C3. Décrire et relier des variables

Olivier Godechot

I. Décrire des variables

De l'importance de la description d'une variable

- Souvent présente en annexe des articles scientifiques
- Nécessaire avant toute mise en relations de variables
  - Bien connaître ses variables. Prendre connaissance de la distribution
  - Vérifier le codage (non réponses, manquantes, etc.)
  - Éviter les erreurs

TABLE A1
DESCRIPTIVE STATISTICS

Variable	Mean	SD	Min	Max
Crime Factors and Neighborhood Decline in Chicago (N = 2,796):				
Perception of neighborhood crime (factor score)	.01	.83	−1.34	2.19
Respondent black	.31	.46	.00	1.00
Respondent Latino	.06	.24	.00	1.00
Male	.34	.47	.00	1.00
Age	42.13	15.73	17.00	91.00
Education (years)	12.45	3.78	.00	20.00
Family income:				
< $10,000	.26	.44	.00	1.00
$10,000–$20,000	.35	.48	.00	1.00
$20,000–$30,000	.18	.38	.00	1.00
> $30,000	.13	.34	.00	1.00
Missing	.09	.28	.00	1.00
Personal victimization experience	.38	.49	.00	1.00
Crime rate (logged)	4.60	.53	3.30	7.17
...				

(Quillian and Pager 2001)

Les types de variables

- Deux grands types :
  - Quantitatives
    - Ex : taille, âge, poids, revenu
    - Toujours représenté par des chiffres
  - Catégorielles
    - Ex : sexe, PCS, etc
    - Représenté soit par des lettres soit par codes numériques (arbitraires).
    - Ex. 1 (=homme), 2 (=femme)

- Types de variables catégorielles
  - Variables nominales : Ex. Sexe
  - Variables ordonnées (relation d'ordre entre les modalités) Ex :
    - 1. Jamais, 2. Parfois. 3. Souvent, 4. Toujours
  - Variables Intervalles. Ex :
    - a. 1 fois par semaine ou plus, b. Entre 1 fois par semaine et une fois par mois, c. entre une fois par mois et une fois par an, d. Moins d'une fois par an

Transformation des variables

- De quantitative en catégorielle :
  - Ex. Tranches de revenu (âge, etc.)
    - Grand/ Petit. Riches/Pauvres
- De catégorielle en quantitative
  - Variables dichotomiques (0 ou 1)
  - Ex : Sexe → Deux variables : Homme, Femme (avec Homme=1-Femme)

Transformation des variables (2)

- De variables ordonnées en score
  - 1. Jamais, 2. Parfois. 3. Souvent, 4. Toujours
  - Valeurs : 1, 2, 3, 4
- Avantage :
  - exprime la relation d'ordre,
  - conserve l'hétérogénéité,
  - facilite les calculs
- Limite : arbitraire des valeurs

- De intervalles en score. Utiliser les valeurs numériques des bornes de tranche.
  - 1 fois par semaine ou plus, b. Entre 1 fois par semaine et une fois par mois, c. entre une fois par mois et une fois par an, d. Moins d'une fois par an
    - A la borne inférieure : 52 , 12 , 1, (0 ?)
    - A la borne supérieure : (365 ??) , 52 , 12, 1
    - Moyenne des bornes : (208.5?), 32, 6.5, (0.5)
- Avantage : similaires + exprime (imparfaitement) les grandeurs sous-jacentes. Limites moindres (arbitraire de la transformation)

Le tableau des effectifs

- Le tableau des effectifs [*Frequency table*]
  - Effectif pour chaque catégorie/valeur

    ⇒ Outil principal de description de variables catégorielles

    ⇒ Outil utile de description de variables quantitatives si nombre de valeurs peu important

I. — *Tableau des effectifs du clergé paroissial de Paris (curés compris) d'avril 1789 à janvier 1791.*

Le premier nombre inclut les ecclésiastiques qui ont assisté aux assemblées préliminaires pour les élections aux États généraux, mais ne figurent plus ensuite sur les listes de 1791; figurent par contre dans ce nombre certains ecclésiastiques dont on ne sait pas s'ils étaient déjà dans la paroisse en avril 1789. Le second nombre représente les effectifs en janvier 1791 au moment du serment.

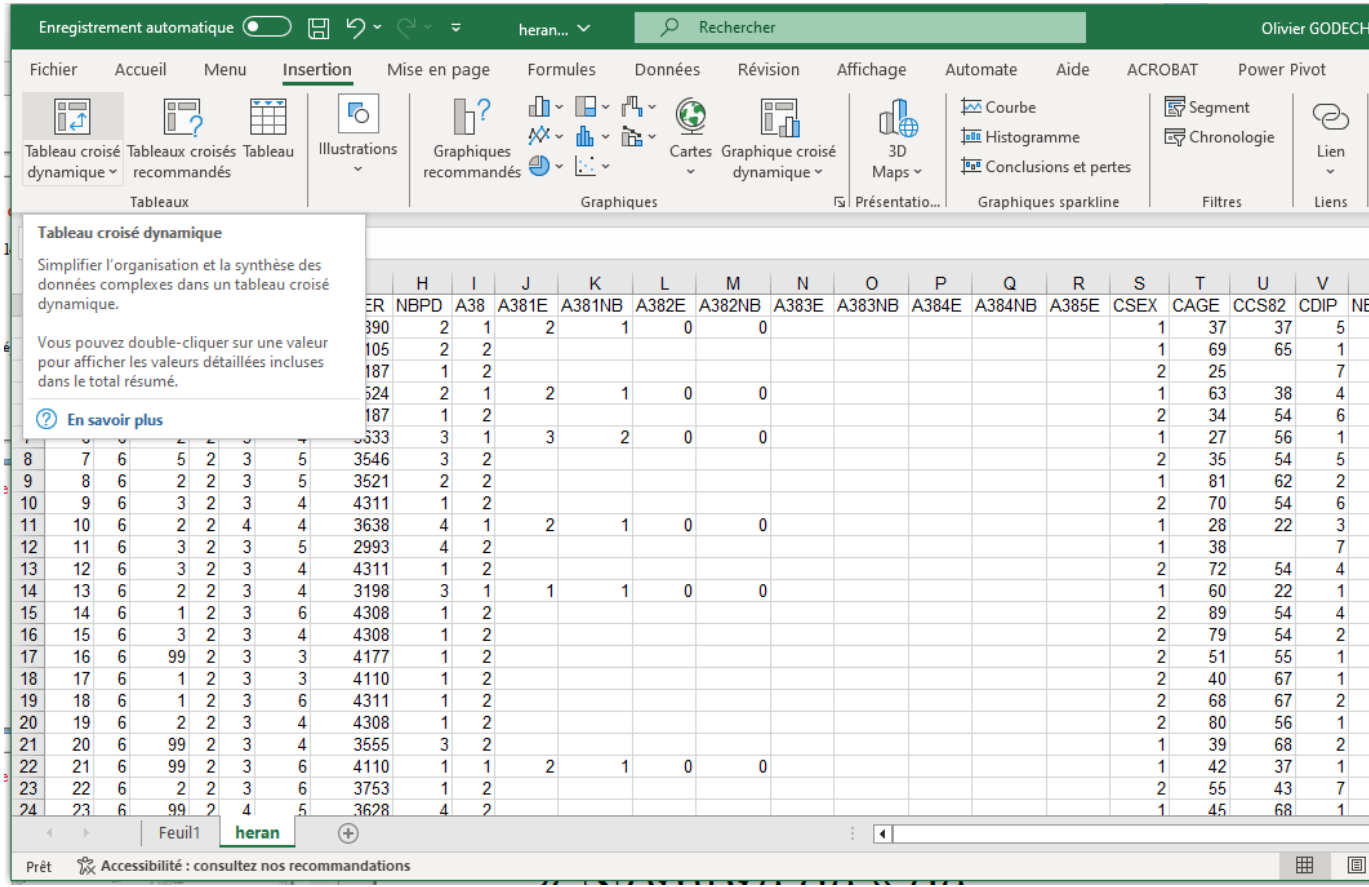
Paroisses	1789-1790	1791
Saint-André-des-Arts [1]	13	13
Saint-Barthélemy	14	14
Saint-Benoît	9	9
Sainte-Chapelle Basse	4	2
Saint-Côme	11	8
Sainte-Croix	2	3
Saint-Denis-du-Pas et Saint-Jean-Baptiste	2	2

...

