Traitement statistique des données d'enquête avec introduction à SPSS

Gilbert Ritschard

Département d'économétrie, Université de Genève

gilbert.ritschard@themes.unige.ch

Bamako, 7-11 octobre 2002

Traitement statistique des données d'enquête avec introduction à SPSS

Gilbert Ritschard
Département d'économétrie, Université de Genève

Plan

http://mephisto.unige.ch

1 Objectifs

1. Savoir gérer (importer, exporter, recoder, transformer, filtrer) les données avec SPSS.
2. Analyse statistique descriptive (notion de distribution, graphiques et indicateurs statistiques)
   (a) analyse univariée
   (b) analyse bivariée
3. Quelques principes de statistique inférentielle
   (a) Estimation ponctuelle : biais et variance
   (b) Intervalle de confiance et marge d'erreur
   (c) Principe du test statistique d'hypothèse

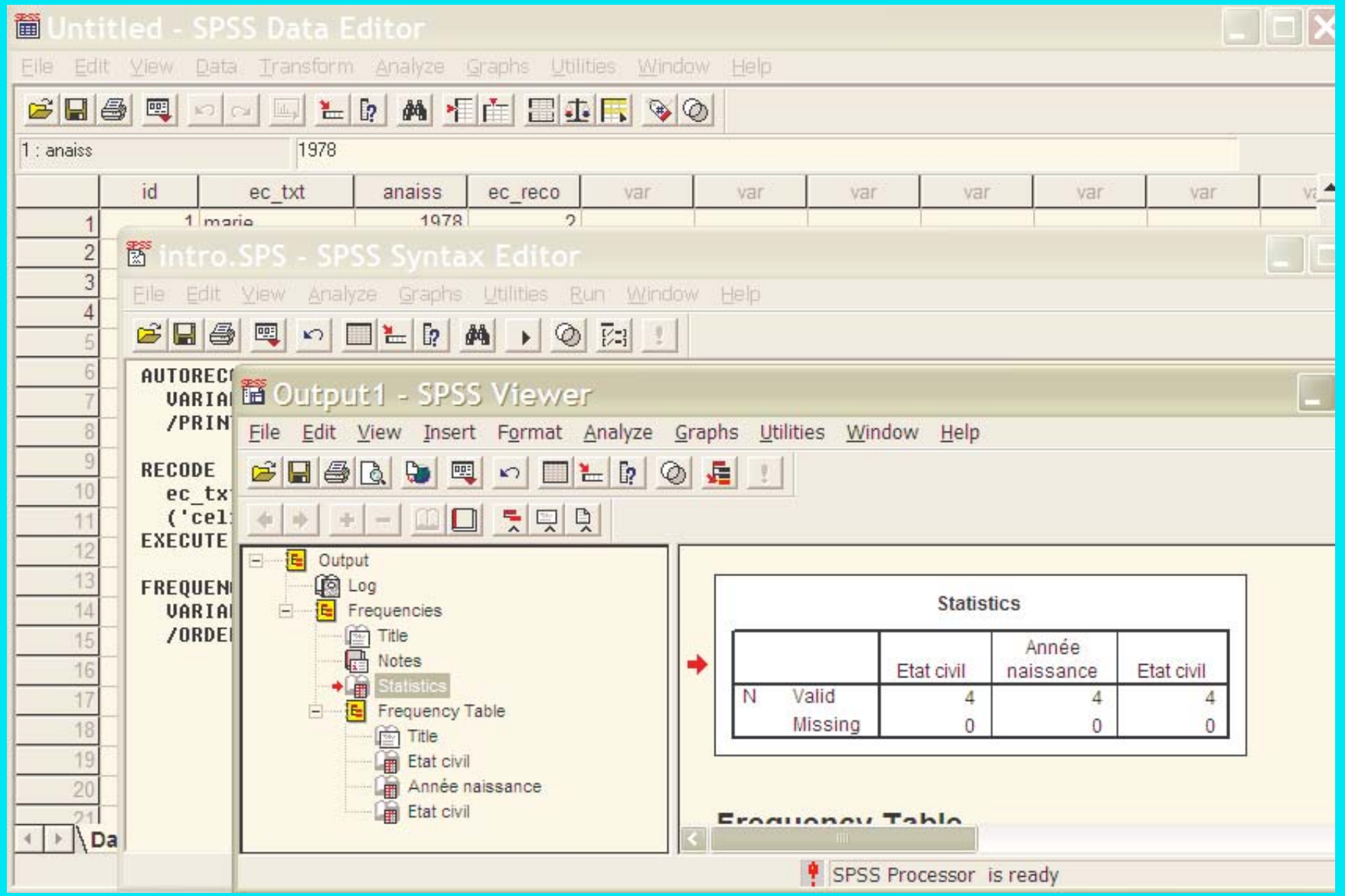
2 Introduction à SPSS

SPSS (Statistical Package for the Social Sciences)

