CELLS=COUNT
```

4.2.1. Éditeur de données -> Vue des variables

- Liste des variables et de leurs caractéristiques (Type, Valeurs, Valeurs manquantes, Mesure ...).
- Permet d'explorer, ajouter, éditer, supprimer, déplacer, trier les variables et leurs valeurs.

Accès rapide aux dernières commandes effectuées

Chaque variable et ses caractéristiques occupent une ligne

Un double-clic sur un numéro de variable transfère vers sa colonne dans l'affichage de données

Bouton Variables: pour visualiser l'ensemble des informations sur les variables

	Nom	Type	Largeur	Décimales	Étiquette	Valeurs	Manquant	Colonnes	Align	Mesure	Rôle
1	REC_NUM	Numérique	7	0	Ordre de l'obser...	Aucun	Aucun	9	⇒ Droite	Echelle	Entrée
2	SURVYEAR	Numérique	4	0	Année d'enquête	{2, 007 "200...	Aucun	10	⇒ Droite	Nominales	Entrée
3	SURVMNTH	Numérique	2	0	Mois d'enquête	{1, Janvier}...	Aucun	10	⇒ Droite	Nominales	Entrée
4	LFSSTAT	Numérique	1	0	Situation vis-à-v...	{1, Personn...	Aucun	9	⇒ Droite	Nominales	Entrée
5	PROV	Numérique	2	0	Province	{10, Terre-N...	Aucun	6	⇒ Droite	Echelle	Entrée
6	CMA	Numérique	1	0	Demeure à Mo...	{1, Montréal...	Aucun	5	⇒ Droite	Nominales	Entrée
7	AGE_12	Numérique	2	0	Âge du réponda...	{1, Âgé entr...	Aucun	8	⇒ Droite	Nominales	Entrée
8	AGE_6	Numérique	2	0	Âge des 15 à 2...	{-1, Sans ob...	-1	7	⇒ Droite	Nominales	Entrée
9	SEX	Numérique	1	0	Sexe du répod...	{1, Hommes...	Aucun	5	⇒ Droite	Nominales	Entrée
10	MARSTAT	Numérique	1	0	État matrimoni...	{1, Mariés}...	Aucun	9	⇒ Droite	Nominales	Entrée
11	ED76TO89	Numérique	2	0	Plus haut nivea...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
12	EDUC90	Numérique	2	0	Plus haut nivea...	{0, 0 à 8 an...	Aucun	8	⇒ Droite	Nominales	Entrée
13	MJH	Numérique	2	0	Cumul d'emploi...	{-1, Sans ob...	-1	5	⇒ Droite	Nominales	Entrée
14	EVERWORK	Numérique	2	0	Ne travaillent p...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
15	FTPTLAST	Numérique	2	0	Situation du der...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
16	COWMAIN	Numérique	2	0	Catégorie de tr...	{-1, Sans ob...	-1	9	⇒ Droite	Nominales	Entrée
17	NAICS_18	Numérique	2	0	Branche d'activi...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
18	NAICS_43	Numérique	2	0	Branche d'activi...	{-1, Sans ob...	-1	10	⇒ Droite	Echelle	Entrée
19	SOC80_49	Numérique	2	0	Profession à l'e...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
20	SOC80_21	Numérique	2	0	Profession à l'e...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
21	NOC01_25	Numérique	2	0	Profession à l'e...	{-1, Sans ob...	-1	10	⇒ Droite	Echelle	Entrée
22	NOC01_47	Numérique	2	0	Profession à l'e...	{-1, Sans ob...	-1	10	⇒ Droite	Echelle	Entrée
23	YABSENT	Numérique	2	0	Personnes occ...	{-1, Sans ob...	-1	9	⇒ Droite	Nominales	Entrée
24	WKSAWAY	Numérique	2	0	Semaines d'ab...	{-1, Sans ob...	-1	9	⇒ Droite	Echelle	Entrée
25	PAYAWAY	Numérique	2	0	Congé rémunér...	{-1, Sans ob...	-1	9	⇒ Droite	Nominales	Entrée
26	UHRMAIN	Numérique	5	1	Heures habituel...	{-1,0, 0 "Sa...	-1,0	10	⇒ Droite	Echelle	Entrée
27	AHRMAIN	Numérique	5	1	Heures effective...	{-1,0, 0 "Sa...	-1,0	10	⇒ Droite	Echelle	Entrée
28	FTPTMAIN	Numérique	2	0	Temps plein ou...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
29	UTOTHRS	Numérique	5	1	Heures habituel...	{-1,0, 0 "Sa...	-1,0	9	⇒ Droite	Echelle	Entrée
30	ATOTHRS	Numérique	5	1	Heures effective...	{-1,0, 0 "Sa...	-1,0	9	⇒ Droite	Echelle	Entrée
31	HRSAWAY	Numérique	5	1	Heures d'absen...	{-1,0, 0 "Sa...	-1,0	9	⇒ Droite	Echelle	Entrée
32	YAWAY	Numérique	2	0	Raison pour ab...	{-1, Sans ob...	-1	7	⇒ Droite	Nominales	Entrée
33	PAIDOT	Numérique	5	1	Heures supplé...	{-1,0, 0 "Sa...	-1,0	8	⇒ Droite	Echelle	Entrée
34	UNPAIDOT	Numérique	5	1	Heures supplé...	{-1,0, 0 "Sa...	-1,0	10	⇒ Droite	Echelle	Entrée
35	XTRAHRS	Numérique	5	1	Nombre d'heure...	{-1,0, 0 "Sa...	-1,0	9	⇒ Droite	Echelle	Entrée
36	WHYPTOLD	Numérique	2	0	Raison pour le t...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
37	WHYPTNEW	Numérique	2	0	Raison pour le t...	{-1, Sans ob...	-1	10	⇒ Droite	Nominales	Entrée
38	TENURE	Numérique	3	0	Durée de l'empl...	{-1, Sans ob...	-1	8	⇒ Droite	Echelle	Entrée

Dictionnaire de données:

- ▶ Menu **Fichier** > Afficher des informations sur un fichier de données > fichier de travail.
- ▶ Menu **Analyse** > Rapports > Livre de codes.
- ▶ Menu **Utilitaires** > Variables.

4.2.1. Éditeur de données -> Vue des données

- Colonnes > variables
- Lignes > **unité d'analyse**: cas, observations, répondants...
- Cellules > valeurs | réponses (modalités, attributs)

Double-clic sur l'intitulé d'une colonne : transfert à sa ligne dans la vue des variables

Clic droit sur l'intitulé d'une colonne : trier les valeurs de la variable

Clic droit sur n'importe quelle cellule: Générer automatiquement un tableau de stats descriptives

Possibilité de scinder la fenêtre en "figeant" une colonne ou une ligne

Remplacer les codes par les étiquettes de valeurs

The screenshot shows the IBM SPSS Statistics 'Data Editor' window. The menu bar includes 'Fichier', 'Édition', 'Affichage', 'Données', 'Transformer', 'Analyse', 'Graphes', 'Utilitaires', 'Fenêtre', and 'Aide'. The toolbar contains various icons for file operations, editing, and analysis. The main area displays a data table with columns: REC_NUM, SURVYEAR, SURVMNTH, LFSSTAT, PROV, CMA, AGE_12, AGE_6, SEX, MARSTAT, ED76TO89, EDUC90, MJH, EVERWORK, and FTPTLAST. The rows are numbered 1 to 36. A red circle highlights the 'AGE_12' column header, and another red circle highlights a cell in the 'AGE_6' column. At the bottom, there are tabs for 'Affichage des données' (selected) and 'Affichage des variables'. A red arrow points to the 'Affichage des données' tab.