(de Dainville-Barbiche 1989)

Tableau d'effectifs avec Excel

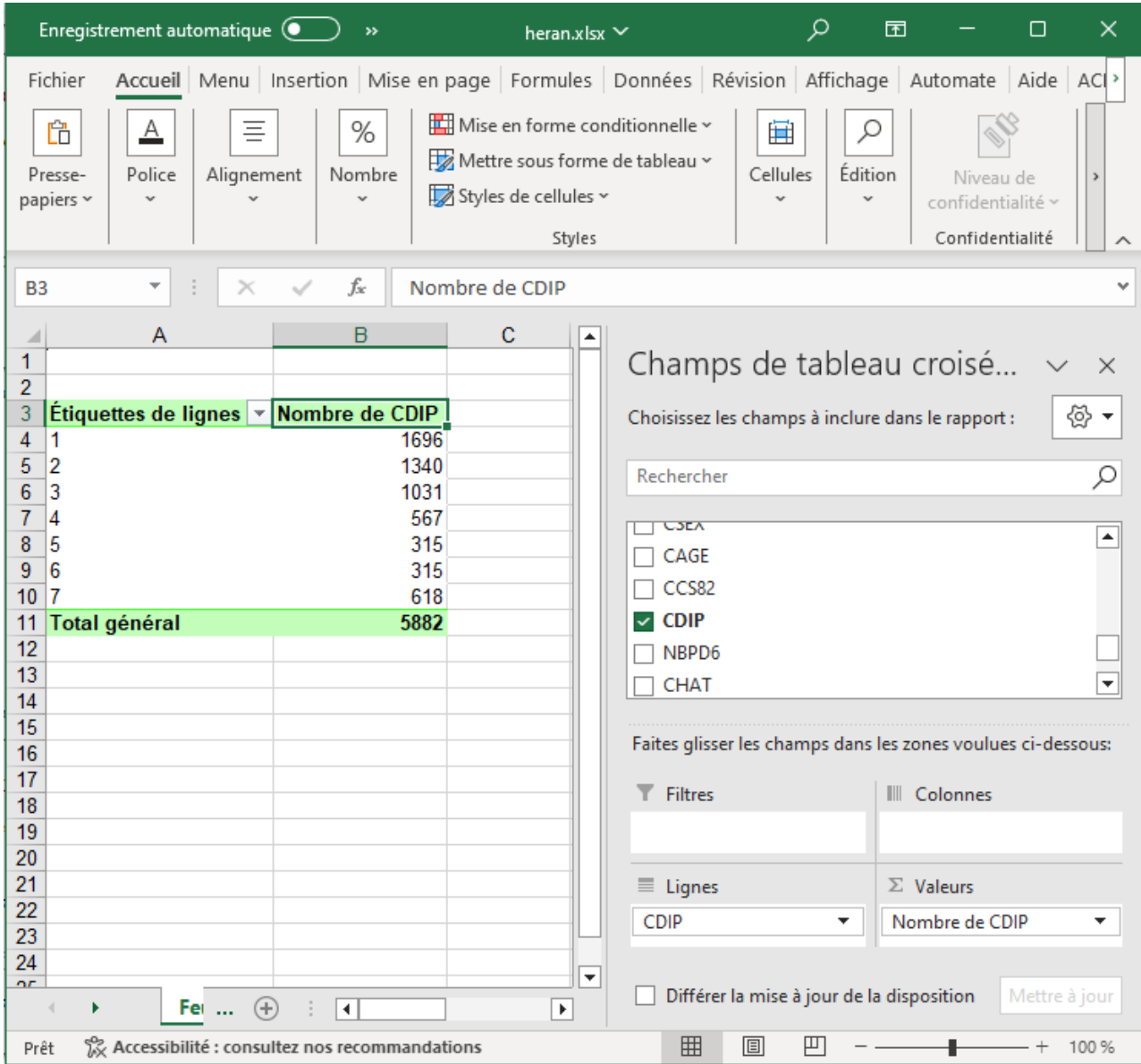
- Menu Insertion
- Tableau Croisé Dynamique



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	NBPD	A38	A381E	A381NB	A382E	A382NB	A383E	A383NB	A384E	A384NB	A385E	CSEX	CAGE	CCS82	CDIP
	2	1	2	1	0	0						1	37	37	5
	2	2										1	69	65	1
	1	2										2	25		7
	2	1	2	1	0	0						1	63	38	4
	1	2										2	34	54	6
	3	1	3	2	0	0						1	27	56	1
	3	2										2	35	54	5
	2	2										1	81	62	2
	1	2										2	70	54	6
	4	1	2	1	0	0						1	28	22	3
	4	2										1	38		7
	1	2										2	72	54	4
	3	1	1	1	0	0						1	60	22	1
	1	2										2	89	54	4
	1	2										2	79	54	2
	1	2										2	51	55	1
	1	2										2	40	67	1
	1	2										2	68	67	2
	1	2										2	80	56	1
	3	2										1	39	68	2
	1	1	2	1	0	0						1	42	37	1
	1	2										2	55	43	7
	4	2										1	45	68	1

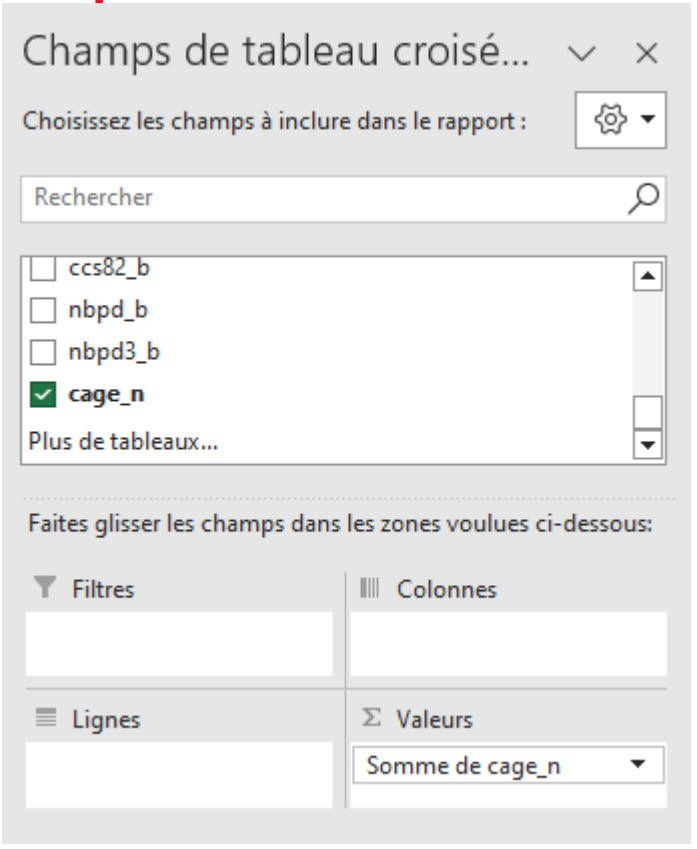
Tableau d'effectifs avec Excel

- Cocher la variable
- La faire glisser en « lignes »
- Changer (double clic) « Somme de » *ma_variable* pour « Nombre de » *ma_variable*