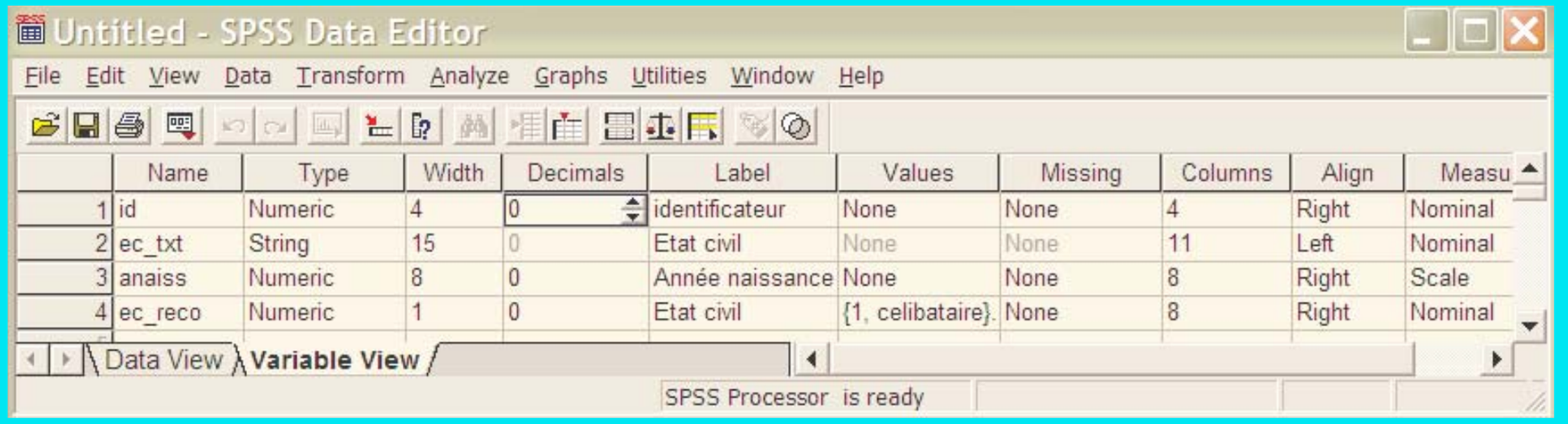
Logiciel commercial pour le traitement et l'analyse statistique de données.

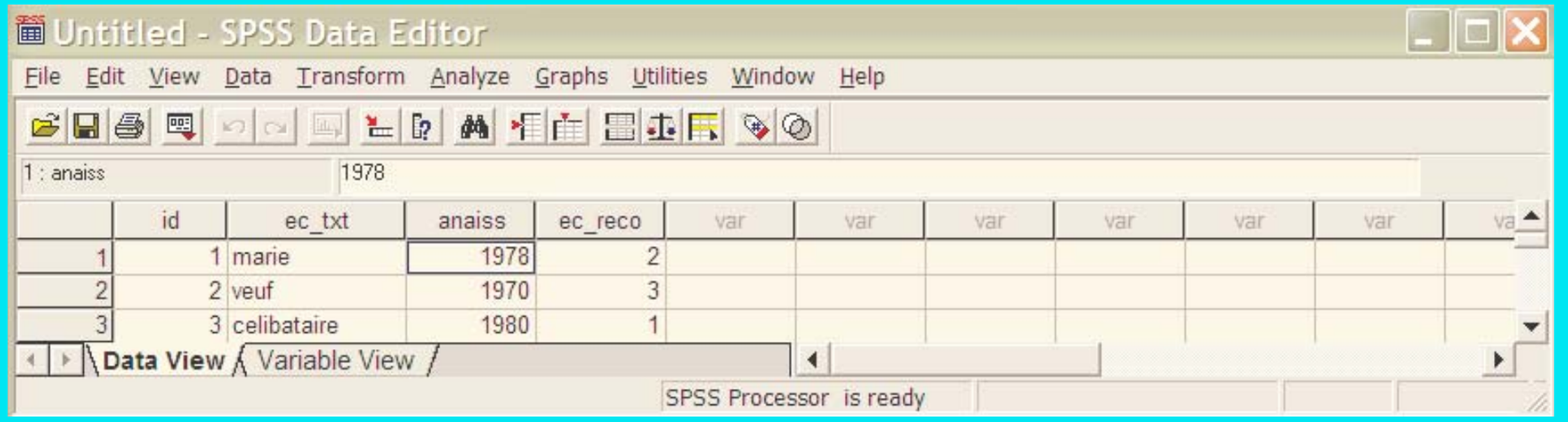
Distribué par SPSS Inc. (http://www.spss.com) sous forme d'un module de base et de plusieurs modules spécialisés (advanced, categories, trend, ...)