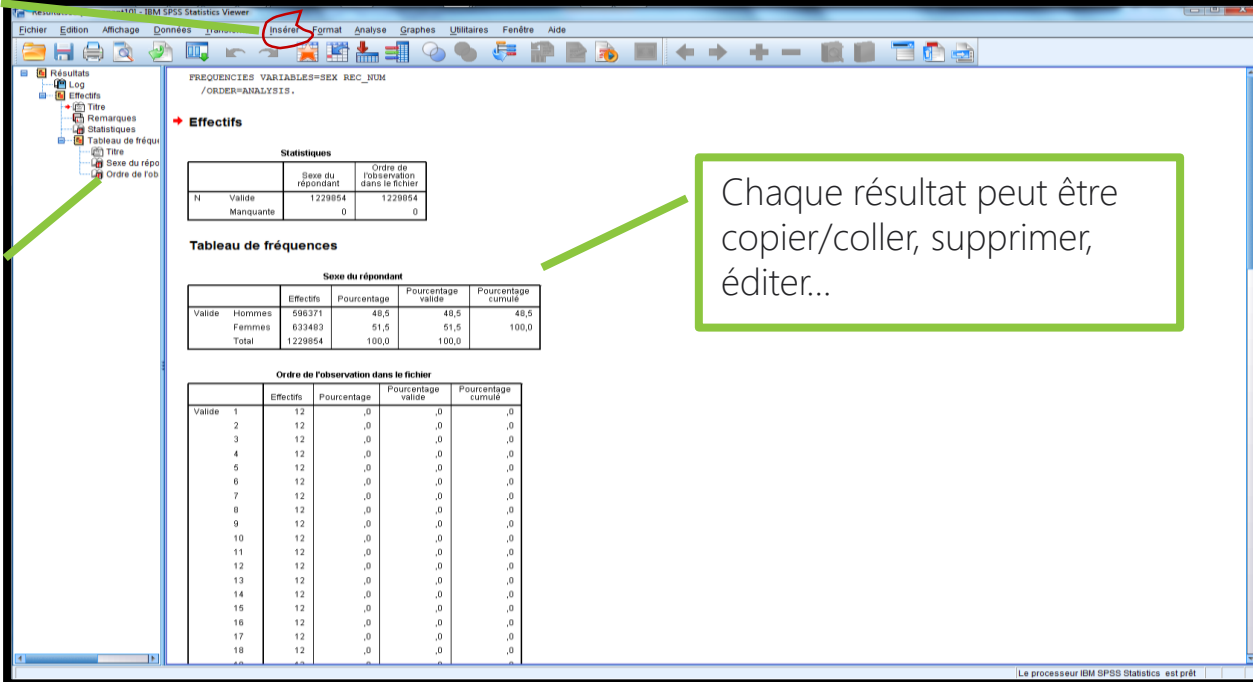
	REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	ED76TO89	EDUC90	MJH	EVERWORK	FTPTLAST
1	1	2007	1	1	13	4	1	2	2	6	-1	1	1	-1	-1
2	2	2007	1	6	35	2	11	-1	1	1	-1	1	-1	2	-1
3	3	2007	1	6	13	4	12	-1	1	1	-1	0	-1	2	-1
4	4	2007	1	6	13	4	12	-1	2	1	-1	1	-1	2	-1
5	5	2007	1	2	46	4	8	-1	1	1	-1	6	1	-1	-1
6	6	2007	1	6	59	3	11	-1	2	1	-1	4	-1	3	-1
7	7	2007	1	1	48	4	8	-1	2	5	-1	5	1	-1	-1
8	8	2007	1	1	47	4	6	-1	1	1	-1	2	1	-1	-1
9	9	2007	1	6	24	4	2	4	2	6	-1	5	-1	1	1
10	10	2007	1	4	48	4	2	4	1	6	-1	5	-1	2	-1
11	11	2007	1	2	48	4	4	-1	1	1	-1	4	1	-1	-1
12	12	2007	1	1	48	4	8	-1	1	6	-1	4	1	-1	-1
13	13	2007	1	1	48	4	2	4	2	6	-1	4	1	-1	-1
14	14	2007	1	6	13	4	1	2	2	6	-1	3	-1	1	2
15	15	2007	1	1	12	4	10	-1	2	1	-1	4	1	-1	-1
16	16	2007	1	2	35	4	6	-1	2	1	-1	5	1	-1	-1
17	17	2007	1	6	47	4	12	-1	2	3	-1	2	-1	2	-1
18	18	2007	1	6	35	4	2	4	1	6	-1	3	-1	3	-1
19	19	2007	1	1	12	4	7	-1	1	2	-1	4	1	-1	-1
20	20	2007	1	1	12	4	4	-1	2	6	-1	4	1	-1	-1
21	21	2007	1	6	35	4	6	-1	2	4	-1	4	-1	2	-1
22	22	2007	1	1	24	4	1	2	1	6	-1	2	1	-1	-1
23	23	2007	1	1	35	4	7	-1	1	1	-1	2	1	-1	-1
24	24	2007	1	6	46	4	11	-1	2	1	-1	1	-1	2	-1
25	25	2007	1	1	35	4	7	-1	2	1	-1	2	1	-1	-1
26	26	2007	1	1	46	4	6	-1	2	2	-1	1	1	-1	-1
27	27	2007	1	2	24	4	10	-1	1	1	-1	1	1	-1	-1
28	28	2007	1	6	35	4	11	-1	1	1	-1	4	-1	2	-1
29	29	2007	1	1	24	4	7	-1	2	1	-1	4	1	-1	-1
30	30	2007	1	6	24	4	8	-1	1	2	-1	4	-1	2	-1
31	31	2007	1	1	47	4	3	6	1	1	-1	4	1	-1	-1
32	32	2007	1	6	35	4	12	-1	1	1	-1	2	-1	2	-1
33	33	2007	1	1	10	4	6	-1	2	1	-1	4	2	-1	-1
34	34	2007	1	6	35	4	4	-1	1	6	-1	4	-1	2	-1
35	35	2007	1	6	10	4	12	-1	2	1	-1	1	-1	2	-1
36	36	2007	1	6	11	4	10	-1	2	1	-1	4	-1	2	-1

4.2.2. SPSS Viewer (résultats)

- Résultats des commandes effectuées > tableaux, graphiques.
- Fichier qui peut être édité et enregistré sous le nom de son choix. L'enregistrement des résultats se fait dans un fichier distinct (fichier **.spv**) de la base de données (fichier **.sav**).
- Les résultats peuvent être copiés/collés dans un document texte (clic droit).

Onglet Insérer : édition de la feuille de résultats (seul onglet distinct de la fenêtre de la base de données)

Document map > « table des matières »: permet de repérer, sélectionner, copier, supprimer des résultats



The screenshot shows the IBM SPSS Statistics Viewer window. The 'Insérer' menu is highlighted in the top toolbar. The left sidebar shows the 'Document map' with a tree structure including 'Résultats', 'Log', 'Effetifs', 'Titre', 'Remarques', 'Statistiques', 'Tableau de fréquences', 'Sexe du répondant', and 'Ordre de l'ob'. The main area displays the 'Tableau de fréquences' output for the variable 'Sexe du répondant'. The table shows counts and percentages for 'Hommes' and 'Femmes'.

		Sexe du répondant	
		Effectifs	Pourcentage
N	Valide	1229854	100,0
	Manquante	0	0

		Sexe du répondant		Ordre de l'observation dans le fichier	
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
N	Valide	1229854	100,0	100,0	100,0
	Manquante	0	0	0	0

Chaque résultat peut être copier/coller, supprimer, éditer...



Il n'est pas recommandé de copier/coller tels quels les sorties SPSS dans vos travaux. Une mise en forme est nécessaire pour respecter les **normes scientifiques de présentation de tableaux** (analyses descriptives, matrices de corrélation) > Consulter les **normes de présentation de tableaux et graphiques** dans un guide méthodologique.

4.2.3. Éditeur de syntaxe

- Fichier texte (.sps) où inscrire le *code* des commandes à réaliser (mise en forme, analyses, ...):
Fichier > Nouveau > Syntaxe
- Les commandes SPSS peuvent être faites de **2 façons**:
 1. Par le biais des options du menu du haut;
 2. En écrivant la **ligne de commande** dans l'éditeur de syntaxe puis en cliquant sur **Exécuter**
- **Avantages**: garder un historique des commandes, automatiser, assurer la reproductibilité, ... Certaines commandes ne sont possibles que par syntaxe.

❗ Compromis: Option COLLER

À partir des commandes du menu du haut, il est possible de copier le code de la commande à exécuter dans l'Éditeur de syntaxe en cliquant sur le bouton **Coller** avant de cliquer sur **OK** pour lancer la commande.

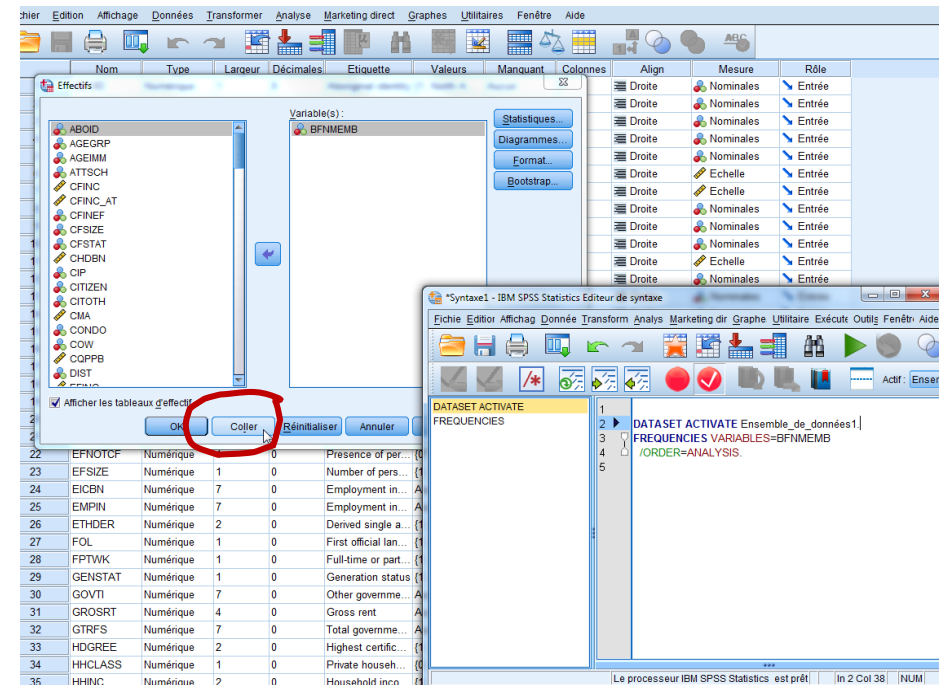
Ex: commande d'un tableau croisé

{

CROSSTABS

```
/TABLES=AGEGRP BY WAGES  
/FORMAT=AVALUE TABLES  
/STATISTICS=CHISQ CORR  
/CELLS=COUNT ROW TOTAL  
/COUNT ROUND CELL.
```

Commenter sa syntaxe en commençant une ligne par *



4.3. Impression et sauvegarde

Impression:

- Chaque fenêtre peut-être imprimée en totalité ou en partie (sélection)
- Préférable d'utiliser **l'Aperçu avant impression** (menu Fichier)

Sauvegarde:

- **Syntaxe:** Enregistrer/Enregistrer sous > **.sps**
- **Résultats:**
 - Enregistrer/enregistrer sous > **.spv** (ou .htm)
 - Exporter > pour enregistrer l'ensemble ou une sélection de résultats en différents formats dont pdf, xls, ppt. (raccourci: clic droit sur un tableau > Exporter)
- **Base de données:**
 - Enregistrer (ctrl + s): enregistrement **.sav** des modifications apportées à l'éditeur de données
 - Enregistrer sous/Exporter: enregistrer la base de données en différents formats: spss, excel (perte d'information), SAS, Stata, ...