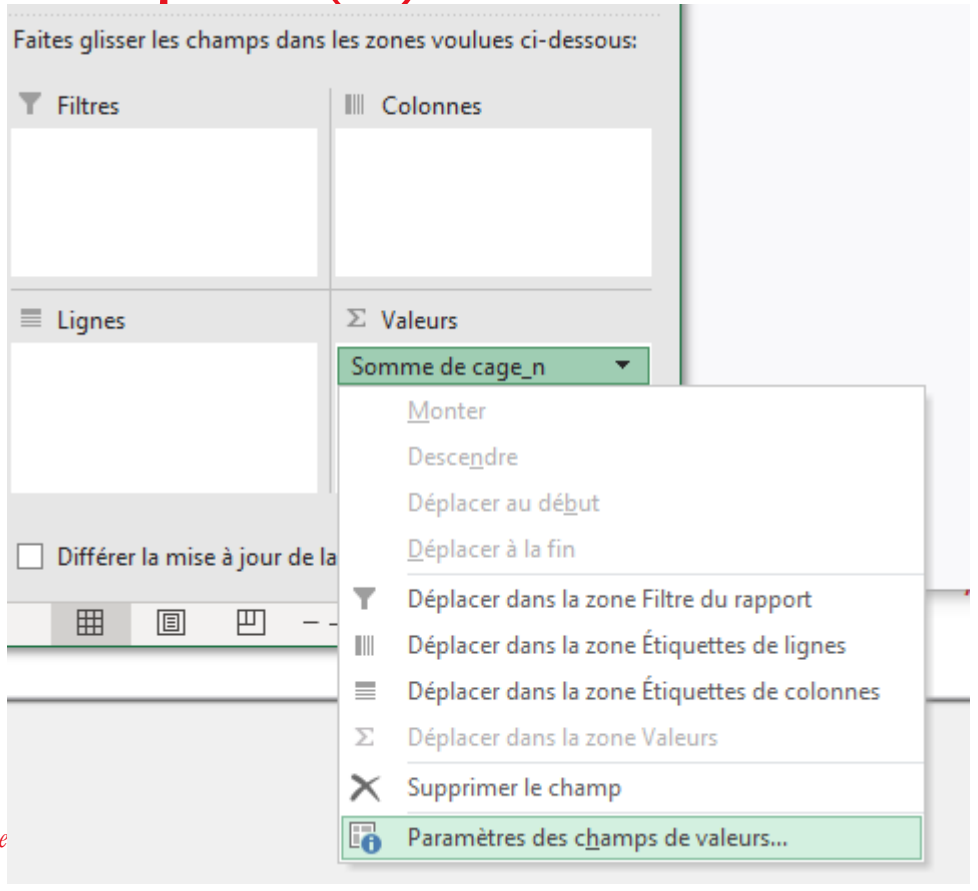
Addendum sur les champs de tableau croisé dynamique

- Les quatres champs dans lesquels on peut glisser des variables. Dans l'ordre d'importance
  - **Valeurs :** Variables pour laquelle on veut Statistiques (Somme, Dénombrement, moyennes, etc.)
  - **Lignes :** Glisser variable catégorielle pour avoir en ligne des statistiques pour différentes catégories
  - **Colonnes :** Glisser variable catégorielle pour avoir en ligne des statistiques pour différentes catégories
  - **Filtres :** Pour filtrer sur un champ particulier d'une variable



Addendum sur les champs de tableau croisé dynamique (2)

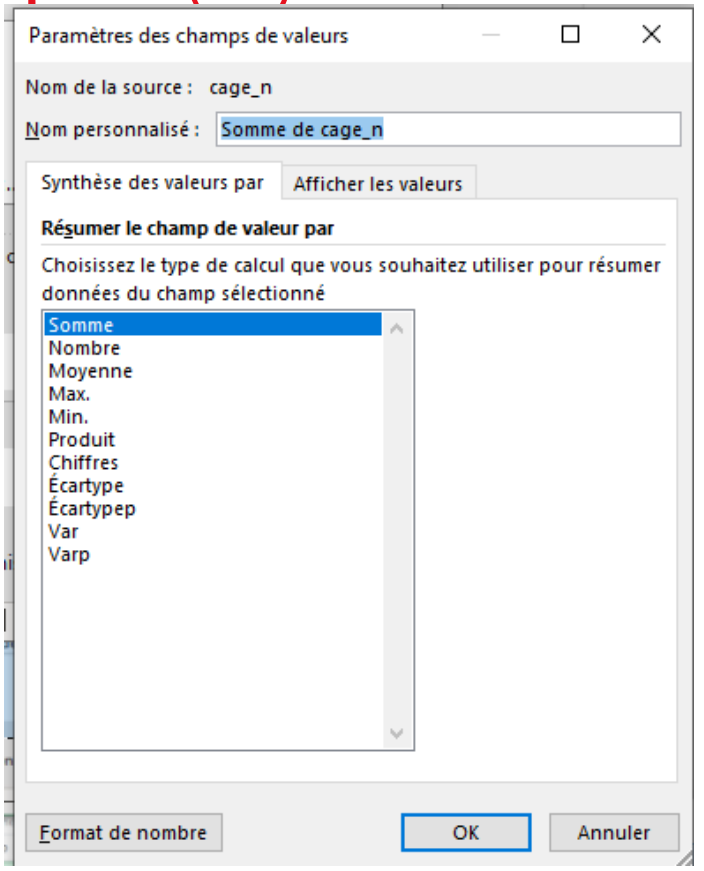
- Vérifier que la plage sélectionnée est la bonne
- Toujours mettre le tableau dynamique dans une nouvelle feuille
- Faire glisser une (ou plusieurs variables, ou plusieurs fois la même) dans « **Valeurs** »
- Double cliquer sur la variable
- Paramètres « **champs de valeur** »



Variables. Décrire

Addendum sur les champs de tableau croisé dynamique (3)

- Pour l'effectif :
  - Variable quantitatives : Utiliser « Chiffres »
  - Variable catégorielles : Utiliser « Nombre »
- Variance et Écart-type
  - Utiliser Var et Ecartype



Aller au-delà du tableau d'effectifs

- Tableau d'effectifs contient toute l'information sur une variable
- Mais effectifs peu lisibles
  - Transformer en pourcentage (du total)
- Trop d'information tue l'information !
  - Variables catégorielles → regrouper les modalités
  - Variables quantitatives → résumer l'information avec un indicateur du milieu (Moyenne, Médiane). Cf. Cours 1

Étiquettes de lignes	Nombre de CAGE
18	6
19	8
20	15
21	30
22	50
23	65
24	75
25	92
26	101
27	112
28	105
29	117
30	140
31	134
32	128
33	146
34	121
35	155
36	122
37	100
38	111
39	104
40	99
41	81
42	82
43	89
44	93
45	97
46	84
47	97
48	92
49	102
50	100
51	87
52	117
53	107
54	116
55	107
56	102
57	96
58	120
59	116
60	110
61	128
62	131
63	94
64	60
65	45

Les limites de la moyenne et l'importance de la dispersion

- Soit deux pays où le revenu moyen est de 30 000 euros par an
  - Pays a) Variation entre 20,000 et 40,000
  - Pays b) Variation entre 5,000 et 500,000
  - Pays équivalents? Préférence pour l'un ou l'autre ?
- Étudier la dispersion d'une variable quantitative
  - Connaître l'hétérogénéité, les inégalités de la variable
  - Représentativité de la moyenne

- Les quantiles
  - Les quartiles (Q1 à Q3)
    - 1er quartile : Seuil supérieur du quart du bas
    - 2ème quartile (= médiane). Seuil supérieur de la moitié du bas
    - 3ème quartile : Seuil supérieur des trois quarts du bas
  - Les déciles (D1 à D9)
    - D1 (seuil supérieur des 10 % du bas)
    - ...
    - D9 (seuil supérieur des 90 % du bas)
  - Les centiles : P1 à P99

Quantiles

- Fonctions pour les quartiles

  =QUARTILE(Plage ; 1)

  ...

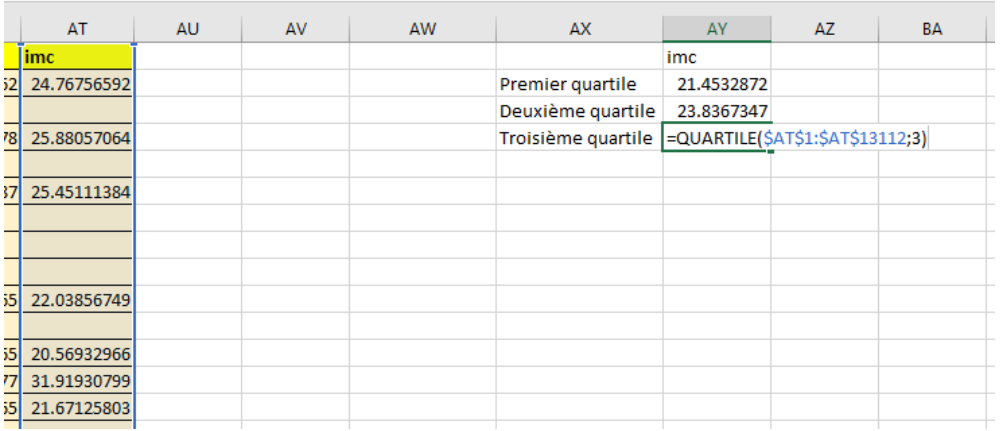
  =QUARTILE(Plage ; 3)

- Pour les déciles et centiles

  =CENTILE(Plage ; 0.1)

  =CENTILE(Plage ; 0.9)

  =CENTILE(Plage ; 0.99)



- Ratios classiques de dispersion et d'inégalités
  - Q3/Q1
  - D9/D1
  - P99/P50