Les trois fenêtres



L'éditeur de données





Utilisation

Les opérations

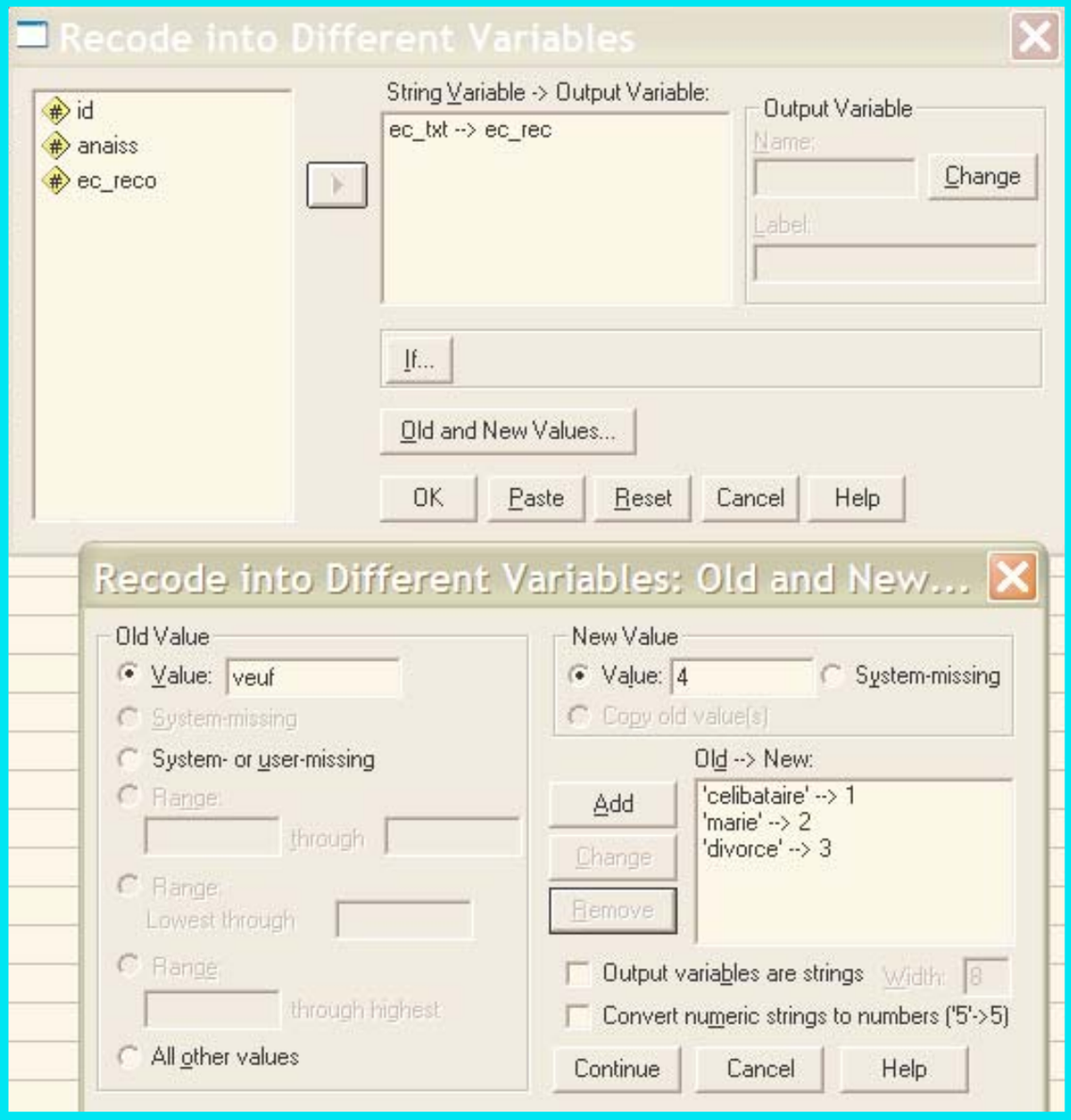
- saisie ou lecture des données
- transformation et construction de variables
- analyse statistique

peuvent se faire

- par le menu et les dialogues appropriés
  avantage : intuitif, rapidement opérationnel
- par la syntaxe
  avantage : peut être mémorisée et donc réutilisée et transférée aux collègues.
  Certaines options ne sont accessibles que par la syntaxe.