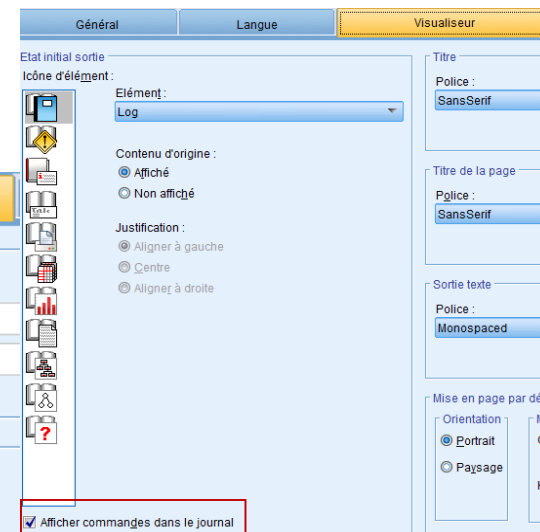
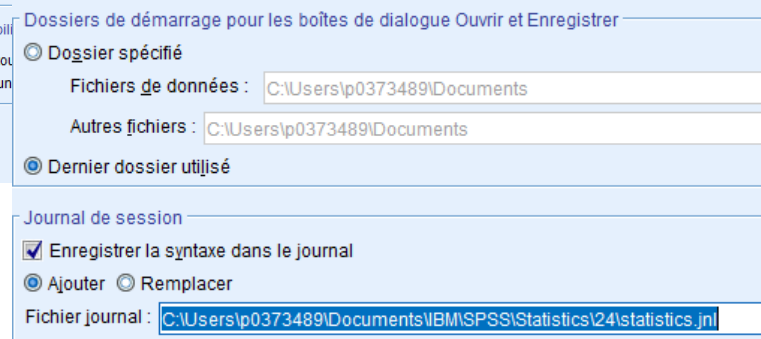
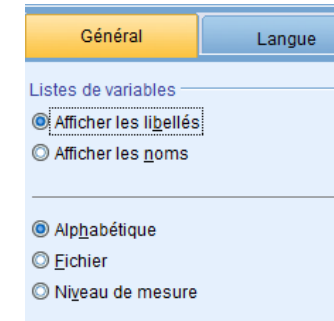
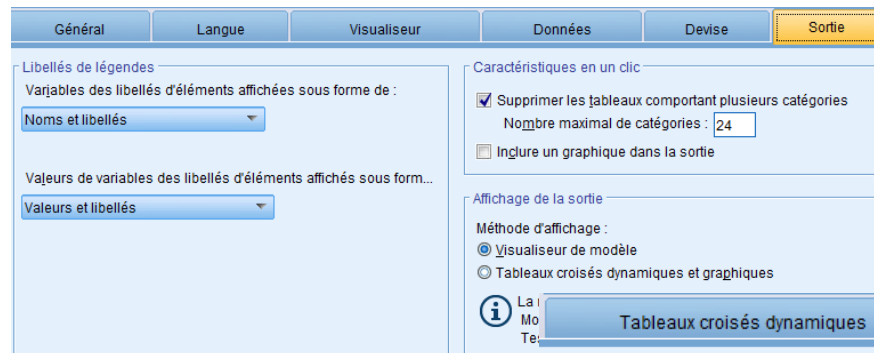


Toujours garder une copie de sa base originale en sécurité

5.1. Principales fonctions -> Menu Édition

Configuration de l'environnement SPSS

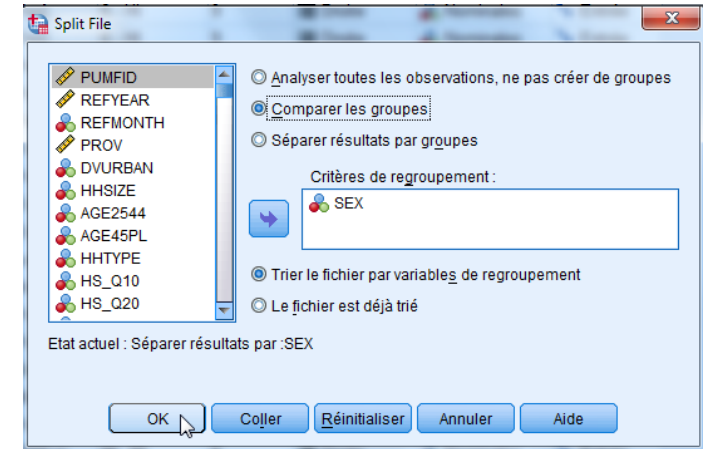
- Faire apparaître les **codes ET noms de valeurs** dans les tableaux.
- Changer la langue de l'interface.
- Changer les formats de tableaux.
- Afficher les commandes effectuées dans les résultats.
- Enregistrement automatique de la syntaxe dans journal.



5.2. Principales fonctions -> Menu Données

Modifications et requêtes sur le fichier de données

- **Fractionner en fichiers** : diviser les observations sur la base des valeurs d'une variable en fichiers distincts.
- **Scinder un fichier** : diviser les résultats des analyses subséquentes en fonction des valeurs d'une variable catégorielle (ex: sexe, âge, ...) [les données doivent d'abord être triées par la var de groupe].
Exemple: diviser tous les résultats par sexe
- **Sélectionner des observations** : sélectionner un échantillon ou sous-groupe d'observations sur lesquelles seront réalisées les traitements statistiques (les autres cas peuvent être conservés ou supprimés).
Ex: limiter les analyses aux répondants du Québec
- **Pondérer les observations** : permet d'associer un poids à chaque observation.
- **Fusionner des fichiers**: ajouter des observations ou variables
- **Agréger**: agréger l'info quantitative selon une nouvelle unité d'observation (en fonction des valeurs d'une variable choisie)



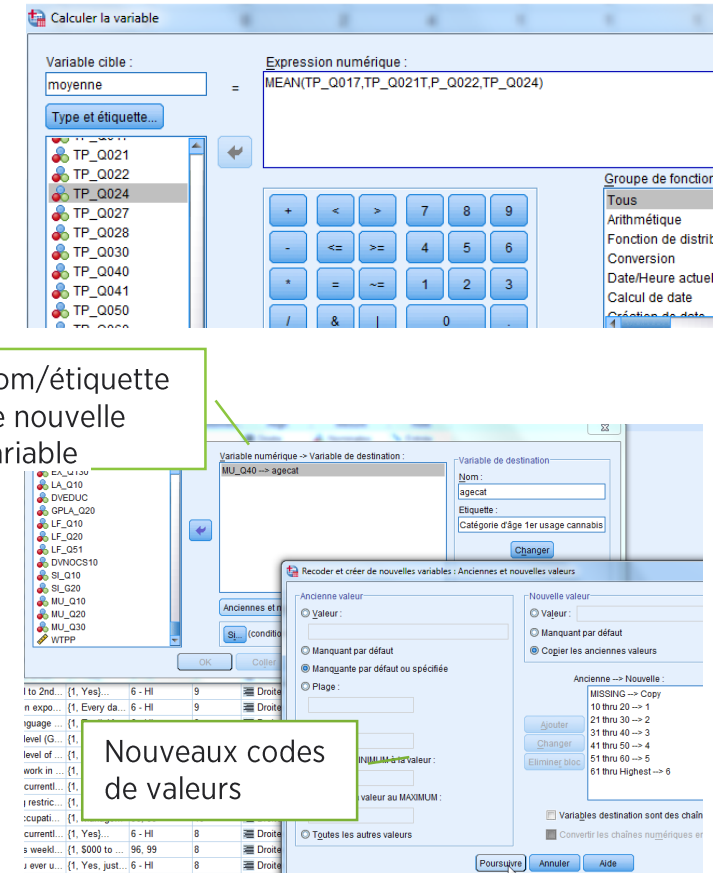
5.3. Principales fonctions -> Menu Transformer

Manipulation et création de variables

- **Transformer > Calculer la variable** : permet de créer une nouvelle variable à partir de calculs effectués sur des variables existantes. Par exemple, variable calculant la somme ou la moyenne de plusieurs résultats d'examens.
- **Transformer > Création de variable**: permet de modifier les valeurs d'une variable – par exemple, fusionner les catégories d'âge ou de revenu, recoder une variable continue en variable catégorielle, variable dichotomique/binaire, valeurs manquantes et extrêmes, transformation logarithmique, ...
- **Regroupement en classes visuelles**: outil visuel pour recoder variable continue en variable catégorielle.



Une fois recodée, toujours vérifier le résultat en faisant un tableau de fréquences.



[La fonction **Recoder des variables** écrase la variable existante – à éviter!]

5.4. Principales fonctions -> Menu Analyse

Statistiques descriptives

Créer des tableaux statistiques et graphiques servant à décrire et analyser des variables quantitatives et qualitatives pour explorer les données, les niveaux de mesure, les valeurs manquantes et erratiques, observer le nombre et le pourcentage de cas pour chaque valeur de variable, s'assurer de leur qualité, normalité, effectuer les pré-tests nécessaires aux analyses inférentielles, ...