Un indice synthétique de dispersion

- Variance : indicateur de dispersion. *Moyenne des carrés des écarts à la moyenne*

$$V(X) = \sum_{i=1}^{i=n} \frac{(X_i - \overline{X})^2}{n} \qquad V(X) = \sum_{i=1}^{i=n} \frac{(X_i - \overline{X})^2}{n-1}$$

Sur une population complète *Sur un échantillon (cas le + fréquent)*

- Ecart-type : Racine carrée de la variance

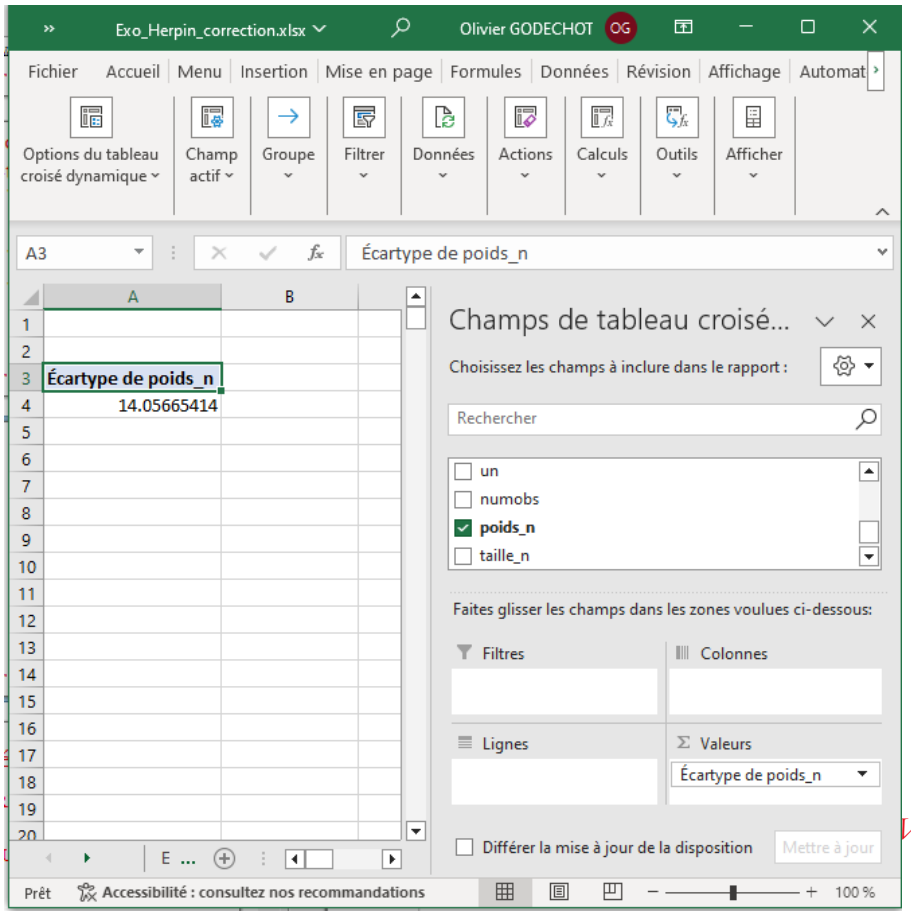
$$ET(X) = \sqrt{V(X)} = \sqrt{\sum_{i=1}^{i=n} \frac{(X_i - \overline{X})^2}{n}} \qquad ET(X) = \sqrt{V(X)} = \sqrt{\sum_{i=1}^{i=n} \frac{(X_i - \overline{X})^2}{n-1}}$$

Sur une population complète *Sur un échantillon (cas le + fréquent)*

Avantages et limites de la variance et de l'écart-type

- Prend en compte toute l'information : tous les écarts à la moyenne comptent
- L'échelle de la variance est peu lisible
- L'écart-type est à peu près sur la même échelle que la moyenne
  - En faveur de l'écart-type
- L'écart-type dépend de l'échelle de mesure (comme la moyenne).
  - Écart-type différent si on parle en euros ou en francs
  - Pas un indicateur d'inégalités
- Le ratio écart-type/moyenne est un indicateur d'inégalités
- Si la distribution d'une variable est normale (ex. La taille)
  - 95 % des observations se trouvent dans l'intervalle [moyenne-2*E-T ; moyenne+2*E-T]

Écarts-types (resp. Variance) avec Excel



- Avec tableau croisé dynamique
  - Sélectionner « Ecartype » dans Valeurs
  - Sélectionner « Variance » dans Valeurs
- Avec les formules

  =ECARTYPE(Plage)

  =VAR(Plage)

Description des variables. En bref

- Toujours mettre l'effectif total (hors valeurs manquantes)
  - Eventuellement l'effectif manquant
- Variables quanitatives
  - Moyenne, Ecart-type, Min, Max
  - Eventuellement : quartiles
- Variables catégoriques
  - Tableau d'effectifs

TABLE A1
DESCRIPTIVE STATISTICS

Variable	Mean	SD	Min	Max
Crime Factors and Neighborhood Decline in Chicago (N = 2,796):				
Perception of neighborhood crime (factor score)	.01	.83	−1.34	2.19
Respondent black	.31	.46	.00	1.00
Respondent Latino	.06	.24	.00	1.00
Male	.34	.47	.00	1.00
Age	42.13	15.73	17.00	91.00
Education (years)	12.45	3.78	.00	20.00
Family income:				
< $10,000	.26	.44	.00	1.00
$10,000–$20,000	.35	.48	.00	1.00
$20,000–$30,000	.18	.38	.00	1.00
> $30,000	.13	.34	.00	1.00
Missing	.09	.28	.00	1.00
Personal victimization experience	.38	.49	.00	1.00
Crime rate (logged)	4.60	.53	3.30	7.17

(Quillian and Pager 2001)