Suggestion : utiliser le menu la première fois et sauver la syntaxe en cliquant sur Paste

Exemple de dialogues : recodage d'une variable



$\Rightarrow$ Paste génère la syntaxe :

```
RECODE ec_txt
 ('celibataire'=1)
 ('marie'=2)
 ('divorce'=3)
 ('veuf'=4)
 INTO  ec_rec .
EXECUTE .
```

Éléments de syntaxe

Commande : commence avec le nom de commande (GET, COMPUTE, FREQUENCY, ...), suivie des arguments et se termine par un point “.” .

Sous-commandes : précédées d’un “/” (peut être omis s’il suit directement le nom de commande) et séparées des éventuels arguments par “=”.

Séparateurs entre arguments : espace ou virgule.

Commentaire : entre /* et */ ou ligne commençant avec * → “.” .

Exemples :

```
GET FILE= 'exemple.sav'.        /* lecture du fichier exemple.sav */
* ceci est un commentaire.
FREQUENCY
  VARIABLES = ec_rec anaiss     /* distribution empirique */
  /ORDER=  ANALYSIS             /* des variables ec_rec et anaiss */
  /BARCHART .
```

Cas et variables

Dans SPSS : données organisées sous forme de tableau

- Lignes : Cas
- Colonnes : Variables

Noms de variables (entêtes de colonnes) : au plus 8 caractères (pas d'espace) dont le premier doit être une lettre (ou @,# ou $).

Les lignes sont numérotées. Il peut être utile de définir une variable (colonne) "identificateur" prenant une valeur différente pour chaque cas.

Remarque : SPSS permet de construire facilement des variables par combinaison de colonnes (combinaison d'éléments d'une même ligne).

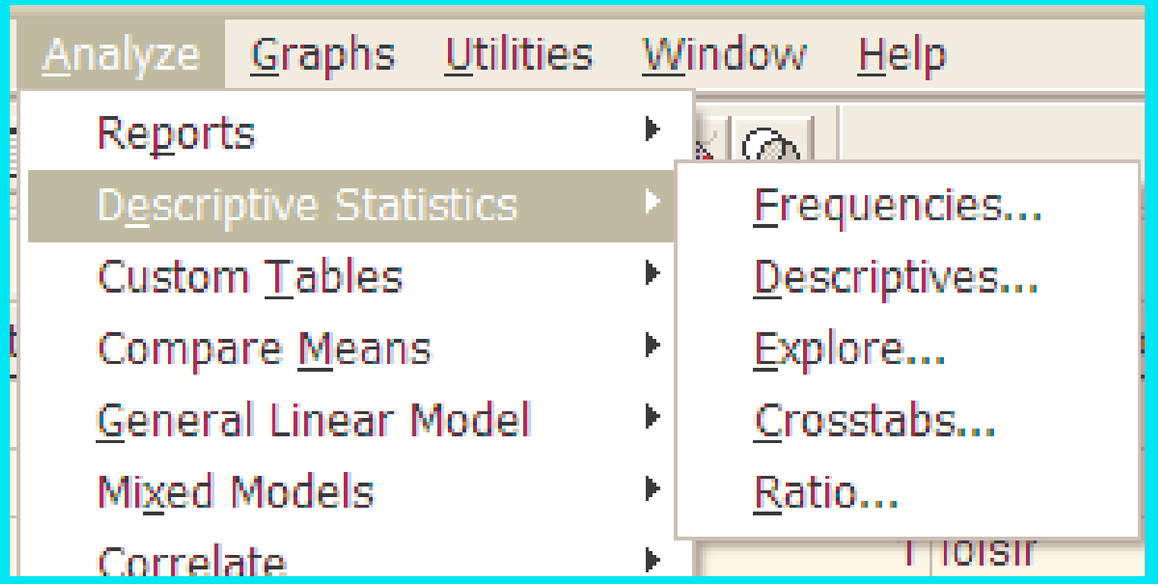
Les transformations nécessitant des fonctions d'éléments d'une même colonne (retrancher la moyenne par exemple) requièrent des opérations avancées avec MATRIX que nous ne traitons pas ici.