Statistiques descriptives univariées – Procédures:

12.1 Analyse > Statistiques descriptives > Fréquences (nominales et ordinales + échelle)

12.2 Analyse > Statistiques descriptives > Descriptives (échelle)

12.3 Analyse > Statistiques descriptives > Explorer (échelle)

Statistiques descriptives bivariées – Procédures:

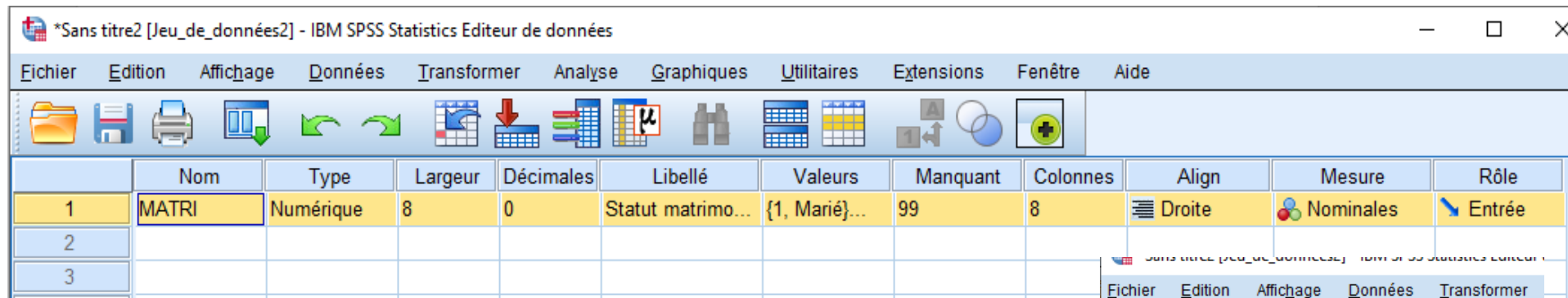
12.4 Analyse > Statistiques descriptives > Tableau croisé (2 var. catégorielles)

12.5 Analyse > Comparaison de moyenne (1 var. catégorielle / 1 var. échelle)

12.6 Analyse > Corrélation (2 var. échelle)

6. Création d'une base de données

- ⇒ **Fichier > nouveau > données**: saisir les informations sur les variables (Vue des variables) puis les données brutes (Vue des données).
OU
⇒ **Importer un jeu de données** (xls, csv, ...) et compléter les informations dans la vue des variables.



Vue des variables ↗

Vue des données ↘

4 ÉTAT MATRIMONIAL
Cochez «X» un seul cercle.

☐ Jamais légalement marié
☐ Légalement marié (et non séparé)
☐ Séparé, mais toujours légalement marié
☐ Divorcé
☐ Veuf ou veuve

	MATRI	var	var
1	1		
2	2		
3	3		
4	2		
5	3		
6	2		
7	99		
8	.		
9	3		
10			
11			

6. Création d'une base de données (suite)

	Eichier	Edition	Affichage	Données	Transformer	Analyse	Graphiques	Utilitaires	Extensions	Fenêtre	Aide
	Nom	Type	Largeur	Décimales	Libellé	Valeurs	Manquant	Colonnes	Align	Mesure	Rôle
1	MATRI	Numérique	8	0	Statut matrimo...	{1, Marié}...	99	8	Droite	Nominales	Entrée
2											

- **Nom**: Donner un nom court et significatif, sans espace, éviter les caractères spéciaux et les accents. Lettre comme 1^{er} caractère. 64 caractères max.
- **Type**: privilégier un codage **numérique** et non alphanumérique/chaîne de caractères (*string*).
- **Libellé (étiquette)**: descriptif au long de la variable.
- **Valeurs**: toujours attribuer des **codes numériques** et associer une étiquette aux valeurs **ordinales, nominales** et manquantes (reste vide pour les variables échelle).
- **Manquant**: définir les codes de valeurs manquantes (ex: 9, 99, 999)
- **Mesure**: définir le type de mesure (échelle, ordinale, nominale)

→ La fonction **Recoder automatiquement (Transformer)** permet de recoder une variable texte en variable numérique. Les valeurs alphanumériques seront recodées par ordre alphabétique par des codes à partir du chiffre 1.

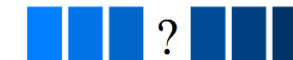
Par ex: 1 = « Femme »

2 = « Homme »

	Sexe1chaîne	Sexe2chaîne	Sexenum
1	Femme	F	1
2	Homme	H	2
3	Homme	H	2
4	Femme	F	1
5	Femme	F	1

→ Les valeurs d'une variable à **réponses multiples** doivent codées en variables distinctes dichotomiques (0/1). Celles-ci pourront ensuite être agrégées avec la fonction **Analyse > Réponses multiples > Définir des jeux de variables**.

6.1. Les valeurs manquantes



2 types de valeurs manquantes:

1. Codes de valeurs définis par l'utilisateur :

97 - Refus

98 - Ne s'applique pas

99 - Ne sait pas

2. Cellules vides (SYSMISS)

➔ Valeurs **exclues des analyses** (on pourrait aussi vouloir les conserver).

NB. Chaque logiciel gère les valeurs manquantes à sa façon. Ex: Stata= ., .a, .b..., R= NA

Essentiel de faire un bilan des valeurs manquantes:

- Sont-elles bien codées ?
- Sont elles trop nombreuses ?
- Problème du biais de non réponse (totale ou partielle) ?



88888888

Revenu : Revenu d'emploi

Revenu : Revenu total

88

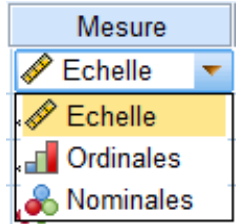
Scolarité : Plus haut certificat, diplôme ou grade

Travail : Travail en 2015

Traitement des données manquantes: ex. par techniques d'imputation ([voir StatCan](#))

Manquant	Non disponible	4725	,5
	Système	5	,0
	Total	4730	,5
Total		887012	100,0

6.2. Échelles de mesure



Catégorielles

Nominale: Variable qualitative dont les valeurs/modalités représentent des catégories sans classement, ordre ou relations hiérarchiques entre elles. Nombre limité de valeurs. Par exemple sexe, état matrimonial, province, ...

Ordinale: Variable qualitative ou quantitative dont les valeurs représentent des catégories associées à un classement. La codification de la variable respecte l'ordre des valeurs. Nombre limité de valeurs. Par exemple: niveaux de satisfaction (0 à 5), niveau d'éducation, catégories d'âge, de revenu, ...

Échelle

Intervalle/ratio : variable quantitative dont les valeurs ne sont pas regroupées en catégories. Permet donc de mesurer la distance exacte entre les valeurs. Ex: l'âge en années et le revenu exact en dollars.

Complexité croissante

Ordinale

Revenu annuel brut en 2015

- 1 - Moins de 10 000 \$
- 2 - 10 000 \$ À 19 999 \$
- 3 - 20 000 \$ À 29 999 \$
- 4 - 30 000 \$ À 39 999 \$

...

Revenu annuel brut en 2015

- 52 500 \$
- 31 280 \$
- 12 187 \$
- 86 200 \$
- ...

Échelle

7. Nettoyer et préparer les données

→ Toujours débiter par un examen approfondie de sa base de données (distributions de fréquence, graphiques).

→ La structure des données doit correspondre aux prérequis des analyses prévues:

- De quels niveaux de mesure sont les variables? (ordinales, nominales, échelle)
- Est-ce que les différents types de valeurs manquantes sont bien codés?
- Y a-t-il des valeurs problématiques (non prévues, erratiques/aberrantes, extrêmes?)
- La distribution des valeurs apparait-elle normale?
- Y a-t-il assez de cas pour procéder aux analyses voulues?
- Est-ce que certaines variables devraient être éliminées, recodées, transformées?

Plusieurs analyses présupposent, par exemple, la normalité des observations. Cette normalité doit être vérifiée, surtout si l'échantillon est petit, à l'aide de statistiques descriptives, de graphiques (histogrammes, boîtes à moustaches, qq plot), ou de tests (Kolmogorov-Smirnov, Shapiro-Wilk...).



Règle d'or: Garbage In, Garbage out!

8. Quelques mots sur les postulats

Pour choisir un test statistique, on tient compte: 1) des caractéristiques de ses données et de son échantillon (format, variance, normalité, ...) et 2) de ses objectifs (analyser les relations entre les variables ou comparer des groupes?)

Les tests statistiques reposent sur différents **postulats relatifs aux données** qu'il faut vérifier.

Exemples:

- indépendance des observations (sélection aléatoire)
- multicollinéarité (lien trop fort entre vars indépendantes)
- distribution normale (+ résidus - qualité de la prédiction des valeurs)
- hétéroscédasticité (variance de la prédiction)

Certains tests sont plus contraignants que d'autres...

- **Tests paramétriques** (anova, corrélation, régression, test T, ...): échantillon aléatoire indépendant, distribution normale, variance égale (test de Levene), min de 30 sujets par groupe.
- **Tests non paramétriques** [Analyse > Tests non paramétrique]: alternatives lorsque les postulats ne sont pas remplis (échantillon trop petit, distribution asymétrique, valeurs extrêmes), qui ne reposent pas sur la moyenne et se servent du rang des observations au lieu des valeurs brutes (ex: Wilcoxon, Krustall-Walis, Friedman, Fisher, Chi-2...)

Arbre décisionnel pour sélection de tests statistiques:

- http://pagesped.cahuntsic.ca/sc_sociales/psy/methosite/consignes/decision.htm
- <http://dl.icdst.org/pdfs/files1/ce2418fcc89682f2d0905bcb6ad93d9a.pdf>

9. Importer un fichier de données en format .csv

9. Importer un fichier de données en format .csv

Démarche :

- Télécharger le tableau *Portrait quotidien des cas confirmés répartis par région, groupe d'âge et sexe (MSSS)*: <https://www.donneesquebec.ca/recherche/dataset/covid-19-portrait-quotidien-des-cas-confirmes>
- Fichier > Ouvrir > Données (sélectionner le format **.csv**)
- Paramétrer les options de l'assistant d'importation: modifier les délimiteurs entre les variables pour la **virgule**.