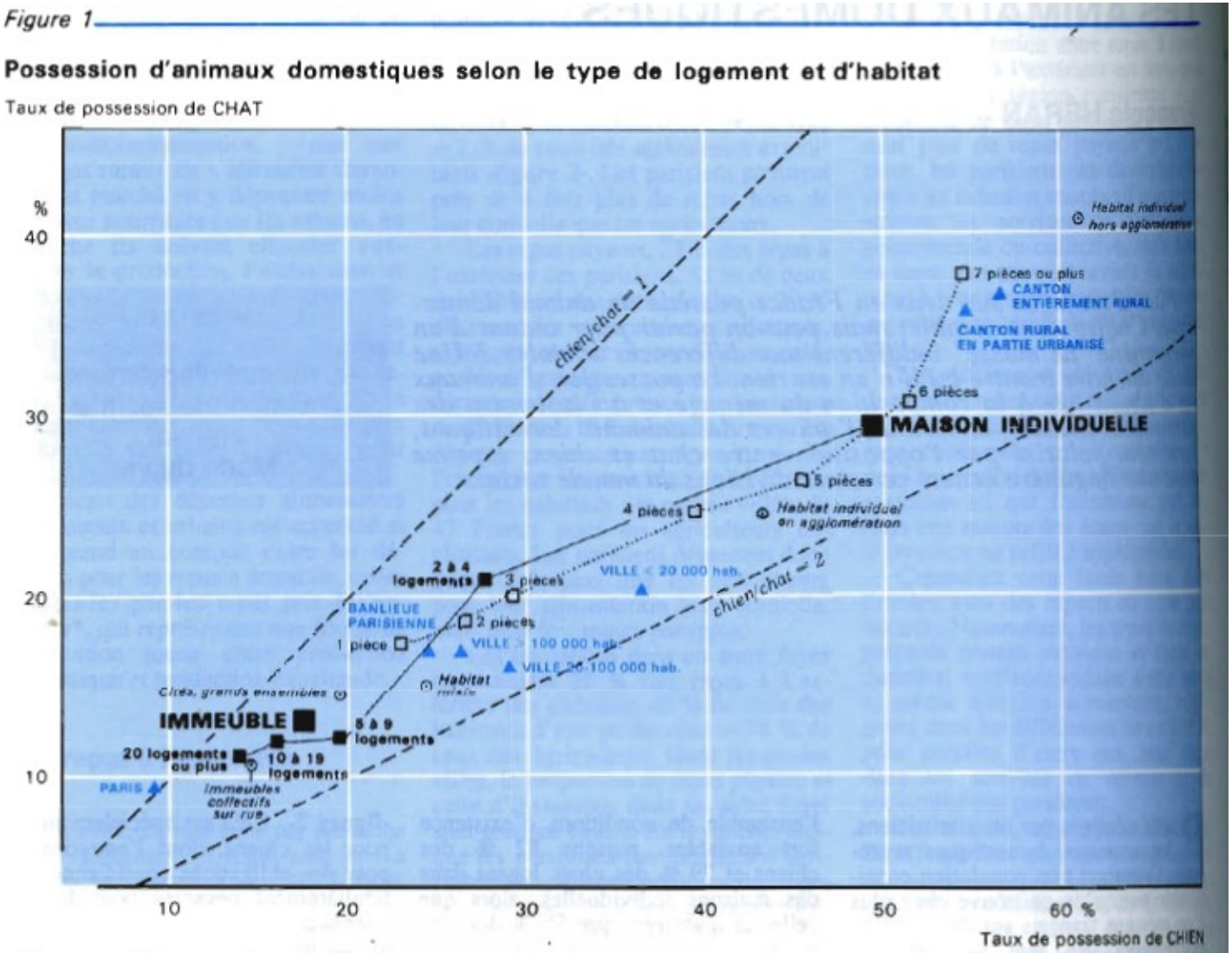
II. Liaison entre deux variables

La statistique c'est comparer

- Milieu et dispersion d'une seule variable n'apportent (presque) rien
- Il faut comparer
  - Ex. comparaison d'une variable (salaire) dans le temps.
    - Salaire * année
  - Liaison entre deux variables
- Trois types de liaison
  - Catégorielle * Catégorielle
    - Tableau croisé
  - Catégorielle * Quantitative
    - Différence de moyennes
  - Quantitative * Quantitative
    - Covariance et coefficient de corrélation

Figure 1

Possession d'animaux domestiques selon le type de logement et d'habitat



Source : INSEE, enquête Contacts entre les personnes, 1982-83.
Lire ainsi : En 1982, sur 100 ménages habitant en immeuble collectif, 13 possédaient un chat et 17 un chien.

- Quelles relations entre quelles variables ?
- 10 tableaux croisés en un graphique
- Type d'immeuble * Chien
- Type d'immeuble * Chat
- Nombre de pièces * Chien
- Nombre de pièces * Chat
- Etc: (Type d'habitat, type d'unité urbaine)* (Chien, Chat)

(Héran, 1987)

Davantage d'animaux domestiques chez les indépendants et les ménages moins diplômés

En %

GROUPES SOCIOPROFESSIONNELS	Proportion de ménages possédant		
	un animal domestique	un chat	un chien
Agriculteurs	77,4	48,7	62,5
Artisans, commerçants, patrons	60,6	20,3	47,2
Cadres et prof. intellectuelles sup.	46,7	20,9	26,4
Professions intermédiaires	48,4	19,1	30,0
Employés	40,5	17,0	23,0
Ouvriers qualifiés	55,7	22,5	38,6
Ouvriers non qualifiés	48,7	20,7	36,4
Jamais eu de profession	28,8	10,5	16,6
DIPLÔME DU CHEF			
Aucun diplôme	53,2	24,2	39,4
Certificat d'études	52,0	22,2	36,1
CAP ou assimilé	60,8	24,5	42,2
BEPC ou assimilé	44,8	18,9	26,6
Diplôme professionnel	49,2	17,5	31,1
Baccalauréat (non technique)	42,9	21,0	26,7
Diplôme supérieur au bac	43,5	20,1	24,3
Ensemble	**51,7**	**22,3**	**35,2**

Classes sociales comme chien et chats ?

- Combien de tableaux croisés ?
- Nature de l'argument
  - Différence sociale en terme de possession de chiens
  - Moindre en terme de possession de chats
  - La différence de classe moins dans la possession de chaque animal que dans le ratio taux de possession de chat / taux de possession de chien

(Héran, 1987)

Tableau croisé initial. Diplôme * Chien

Nombre de cdip_b	Étiquettes de colonnes		
Étiquettes de lignes	1. Oui	2. Non	Total général
1.Aucun Dip	669	1027	1696
2.CE	484	856	1340
3.CAP	435	596	1031
4.BEPC	151	416	567
5.Bac	84	231	315
6.BacPro	98	217	315
7.Supérieur	150	468	618
Total général	**2071**	**3811**	**5882**

Diplôme * Chien. Pourcentage en ligne

Nombre de cdip_b	Étiquettes de colonnes		
Étiquettes de lignes	1. Oui	2. Non	Total général
1.Aucun Dip	39.4%	60.6%	100.0%
2.CE	36.1%	63.9%	100.0%
3.CAP	42.2%	57.8%	100.0%
4.BEPC	26.6%	73.4%	100.0%
5.Bac	26.7%	73.3%	100.0%
6.BacPro	31.1%	68.9%	100.0%
7.Supérieur	24.3%	75.7%	100.0%
Total général	**35.2%**	**64.8%**	**100.0%**

Diplôme * Chien. Pourcentage en colonne

Nombre de cdip_b	Étiquettes de colonnes		
Étiquettes de lignes	1. Oui	2. Non	Total général
1.Aucun Dip	32.30%	26.95%	28.83%
2.CE	23.37%	22.46%	22.78%
3.CAP	21.00%	15.64%	17.53%
4.BEPC	7.29%	10.92%	9.64%
5.Bac	4.06%	6.06%	5.36%
6.BacPro	4.73%	5.69%	5.36%
7.Supérieur	7.24%	12.28%	10.51%
Total général	**100.00%**	**100.00%**	**100.00%**