2.1 Trois commandes statistiques fondamentales

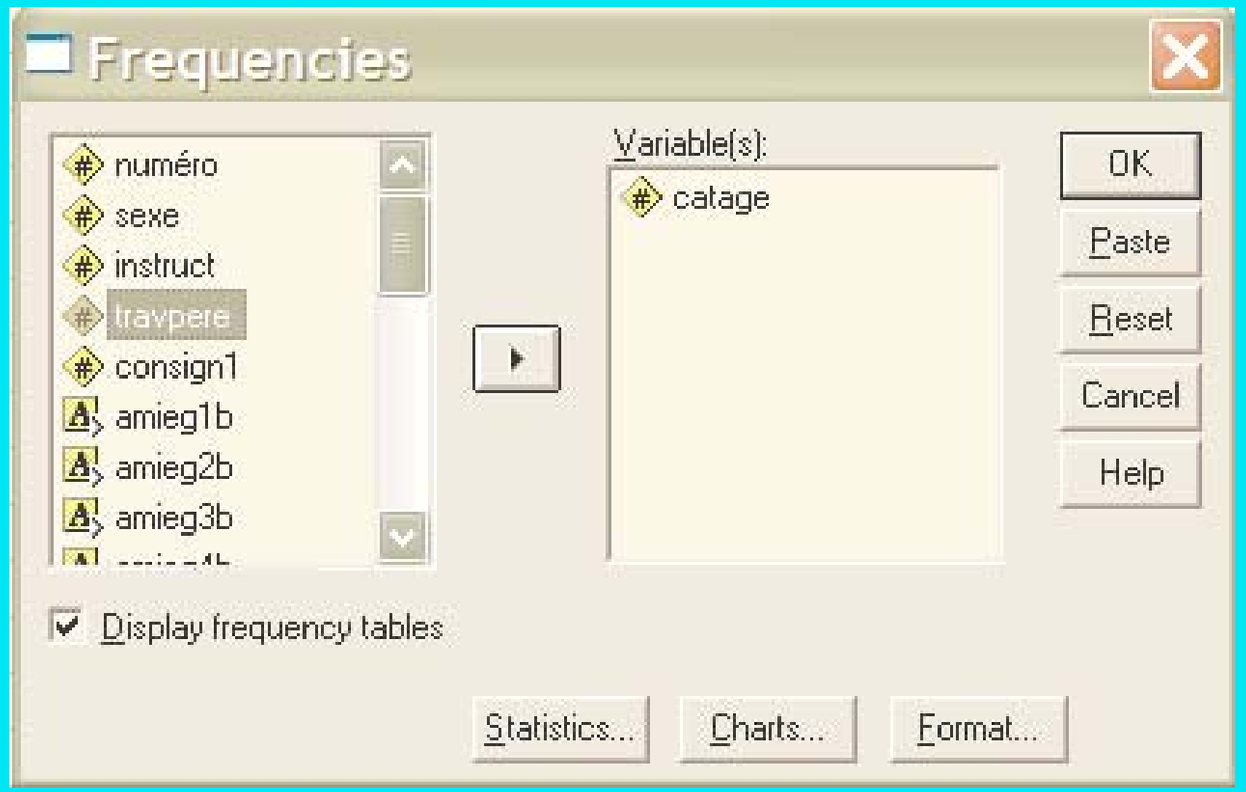
Commandes fondamentales pour l'exploration initiale des données : FREQUENCIES, DESCRIPTIVES, GRAPHS

FREQUENCIES : tableau des fréquences de chaque valeur (+ graphique en barres)

DESCRIPTIVEs : nombre valeurs valides, minimum, maximum, moyenne, écart type, ...