Quelques bonnes pratiques de structuration de fichiers	
Chaque colonne est une variable	Attention aux valeurs nulles et manquantes
Chaque rangée est une observation	Attention aux formats de dates (peut être préférable de diviser en 3 colonnes)
Ne pas combiner d'information dans une cellule	Un seul tableau par onglet
Première ligne (en-tête de colonne) composée des noms de variables	Un seul onglet par feuille
Noms de variables descriptifs sans espace, caractères spéciaux	Enregistrer en format ouvert (csv)*
Pas de commentaires/notes dans les cellules	Documenter ses données dans un fichier dans même dossier (description des variables, valeurs possibles, questions, codes de réponse...)
Attention au format <i>long</i> vs <i>wide</i> (ex: données de panel)	Documenter tous changements/modifications

10. Exercices

1. Tableau de fréquence (+ pondération)

2. Sélectionner des sous-groupes.
3. Recoder une variable catégorielle
4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

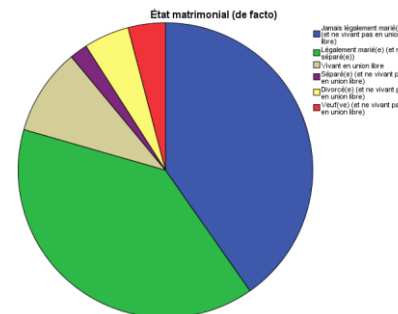
10.1. Tableau de fréquence (+ pondération)

- Un **tableau de fréquences (ou de distribution)** permet de connaître le nombre et la proportion (%) d'effectifs (répondants) dans chaque catégorie de **variables catégorielles**.
- Pour les **variables continues**: permet d'obtenir les mesures 1) de tendance centrale, 2) de distribution (forme), 3) de dispersion (variation) et 4) de position (**bouton Statistiques**).
- Permet de créer des diagrammes. Par ex: histogrammes pour **données continues** (option courbe normale) et pointes de tarte pour **données catégorielles**.

Raccourci SPSS : il est possible d'accéder aux statistiques descriptives via un clic droit sur n'importe quelle cellule de données ou variables (mais moins d'options).

État matrimonial (de facto)					
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Jamais légalement marié(e) (et ne vivant pas en union libre)	357251	40,3	40,3	40,3
	Légalement marié(e) (et non séparé(e))	347427	39,2	39,2	79,4
	Vivant en union libre	84541	9,5	9,5	89,0
	Séparé(e) (et ne vivant pas en union libre)	17412	2,0	2,0	90,9
	Divorcé(e) (et ne vivant pas en union libre)	44110	5,0	5,0	95,9
	Veuf(ve) (et ne vivant pas en union libre)	36271	4,1	4,1	100,0
	Total	887012	100,0	100,0	

Var. ordinales/
nominales



Var. échelle

Statistiques

EMPIN Revenu - Revenu d'emploi

N	Valide	
	Manquant	
Moyenne		29942,46
Médiane		14000,00
Mode		0
Ecart type		50609,340
Variance		2561305258
Asymétrie		6,769
Erreur standard d'asymétrie		,003
Kurtosis		84,857
Erreur standard de Kurtosis		,006
Plage		1096295
Minimum		-50000
Maximum		1046295
Somme		21980128860
Percentiles	10	,00
	25	,00
	50	14000,00
	75	45000,00

10.1. Tableau de fréquence (+ pondération)

- Combien y a-t-il d'immigrants dans la base de données?

Démarche :

- Analyse > Statistiques descriptives > Fréquences;

Variable : Immigration - Statut d'immigrant [IMMSTAT].

=> 202 320

Échantillon vs population ?!

- Activer la variable de pondération et refaire la démarche précédente.
- Combien y a-t-il d'immigrants au Canada? Quel pourcentage de la population canadienne représentent-ils?

Démarche :

- Données > Pondérer les observations > Facteur de pondération pour les particuliers;
- Analyse > Statistiques descriptives > Fréquences;

Variables :

- Facteur de pondération pour les particuliers [WEIGHT];
- Immigration - Statut d'immigrant [IMMSTAT].

=> 7 493 197
21,8%

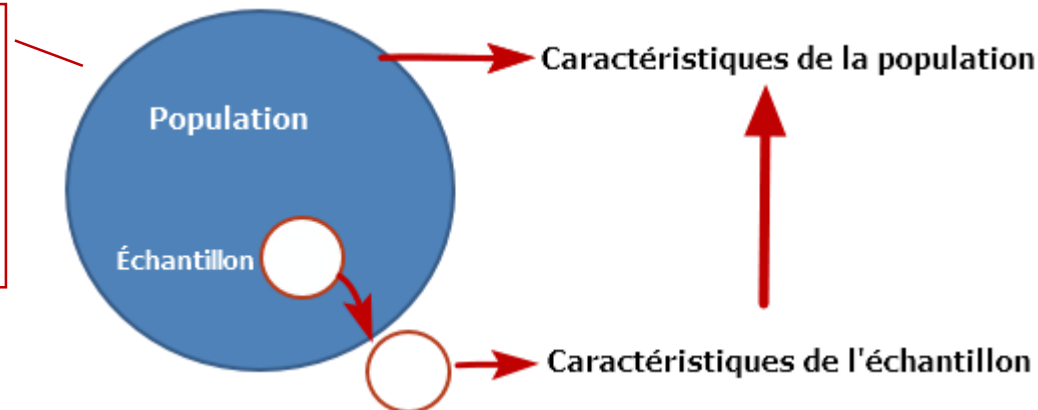
10.1. Tableau de fréquence (+ pondération)

L'estimation des caractéristiques de la population à partir d'une enquête repose sur l'hypothèse selon laquelle chaque unité échantillonnée représente, en plus d'elle-même, un certain nombre d'unités non échantillonnées dans la population.

❗ Pour les enquêtes de Statistique Canada, il faut toujours pondérer les résultats avant d'en rendre compte.

Variable(s) de poids fournies dans le fichier d'enquête (poids déterminés selon la méthode d'échantillonnage: rééquilibrer en cas de suréchantillonnage, ajuster pour la non réponse, calibrer selon estimations démographiques...)

MAIS! Sensibilité des tests au nombre de répondants = création d'un **poids normalisé** (moyenne 1) par règle de 3
Calculer la variable:
 $\text{var de poids} / \text{nb répondants pondérés} * \text{nb répondants réels}$



Estimation: tirer des conclusions sur la population en utilisant des poids et des données d'échantillon

=

Relier l'information de l'échantillon à la population de l'enquête

« Les utilisateurs doivent s'assurer de ne pas diffuser des estimations non pondérées ni de faire des analyses fondées sur des données non pondérées du fichier parce que les résultats non pondérés ne sont pas représentatifs de la population, mais de l'échantillon ». [FMGD du recensement de 2016 - guide de l'utilisateur](#)

1. Tableau de fréquence (+ Pondérer des résultats)

2. Sélectionner des sous-groupes

3. Recoder une variable catégorielle
4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

10.2. Sélectionner (filtrer) des observations

- Limiter l'échantillon à la population du Québec.
- Les immigrants représentent quel pourcentage de la population au Québec ?
- Créer un graphique circulaire des données avec les pourcentages affichés.

Démarche:

- Données > Sélectionner des observations > Selon une condition logique – Si... **Province = 24** ;
- Analyse > Statistiques descriptives > Fréquences + Bouton Graphique > Graphiques circulaires & Pourcentages ;
- Double cliquer sur le graphique > clic droit > Afficher les libellés de données.

Variables :

- **Filtre**: Province ou territoire de résidence actuelle (2016) [PR= 24].
- **Fréquence**: Immigration : Statut d'immigrant [IMMSTAT]

=> 13,6%

10.2. Sélectionner (filtrer) des observations

- Limiter l'échantillon aux Québécois de 25 à 64 ans qui ont travaillé en 2015
- Quel pourcentage de cette population possède un diplôme universitaire (BACC minimum) ?
- Créer un graphique à barres (avec %) pour illustrer cette distribution.
- Quel est le niveau de scolarité le plus fréquent ?

Démarche :

- Données > Sélectionner des observations > Selon une condition logique – Si...
- Analyse > Statistiques descriptives > Fréquences;
- Bouton Statistiques > **Cocher Mode** ;
- Cliquer sur le bouton **Graphiques** > cocher Graphiques à barres et Pourcentage.;
- Vérifier le résultat avec trois tables de fréquence (PR, WRKACT et Âge);
- Analyse > Statistiques descriptives > Fréquences > HDEGREE.

Variables :

- Filtre:
 - Province ou territoire de résidence actuelle (2016) [PR=24].
 - Travail : Travail en 2015 [WRKACT > 2].
 - Âge [AGEGRP > 8 & < 17].
- **Fréquence**: Scolarité : Plus haut certificat, diplôme ou grade [HDEGREE].