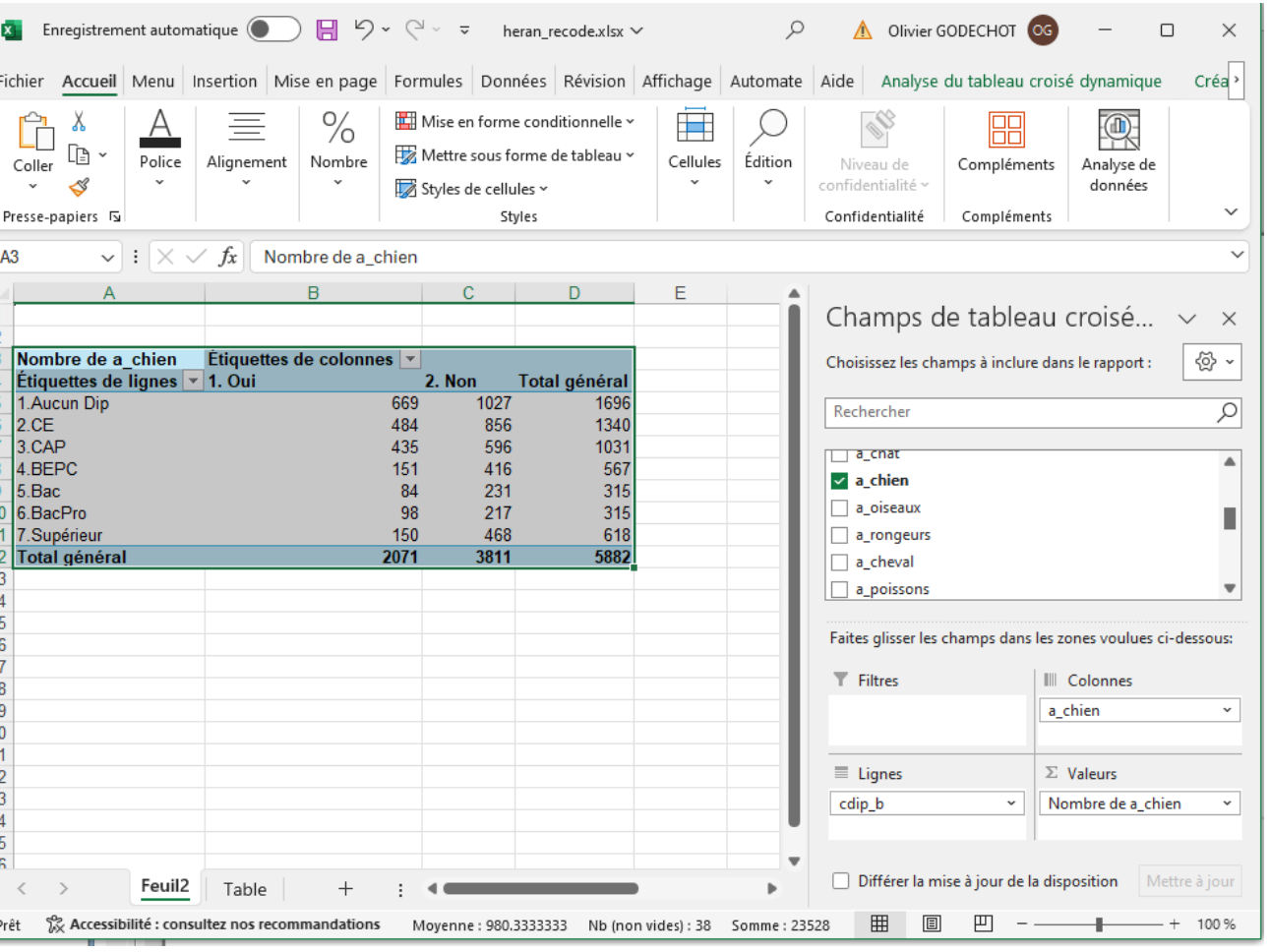
Diplôme * Chien. Pourcentage du total

Nombre de cdip_b	Étiquettes de colonnes		
Étiquettes de lignes	1. Oui	2. Non	Total général
1.Aucun Dip	11.37%	17.46%	28.83%
2.CE	8.23%	14.55%	22.78%
3.CAP	7.40%	10.13%	17.53%
4.BEPC	2.57%	7.07%	9.64%
5.Bac	1.43%	3.93%	5.36%
6.BacPro	1.67%	3.69%	5.36%
7.Supérieur	2.55%	7.96%	10.51%
Total général	**35.21%**	**64.79%**	**100.00%**

Liaison entre deux variables qualitatives

- Tableau croisé
- Les différents éléments d'un tableau croisé :
  - Effectifs
  - Pourcentage total
  - Pourcentage en ligne
  - Pourcentage en colonne
- Lecture :
  - On compare les pourcentages en ligne au sein d'un même colonne.
  - On compare les pourcentages en colonne au sein d'une même ligne
  - Avoir l'œil sur l'effectif.

Excel La manière de faire

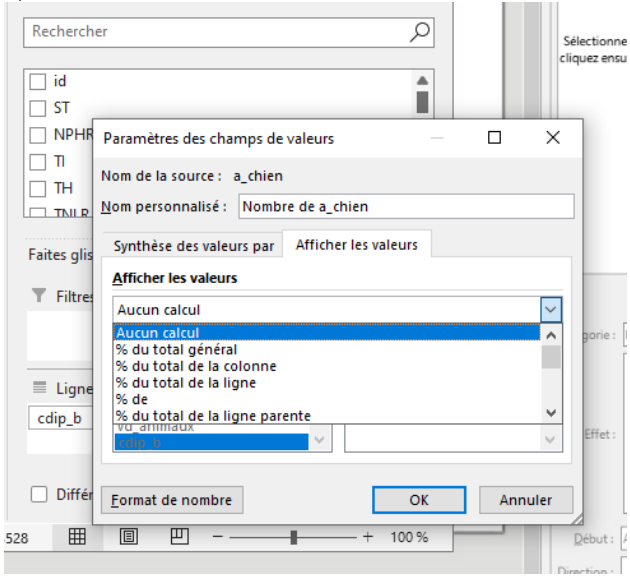


- Choisir la variable en ligne dans le champ **Lignes**
- Choisir la variable en colonne dans le champ **Colonnes**
- Mettre l'une des deux dans **Valeurs** (peu importe)
  - Sélectionner **Nombre**

Excel. Calculer les pourcentages

- Avec les formules
  - Préalable copier (Ctrl+C / Command +C) le tableau et collage spécial en « valeur » (Ctrl+Maj + V / Command + Maj +V). Par ex. dans une nouvelle feuille
    - Pctage en ligne : =B3/$D3
    - Pctage en colonne : =B3/B$10
    - Pctage du total : =B3/D10

- Avec le menu tableau croisé dynamique
  - Cliquer sur afficher les valeurs
  - Choisir le % du total



Présentation des tableaux croisés

- Conventions fréquentes
  - Mettre les variables « explicatives » en ligne. Les variables expliquées en colonne
  - Privilégier les pourcentages en ligne
  - Mettre les effectifs pour reconstituer le tableau complet
  - Ajouter une note de lecture
  - Faire du tableau/ du graphique un objet détachable.

Exemple. Suivre son chef s'il part dans une autre entreprise.

	1. Oui	2. A voir	3. Non	Total	Effectif
1. Front office BFI	31%	50%	19%	100%	26
2. Ingénierie, informatique BFI	18%	82%	0%	100%	11
3. Support BFI	29%	64%	7%	100%	14
4. Front office Autre	11%	69%	20%	100%	45
5. Ingénierie, informatique Autre	23%	68%	9%	100%	22
6. Support Autre	11%	71%	18%	100%	89
Ensemble	16%	68%	16%	100%	207

Liaison entre une variable catégorielle et une variable quantitative

- Solution 1.
  - Transformer la variable quantitative en variable catégorielle (en particulier si la variable quantitative est de type explicative (âge, revenu)) et faire un tableau croisé.
- Solution 2.
  - Comparer les moyennes (médianes, quartiles) par groupe de la variable catégorielle.
    - En particulier si variable catégorielle est explicative (genre, origine sociale) et la variable quantitative est expliquée (poids, taille, salaire)

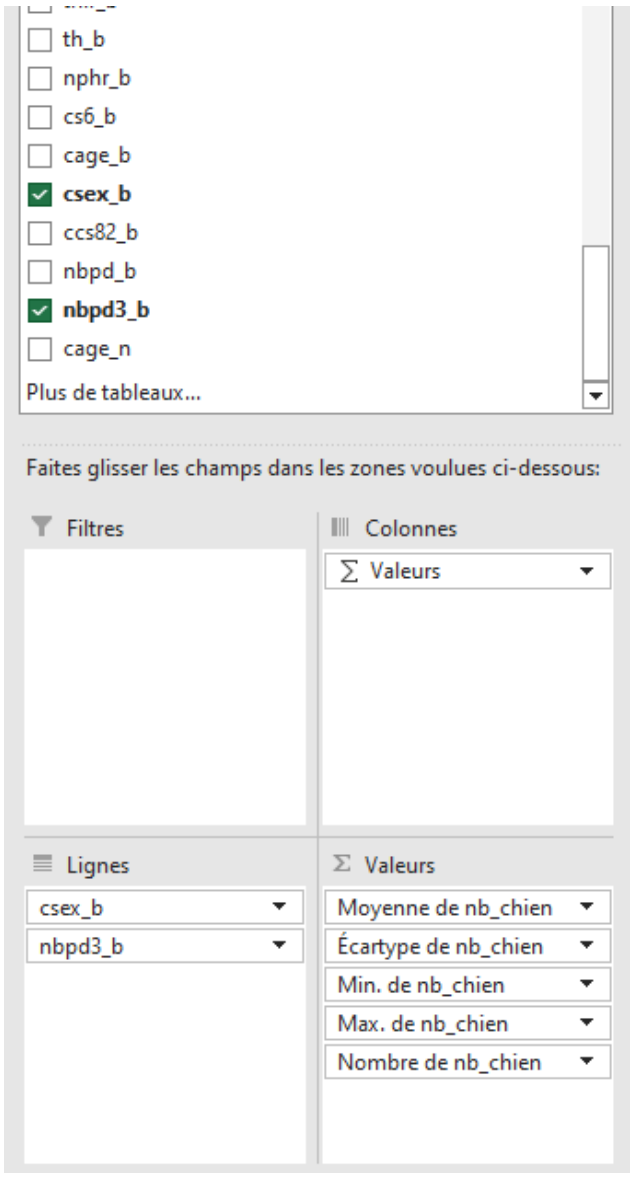
Example : nombre moyen de chiens

- Comparaison « **chef de ménage »** homme-femme
- Et comparaison à taille de ménage identique

Chef de ménage	Taille du ménage	Nombre moyen de **CHIENS** par ménage	Nombre moyen de **CHIENS** par personne	Nombre moyen de **CHATS** par ménage	Nombre moyen de **CHATS** par personne
HOMME	1	0,19	0,19	0,19	0,19
	2	0,47	0,23	0,34	0,17
	3	0,59	0,20	0,40	0,13
	4	0,58	0,14	0,37	0,09
	5	0,78	0,16	0,56	0,11
	6 ou plus	1,11	0,16	0,69	0,11
	Ensemble	**0,56**	**0,18**	**0,39**	**0,12**
FEMME	1	0,15	0,15	0,18	0,18
	2	0,42	0,21	0,35	0,17
	3 ou plus	0,65	0,18	0,46	0,13
	Ensemble	**0,27**	**0,17**	**0,25**	**0,16**
TOUS MÉNAGES		**0,50**	**0,18**	**0,36**	**0,13**