Dialogue FREQUENCIES et syntaxe



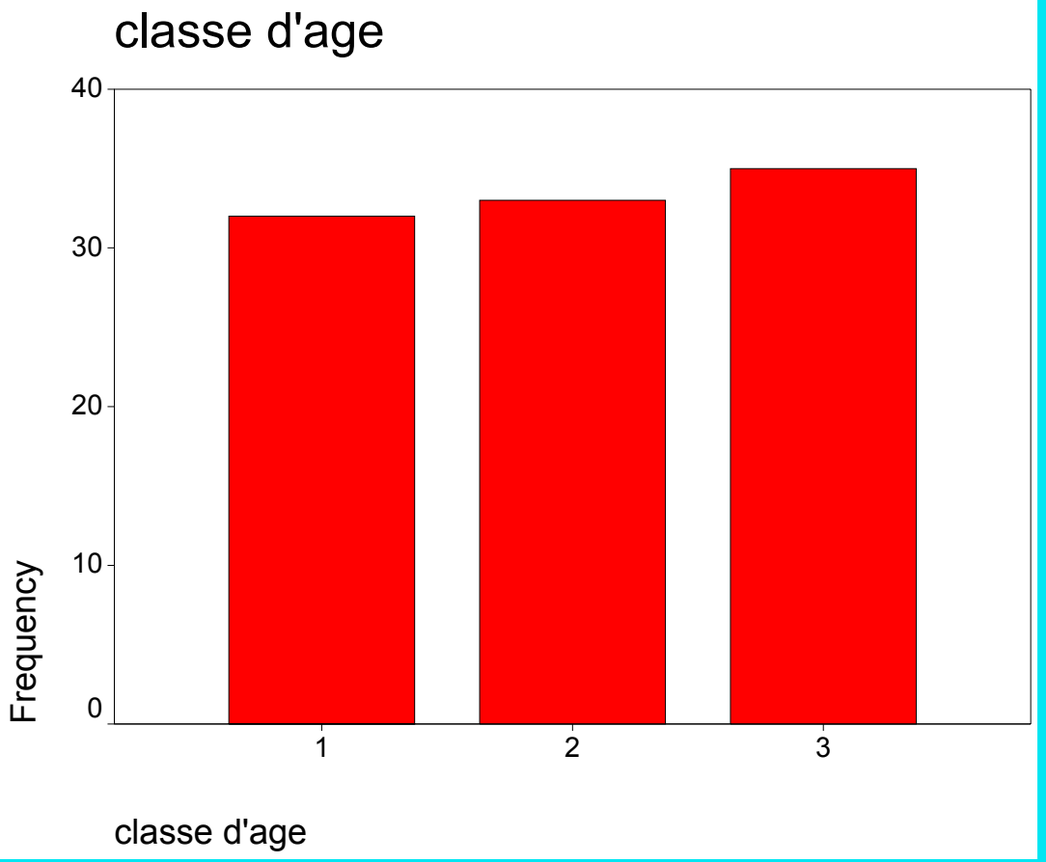

```
FREQUENCIES
   VARIABLES=catage
   /BARCHART  FREQ
   /ORDER=  ANALYSIS .
```

FREQUENCIES

exemple

classe d'age

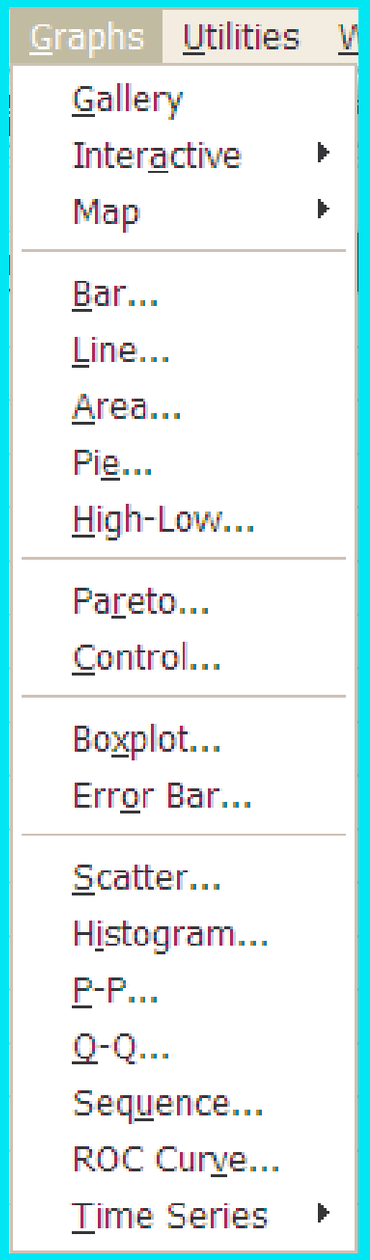
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	32	32.0	32.0	32.0
	2	33	33.0	33.0	65.0
	3	35	35.0	35.0	100.0
	Total	100	100.0	100.0	