=> 27,7%


1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes

3. Recoder une variable catégorielle

4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

10.3. Le recodage de variables -> quelques exemples

Recoder une variable d'échelle en variable catégorielle

 Revenu
13000
260000
43000
30000
86000
8000
4000

Valeur	Libellé
1	0 à 20000
2	20001 à 40000
3	40001 à 60000
4	60001 à 80000
5	80001 à 100000
6	plus de 100000

Combiner des catégories de réponses ou inverser des échelles

Valeur	Libellé
1	Aucun certificat, diplôme ou grade
2	Diplôme d'études secondaires ou attestation d'équivalence
3	Certificat ou diplôme d'une école de métiers autre qu'appren
4	Certificat d'apprenti ou Certificat de qualification
5	Programme d'une durée d'au moins trois mois mais inférieu
6	Programme d'une durée de un à deux ans
7	Programme d'une durée de plus de deux ans
8	Certificat ou diplôme universitaire inférieur au baccalauréa
9	Baccalauréat
10	Certificat ou diplôme universitaire supérieur au baccalauréa
11	Diplôme en médecine, en médecine dentaire, en médecine
12	Maîtrise
13	Doctorat acquis

Valeur	Libellé
1	Primaire ou moins
2	Secondaire
3	Collégiale
4	Universitaire (1er)
5	Universitaire (2-3)

Créer des variables factices

à partir d'une variable catégorielle ou d'échelle

Valeur	Libellé
1	Premières Nations (Indiens de l'Amérique du Nord)
2	Métis
3	Inuk (Inuit)
4	Réponses autochtones multiple
5	Réponses autochtones non comprises ailleurs
6	Identité non autochtone

Valeur	Libellé
0	Non autochtones
1	Autochtones

10.3. Recoder une variable catégorielle

- Créer une variable dichotomique du plus haut niveau de scolarité en divisant les répondants entre ceux qui ont un diplôme universitaire complété (BACC minimum) et les autres.

Démarche :

- Transformer > Création de variables > Scolarité - Plus haut certificat, ...;
- Donner un nouveau Nom et libellé a la nouvelle variable (ex : RHDGREE–Universitaires) > Changer ;
- Entrer les Anciennes et nouvelles valeurs : **1 à 8 = 0** & **9 à 13 = 1** + **88** et **99** (manquantes) ;
- Dans la vue des variables, ajouter les libellés de valeurs et déclarer valeurs manquantes ;
- Faire un **tableau de fréquence** avec l'ancienne et la nouvelle variable.

Variable :

- Scolarité - Plus haut certificat, diplôme ou grade HDGREE.

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle

4. Tableau croisé

5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

10.4. Les tableaux croisés

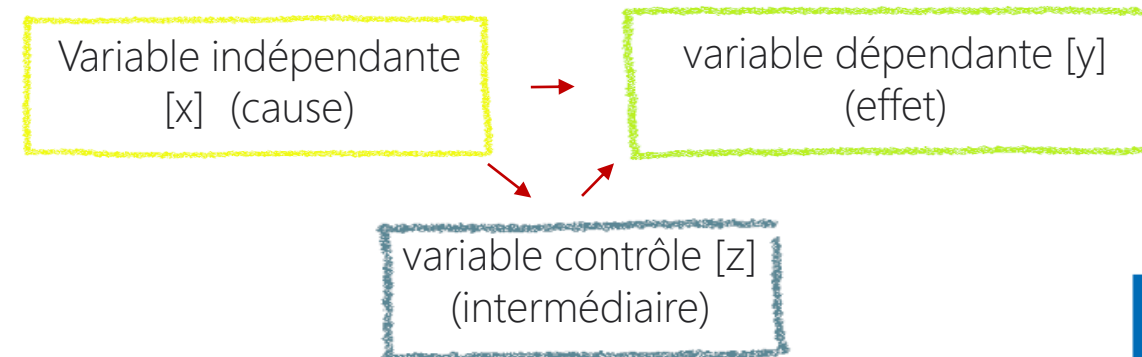
Statistiques bivariées : pour examiner les **relations** entre **variables catégorielles** (nominales ou ordinales)

-> ventiler les valeurs d'une variable en fonction d'une autre.

Tableau croisé SEX Sexe * FOL Langue - Première langue officielle parlée

			FOL Langue - Première langue officielle parlée				
			1 Anglais seulement	2 Français seulement	3 Anglais et français	4 Ni anglais ni français	Total
SEX Sexe	1 Femme	Effectif	333853	103547	4764	9432	451596
		% dans SEX Sexe	73,9%	22,9%	1,1%	2,1%	100,0%
	2 Homme	Effectif	324543	99688	5099	6086	435416
		% dans SEX Sexe	74,5%	22,9%	1,2%	1,4%	100,0%
Total		Effectif	658396	203235	9863	15518	887012
		% dans SEX Sexe	74,2%	22,9%	1,1%	1,7%	100,0%

- Permet également de croiser 2 variables en tenant compte d'une 3e variable catégorielle. Par ex: examiner la relation entre la consommation de cannabis et l'âge en tenant compte du sexe (Strate = variable contrôle [z]).



10.4. Les tableaux croisés

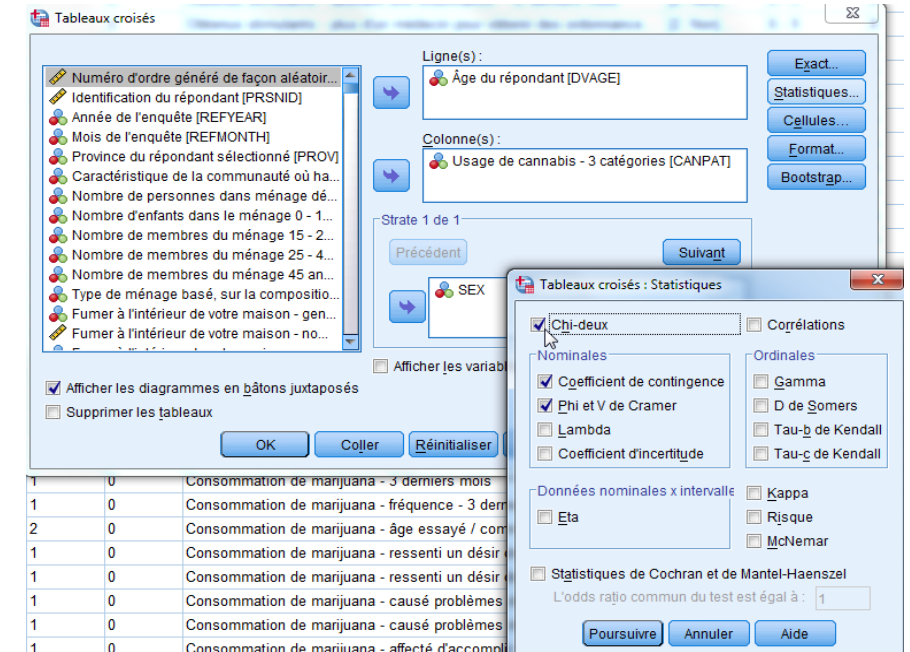
Mesures d'association: déterminer s'il y a relation entre variables (**signification**), le **sens** et la **force** de celle-ci.

Tableaux croisés (2 var. catégorielle)

Comparaison de moyenne (1 var. catégorielle / 1 var. échelle)

Corrélation (2 var. échelle)

- Bouton **Statistiques**: Khi-deux, corrélations, Phi et V de Cramer, Coefficient de contingence, ...
- Bouton **Cellules**: % (côté de la variable indépendante)



- Nombre qui évalue la force de la relation entre 2 variables (au-delà des comparaisons de % dans un tableau croisé).
- Il existe des dizaines de mesures d'association (C de Pearson, V de Cramer, Gamma, Phi, ...) allant généralement de 0 à 1 pour les var. catégorielles et de -1 à +1 pour les var. d'échelle.
- Choix dépend, entre autres, du niveau de mesure, du nombre de catégories, du nombre de cas.

10.4. Les tableaux croisés: Chi2 et valeur p

- Permet le **test du Chi2** (var nominales ou ordinales, effectif min de 5 par cellule): test de validation d'hypothèses qui permet de déterminer s'il existe une relation significative entre les variables, ie qui n'est pas due au hasard. Significative = généralisable à la population totale.
- Résultat influencé par le nb d'observations (effet du nombre) et **ne dit rien sur la force de la relation**.
- La **valeur *p* ou sig (signification)** = si *p* est inférieur à **0,05 (généralement)**, il y a une relation significative et donc on **rejette l'hypothèse nulle** (H0) selon laquelle les variables sont indépendantes, ie pas d'association.
- Calcul basé sur la différence entre fréquence attendue et observée.
- Le **Sig** accompagnant les tests statistiques s'interprète toujours de la même façon: évaluation (en %) du risque de se tromper (**ex < 5%**) en disant que la relation observée dans l'échantillon peut être généralisée à la population entière.
- Si la relation est significative, l'option **Statistiques** permet de choisir les coefficients précisant la relation entre les variables: 1) signification > 2) sens > 3) force.
- **Coefficients d'association dérivées du Chi2**: Phi (élimine effet de taille, 2x2), V Cramer (tableau + de 2x2), de contingence, mesurent la force de la relation (entre 0 et 1).

Tests du khi-deux			
	Valeur	ddl	Signification asymptotique (bilatérale)
khi-deux de Pearson	642,841 ^a	3	,000
Rapport de vraisemblance	648,318	3	,000
Association linéaire par linéaire	229,955	1	,000
N d'observations valides	887012		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 4841,54.

10.4. Les tableaux croisés

- Les immigrants sont-ils proportionnellement plus nombreux que les non immigrants à posséder un diplôme universitaire au Canada ?
- Chez les hommes et chez les femmes?

Démarche :

- Analyse > Statistiques descriptives > Tableaux croisés > Colonne : IMMSTAT / Ligne : RHDGREE
- Couche : Sexe ;
- Bouton Cellules > Pourcentage [Colonne=position].