(Héran, 1987)

Variables. Dé



Faire une comparaison sous Excel

Étiquettes de lignes	Moyenne	Écartype	Min.	Max.	Nombre
femme	**0.27**	**0.63**	**0**	**6**	**1303**
1 personne /ménage	0.15	0.42	0	4	878
2 personnes /ménage	0.42	0.80	0	6	239
3 personnes ou + / ménage	0.65	0.94	0	5	186
homme	**0.56**	**0.88**	**0**	**11**	**4579**
1 personne /ménage	0.19	0.52	0	6	383
2 personnes /ménage	0.47	0.76	0	7	1469
3 personnes ou + / ménage	0.67	0.96	0	11	2727
Total général	**0.50**	**0.84**	**0**	**11**	**5882**

- Ici comparaison complexe, à double entrée
- Ne pas oublier l'écart-type, le min, le max et le nombre (« **CHIFFRES** »)

Liaison entre deux variables quantitatives: covariance et coefficient de corrélation

- Un écart à la moyenne sur une variable est-il accompagné d'un écart à la moyenne sur une autre variable ?
- Covariance : *Moyenne du produit des écart à la moyenne de deux variables*

$$COV(X,Y) = \frac{\sum_{i=1}^{i=n} (X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

- Coefficient de corrélation linéaire.

$$r_{X,Y} = \frac{\sum_{i=1}^{i=n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{i=n} (X_i - \overline{X})^2 \sum_{i=1}^{i=n} (Y_i - \overline{Y})^2}}$$

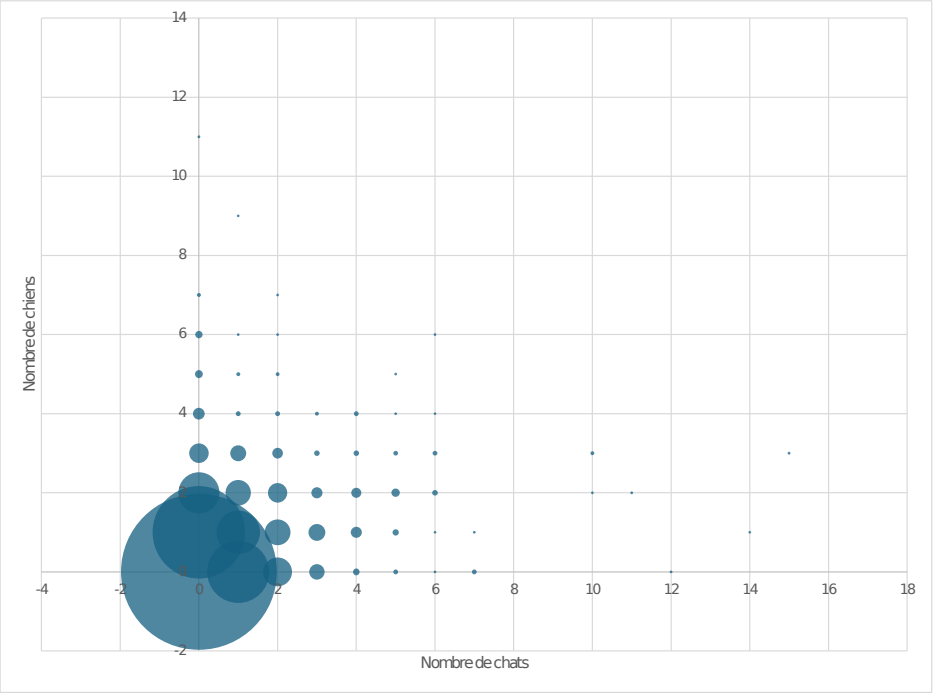
  - Varie entre -1 et 1

$$r_{X,Y} = \frac{COV(X,Y)}{ET(X).ET(Y)}$$

Utilisation

- Covariance
  - Dépend de l'échelle de x et de y
  - Mais le signe informatif
    - Positif
    - Négatif
- Coefficient de corrélation linéaire
  - Interprétation : Effet de variation d'un écart type de x sur la variation de y en proportion de son écart-type
  - Indique l'intensité d'une liaison (valeur absolue)
  - Et la nature du lien : positif et négatif

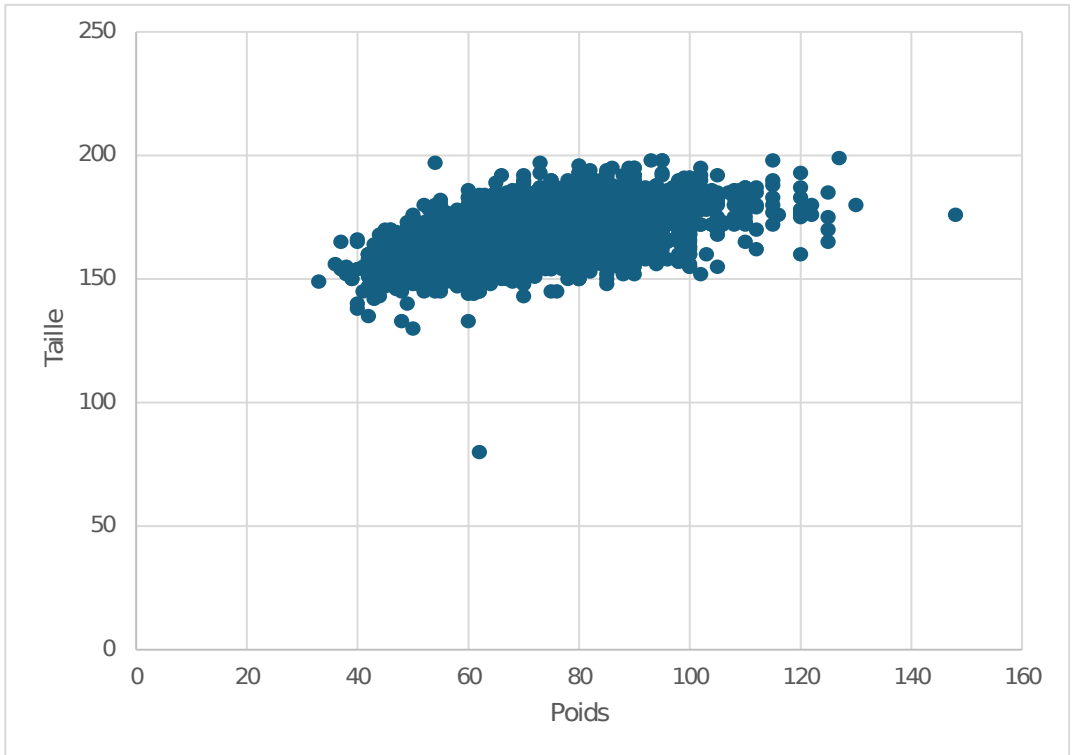
Exemple



Lien nombre de chiens et nombre de chats	Tous les ménages	Ceux qui ont un chat OU un chien
Covariance	0.186	-0.048
Coefficient corrélation	0.245	-0.042