DESCRIPTIVES : exemple

Descriptive Statistics

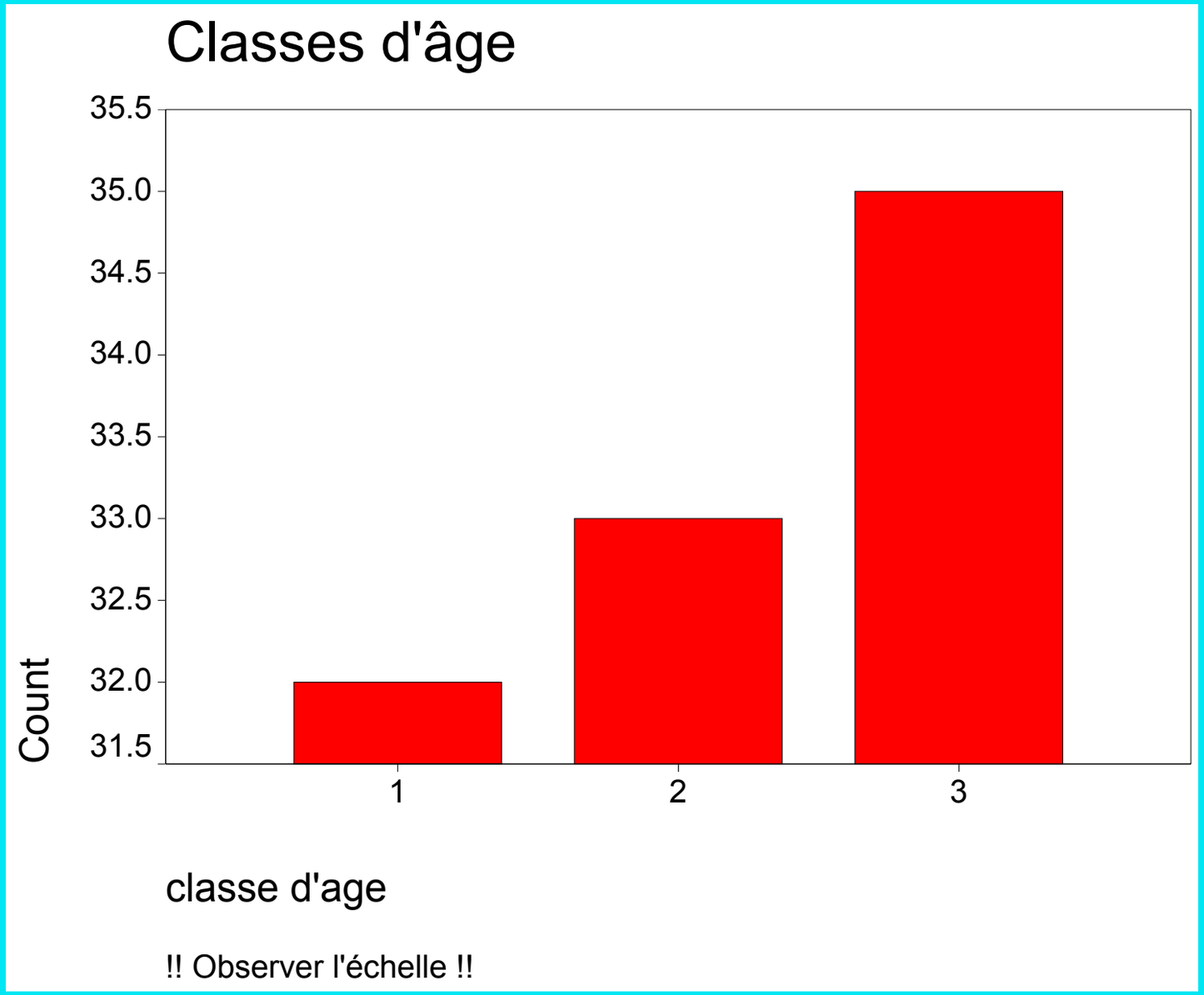
	N	Minimum	Maximum	Mean	Std. Deviation
NUMÉRO	100	102	1101	160.50	99.205
SEXE	100	1	3	1.58	.516
classe d'age	100	1	3	2.03	.822
INSTRUCT	100	1	2	1.38	.488
profession du père	100	1	4	3.02	.864
NUMERO	100	101	200	150.50	29.011
SEX	100	1	2	1.56	.499
classe d'age	100	1	3	2.01	.823
ETUDE	100	1	2	1.38	.488
profession du père	100	1	4	3.05	.869
Valid N (listwise)	100				