Variables :

- Immigration - Statut d'immigrant [IMMSTAT];
- Universitaires [RHDGREE].
- Couche: Sexe [SEX].

=> Oui!

=> 41,9%

variable dépendante en ligne / variable indépendante en colonne (%)

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle
4. Tableau croisé

5. Calculer une variable

6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

10.5. Calculer une variable (+ procédure *Descriptives*)

- Créer une nouvelle variable faisant la somme des 3 variables de revenus de Prestations [EICBN], [CQPPB], [CHDBN]
- Quelle est la moyenne de ces revenus?

Démarche :

- Transformer > Calculer la variable ;
- Donner un nom à la nouvelle variable cible : PRESTATIONS ;
- Créer l'expression numérique pour faire la somme des trois variables avec la commande **SUM**;
- Analyse > Statistiques descriptives > Descriptives

Variables :

- Revenu : Prestations d'assurance-emploi (AE) [EICBN]
- Revenu : Prestations du Régime de rentes du Québec (RRQ) [CQPPB]
- Revenu : Prestations pour enfants [CHDBN].

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle
4. Tableau croisé
5. Calculer une variable

6. Tableau de variables d'échelle

7. Comparer des moyennes de groupes
8. Corrélation
9. Graphiques à barres

10.6. Tableau de variables d'échelle: *Explorer & Descriptives*

Présentent **les caractéristiques d'une variable quantitative** regroupant les mesures de tendance centrale, dispersion et de distribution (pas de fréquence): moyenne, minimum, maximum, écart-type, variance, intervalle, valeurs standardisées (*score Z) ...

Statistiques descriptives												
	N	Plage	Minimum	Maximum	Somme	Moyenne	Ecart type	Variance	Skewness		Kurtosis	
	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques	Erreur std.	Statistiques	Erreur std.
CAPGN Revenu - Gains ou pertes en capital nets	734079	1108235	-50000	1058235	603648253	822,32	17582,767	309153691,3	36,749	,003	1655,161	,006
N valide (liste)	734079											

Mesures de dispersion:

- **Étalement des valeurs:**
 - **Étendue (plage):** distance entre le minimum et maximum.
- **Variabilité des valeurs:**
 - **Écart-type:** distance de chaque valeur à la moyenne (+ est grand, plus données sont hétérogènes)
 - **Variance:** Écart-type au carré.
- **Homogénéité:**
 - **CV:** écart-type divisé par la moyenne * 100 (+ CV est petit (près de 0), + données sont homogènes, en %)
- **Mesures de distribution (forme de la courbe):**
 - **Kurtosis:** coefficient d'aplatissement, mesure la concentration des résultats, ie l'aplatissement de la courbe.
 - **Skewness:** coefficient de symétrie, donne un indice de la normalité de la forme de la courbe (droite ou gauche, symétrie parfaite: moyenne, médiane et mode au même endroit). Normale = 0.
- ***Score z (valeurs standardisées):** nombre d'écart-type séparant observation de la moyenne (score enregistré dans nouvelle variable). $(Nb - \text{moyenne}) / \text{écart type}$. Pour comparer variable sur même échelle.

10.6. Tableau de variables d'échelle: *Explorer & Descriptives*

> Selon les modalités d'une variable catégorielle

- Comparer la moyenne de revenu d'emploi des 2 groupes Universitaires et Non universitaires [scinder un fichier]

Démarche :

Données > Scinder un fichier > Comparer les groupes > Critères de regroupement > Scolarité - Plus haut certificat, diplôme ou grade ;
Analyse > Statistiques descriptives > Descriptives.

Variables :

Revenu : Revenu d'emploi [Empln] ;

Scolarité : Plus haut certificat, diplôme ou grade [HDGREE].

- Comparer et visualiser (boîtes à moustache, histogramme) les caractéristiques du Revenu total chez les hommes et les femmes

Démarche :

- Analyse > Statistiques descriptives > Explorer ;

Variables :

- Revenu : Revenu total [Totinc_AT] ;

- Sexe : Sexe [SEX].

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle
4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle

7. Comparer des moyennes de groupes

8. Corrélation
9. Graphiques à barres

10.7. Comparer des moyennes de groupes

Évaluer si des **groupes** ont des **moyennes différentes** : 1 variable catégorielle + 1 continue (ou ordinale)

- Permet d'obtenir des statistiques sommaires par groupe (Moyenne, écart type, tableau Anova, ...)
- Les tests de différence de moyennes – est-ce que les différences sont significatives ou non? (choix du test selon postulats et nombre de groupes) :
 - **Test T**: Comparer moyennes d'une variable d'échelle (ou ordinale) de 2 gr. (dichotomique) = Statistiques descriptives, **Test de Levene** (différence de variances), valeurs t et intervalle de confiance 95% pour différence de moyenne.
 - **Test F (ANOVA)**: extension du test T pour + de 2 groupes (différence significative entre la moyenne la + haute et la + basse (contrastes pour intermédiaires).
- Les graphiques à barres et boîtes à moustaches permettent de visualiser et comparer ces statistiques entre groupes.

Test d'homogénéité des variances

TOTINC Revenu total

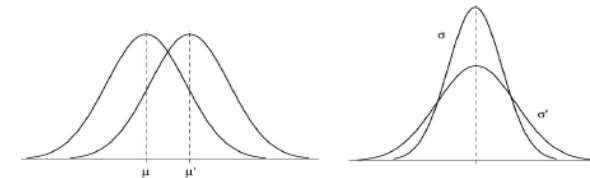
Statistique de Levene	ddl1	ddl2	Sig.
14465,269	1	734077	,000

ANOVA

TOTINC Revenu total

	Somme des carrés	ddl	Carré moyen	F	Sig.
Inter-groupes	5,007E+13	1	5,007E+13	16040,122	,000
Intragroupes	2,292E+15	734077	3121850277		
Total	2,342E+15	734078			

Pour savoir si la différence de moyenne est significative (ie généralisable à la population), il faut déterminer si les variances des gr sont égales= Test de Levene (test F sur écarts à moyenne)



Test des échantillons indépendants										
		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes						
		F	Sig.	t	ddl	Sig. (bilatéral)	Différence moyenne	Différence erreur standard	Intervalle de confiance de la différence à 95 %	
									Inférieur	Supérieur
EMPIN Revenu - Revenu d'emploi	Hypothèse de variances égales	17536,887	,000	-126,445	734077	,000	-14783,441	116,916	-15012,592	-14554,289
	Hypothèse de variances inégales			-124,503	534509,102	,000	-14783,441	118,740	-15016,167	-14550,714

10.7. Comparer des moyennes de groupes

- Comparer la moyenne de revenu d'emploi selon les catégories d'identité autochtone

Démarche:

- Analyse > Comparer les moyennes > Moyennes

Variables:

- Revenu: Revenu d'emploi [Empln]
- Immigration: Statut d'immigration [IMMSTAT]

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle
4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes

8. Corrélation

9. Graphiques à barres

10.8. Corrélation

Calculer l'intensité et le sens d'une relation entre deux **variables continues (ou ordinales)**.

- Coefficient de corrélation de Pearson

Coefficient (r): test statistique (paramétrique) pour mesurer le lien entre deux variables quantitatives. Indice qui décrit la **force** de la relation **linéaire** entre deux variables > varie entre -1 et 1> Plus la **valeur est proche de +1 ou -1**, plus les 2 variables sont associées fortement. Absence de lien si 0

Revenu après impôt	Corrélation de Pearson	1	,300**
	Sig. (bilatérale)		,000
	N	53474	53474
Nombre d'années d'études complétées par la personne (élémentaire/secondaire/post secondaire)	Corrélation de Pearson	,300**	1
	Sig. (bilatérale)	,000	
	N	53474	53474

** . La corrélation est significative au niveau 0.01 (bilatéral).

$p < .05$ = seuil de signification (test d'hypothèse) – détermine si ce lien (r) est **significatif**, i.e. la corrélation observée entre X et Y dans l'échantillon existe bel et bien dans la population ou est due au hasard.

R de Pearson (degré de liaison)

- Corrélation parfaite si $r = 1$
- très forte si $r > 0,80$. (louche!)
- r entre 0,50 et 0,80: forte
- r entre 0,30 et 0,50: moyenne.
- r entre 0,10 et 0,30: faible.
- $r \leq 0,10$: absence de corrélation

NB. chaque domaine de recherche établit des seuils non officiels pour déterminer la force du lien

10.8. Corrélation

- Quel est le coefficient de corrélation entre le revenu d'emploi [Empln] et le niveau de scolarité

Démarche :

- Analyse > Corrélation> Bivariée

Variables:

- Revenu : Revenu d'emploi [Empln]
- Scolarité : Plus haut certificat, diplôme ou grade [HDGREE]

=> 0,277

$0,277^2 * 100 = \% \text{ de}$
la variance expliquée

- Est-ce que cette corrélation est la même pour les hommes et les femmes?