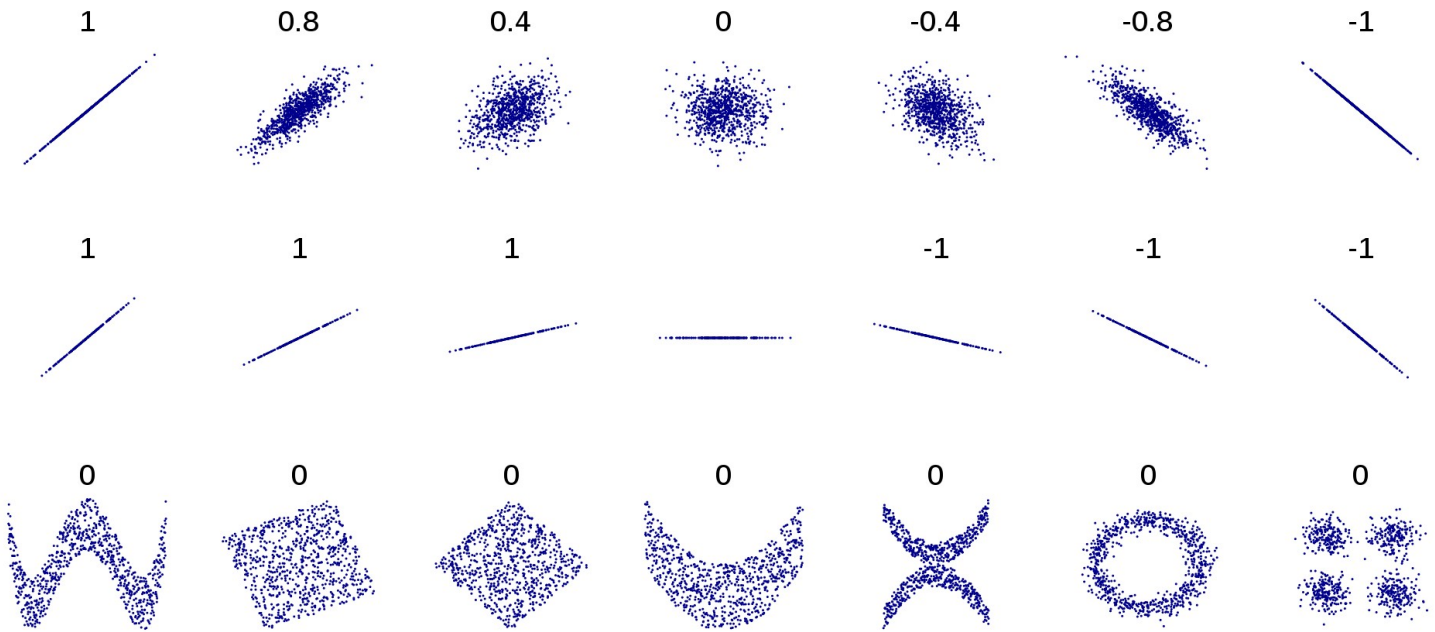
- Importance du point 0,0
  - 3000 ménages

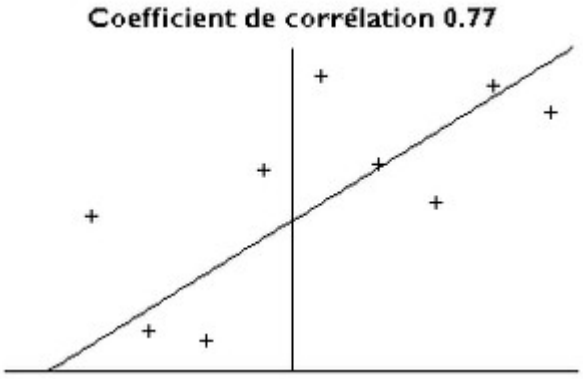
Poids et taille

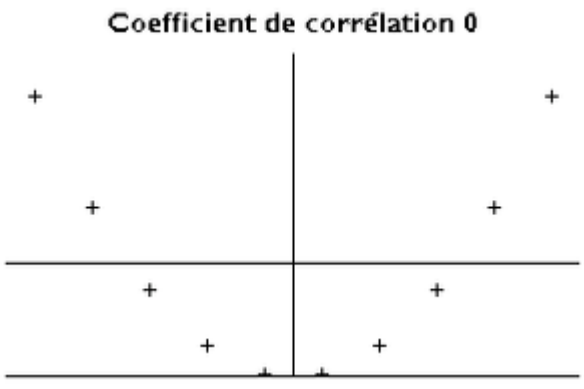


Poids et taille	
Covariance	71.934
Coefficient de corrélation	0.556

Nuage de points et coefficients de corrélation. Exemples







Sources : Wikipedia

Covariance et coefficient de corrélation sous Excel

- Utilisez formules

  Covariance : « =COVARIANCE(A1:A5883;B1:B5883) »

  Coefficient de corrélation :
  « =COEFFICIENT.CORRELATION(A1:A5883;B1:B5883) »

  - Plage 1, plage 2

Quelques conseils pour un nuage de points sous Excel

- Sans valeurs manquantes :
  - Sélectionner la plage des deux colonnes.
  - Insertion / Graphique / Nuage de points
  - La première variable est en abscisse (axe horizontal), la deuxième en ordonnée (axe vertical)
- S'il y a des valeurs manquantes
  - Mettre les deux colonnes dans une nouvelle feuille
  - Trier et supprimer les valeurs manquantes
  - Faire le graphique avec la méthode décrite au-dessus en sélectionnant uniquement les valeurs non vides.

Références

Dainville-Barbiche, Ségolène de. 1989. “Le clergé paroissial de Paris à la fin de l’Ancien Régime (1789-1791).” *Bibliothèque de l’École des chartes*, 539–61.

Héran, François. 1987. “Les animaux domestiques.” *Données sociales*, 417–23.

Quillian, Lincoln, and Devah Pager. 2001. “Black Neighbors, Higher Crime? The Role of Racial Stereotypes in Evaluations of Neighborhood Crime.” *American Journal of Sociology* 107 (3): 717–67. https://doi.org/10.1086/338938.

Appendices. Avec le logiciel R

Tableau d'effectifs avec R

- Syntaxe de base

  table(mabase$mavar)

```
   1    2    3    4    5    6    7
1696 1340 1031  567  315  315  618
```

- Avec les non-réponses

  table(mabase$mavar, useNA= "ifany")

- Avec les non-réponses et l'effectif total

  addmargins(table(mabase$mavar, useNA= "ifany"))

- Organiser les catégories en lignes

  data.frame(addmargins(table(mabase$mavar, useNA= "ifany")))

Tableaux croisés sous R

- Tableau croisé : table

```
a<-table(fic$Sexe,fic$ChgEntr)
a
```

```
          non 1 fois 2à3fois >4fois
1. Homme 137    133     151     55
2. Femme  67     50      38     20
```

- Avantage : en sortie une matrice, permet les calculs matriciels
- Inconvénient : Pas très pratique. On a uniquement les effectifs, sans les marges.

La variance et les écart-types sous R

```
var(mabase$mavar,na.rm=TRUE)
    [1] 5715.055
sd(mabase$mavar,na.rm=TRUE)
    [1] 75.59798
```

Tableaux croisés sous R. Améliorer les sorties

- addmargins(a) ajoute les marges ; addmargins(a,1) ajoute la ligne de total. addmargins(a,2) ajoute la colonne de total.

addmargins(a)

	non	1 fois	2à3fois	>4fois	Sum
1. Homme	137	133	151	55	476
2. Femme	67	50	38	20	175
Sum	204	183	189	75	651

- prop.table(a) ajoute les pourcentages du total; prop.table(a,1) les pourcentages en ligne et prop.table(a,2) les pourcentages en colonne.

prop.table(a,1) % en ligne

	non	1 fois	2à3fois	>4fois
1. Homme	0.2878151	0.2794118	0.3172269	0.1155462
2. Femme	0.3828571	0.2857143	0.2171429	0.1142857

Tableaux croisés sous R (suite)

- Comment combiner tout ça ?

```
addmargins(prop.table(addmargins(a,1),1),2)
```

```
          non    1 fois   2à3fois    >4fois       Sum
1. Homme 0.2878151 0.2794118 0.3172269 0.1155462 1.0000000
2. Femme 0.3828571 0.2857143 0.2171429 0.1142857 1.0000000
Sum      0.3133641 0.2811060 0.2903226 0.1152074 1.0000000
```

- Et le rajout des effectifs totaux ?

```
cbind(addmargins(prop.table(addmargins(a,1),1),2), rowSums(addmargins(a,1)
```

```
                non    1 fois   2à3fois    >4fois Sum
1. Homme 0.2878151 0.2794118 0.3172269 0.1155462   1 476
2. Femme 0.3828571 0.2857143 0.2171429 0.1142857   1 175
Sum      0.3133641 0.2811060 0.2903226 0.1152074   1 651
```

L'indice de liaison entre une variable qualitative et une variable quantitative sous R

- Différence de moyennes par groupe

```
tapply(mabase$mavarnum, mabase$mavarqual,
   mean,na.rm=TRUE)
```

```
      1. Homme       2. Femme 9. Non réponse
      87.73854       72.10076       60.50000
```

Covariance et corrélation avec R

```
cov(na.omit(cbind(mabase$mavar1, mabase$mavar2, mabase$mavar3
    … )))
```

```
             [,1]     [,2]
[1,] 5715.055 4714.221
[2,] 4714.221 4361.620
```

na.rm=TRUE ne semble pas fonctionner sour R 2.8

```
cor(na.omit(cbind(mabase$mavar1, mabase$mavar2, mabase$mavar3
    … )))
```

```
             [,1]      [,2]
[1,] 1.0000000 0.9442256
[2,] 0.9442256 1.0000000
```