Démarche :

- Données > Scinder un fichier > Comparer les groupes > Sexe
- Analyse > Corrélation> Bivariée > relancer le même tableau

=> F: 0,378

=> H: 0,275

1. Tableau de fréquence (+ Pondérer des résultats)
2. Sélectionner des sous-groupes
3. Recoder une variable catégorielle en variable dichotomique
4. Tableau croisé
5. Calculer une variable
6. Tableau de variables d'échelle
7. Comparer des moyennes de groupes
8. Corrélation

9. Graphiques à barres

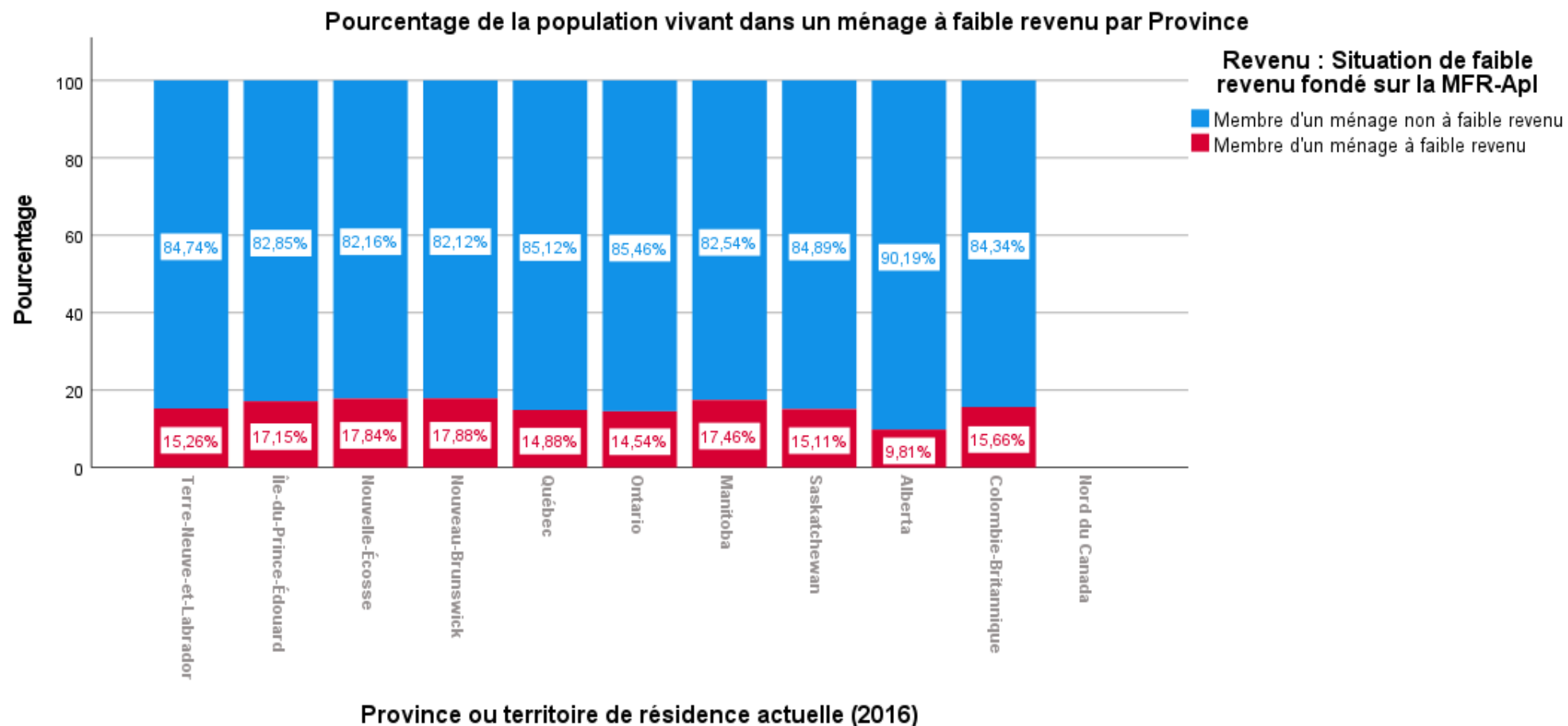
10.9. Diagramme à barres superposées

Observer à l'aide d'un diagramme à barres superposées les pourcentages de ménages vivant en situation de faible revenu (MFR-Apl) selon la province de résidence

Démarche :

- Supprimer la sélection d'observations ;
- Graphiques > Générateur de graphiques... ;
- Dans la galerie, cliquer sur l'icône du *Diagramme en Barres : superposé* ;
- Glisser la variable indépendante **Province** sur la barre des x et la variable dépendante **Revenu : Situation de faible revenu fondé sur la MFR-Apl** sur la case *Empiler* en haut à droite ;
- Dans la fenêtre *Propriétés des éléments* à droite, sous Modifier les propriétés de, sélectionner **Barres1** puis choisir dans l'encadré Statistiques *Pourcentage()* ;
- Cliquer sur le bouton *Définir les paramètres...* et choisir *Total pour la catégorie de chaque axe des x*.

10.9. Diagramme à barres superposées



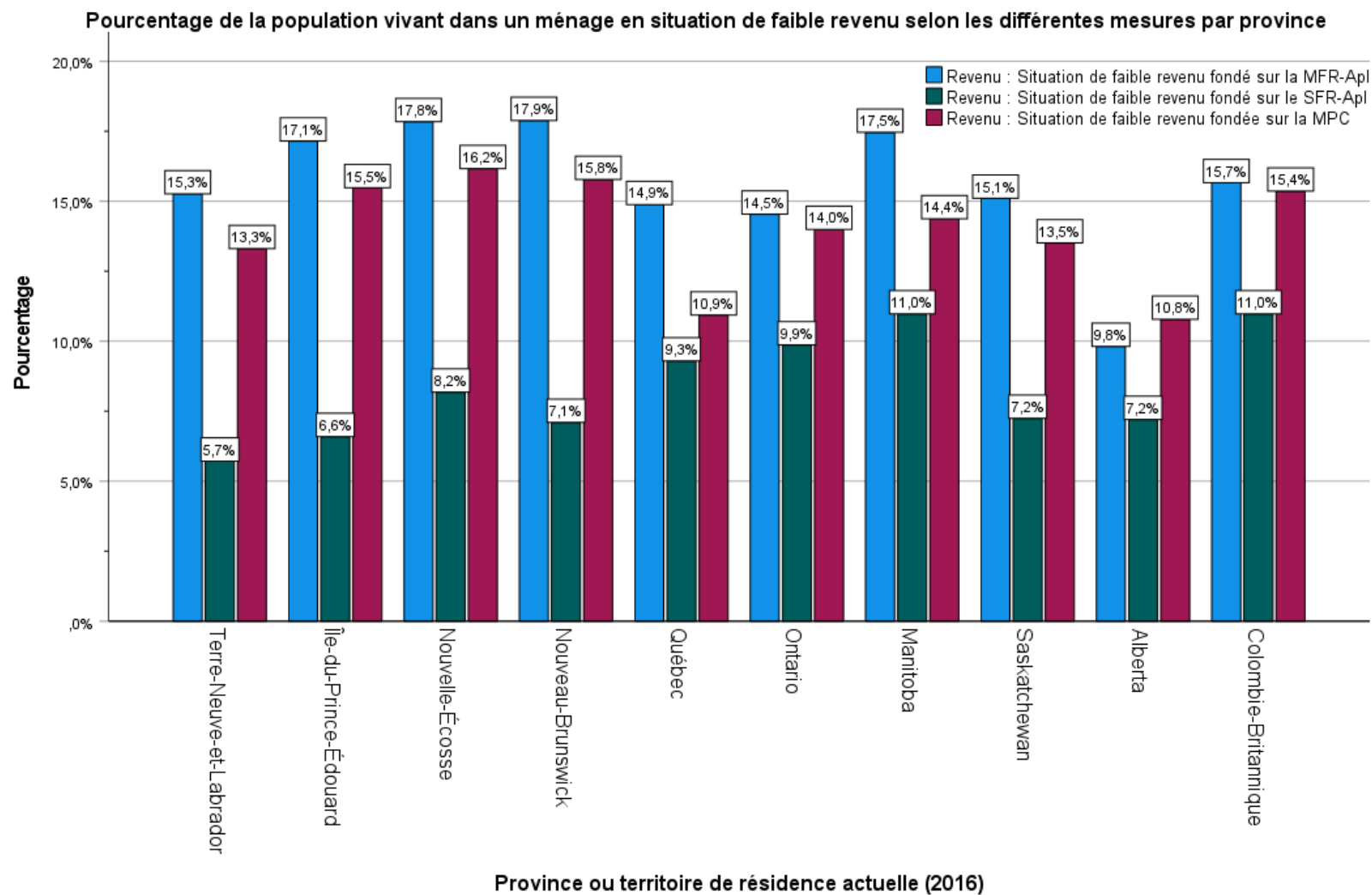
10.9. Diagramme à barres en grappes

Comparer dans un graphique à barres (en grappes) les trois mesures de faible revenu (MPC, MFR-Apl, MFR-Avl) par province.

Démarche :

- Graphiques > Boîtes de dialogue ancienne version > Barres > *En cluster* > *Récapitulatifs pour variables distinctes*.
- Les barres représentent : Ajouter les 3 variables - Revenu : Situation de faible revenu fondé sur la MFR-Apl, Revenu : Situation de faible revenu fondé sur la MFR-Avl, Revenu : Situation de faible revenu fondée sur la MPC.
- Sur chacune des variables, cliquer sur *Changer les statistiques* > Dans l'encadré *Valeur* cocher *Pourcentage au-dessus* et indiquer 1.
- Axe des catégories : variable indépendante **Province**.

10.9. Diagramme à barres en grappes



Merci!

Pour aller plus loin...

- [Capsules d'introduction à SPSS](#)
- [SPSS à l'Usherbrooke](#)
- [Guide d'utilisation SPSS \(cegep Ahuntsic\)](#)
- [Cours Claire Durand \(enregistrements\)](#)
- [Capsules prof Marc Ouimet](#)
- [SPSS Andy Fields](#)
- [SPSS Tutorials](#)
- [Numea \(\\$\)](#)
- [SPSS dans le catalogue Sofia](#)
- [Sage Research Methods \(SRM\)](#)