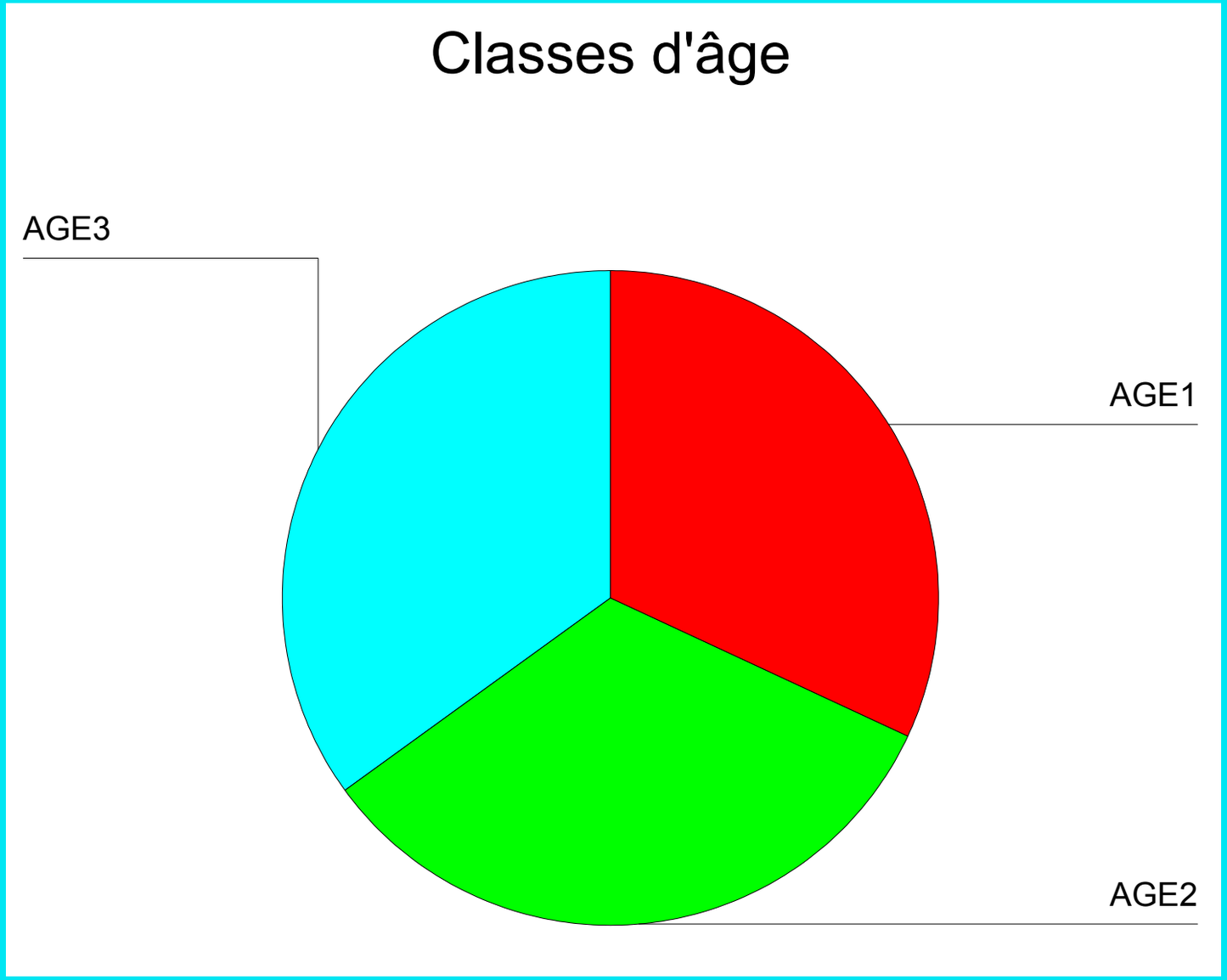
GRAPHS : divers graphiques :

- Données catégorielles : barres (bars), circulaire (pie)
- Données ordinales : lines, surfaces (area), boxplot
- Données quantitatives : histogrammes (histogram), dispersion (scatterplot)

GRAPHS/BAR : exemple



GRAPHS/PIE : exemple



3 Gestion des données avec SPSS

1. Définition des variables et entrées des données
2. Pré-traitement des données
   (a) Codage de données
   (b) Tests de cohérence de données
   (c) Filtre et sélection de variables
   (d) Données manquantes
3. Exportation et importation de fichiers
4. Agrégation et fusion de fichiers

3.1 Définition des variables et entrées des données

Échelles de mesure des variables

- nominale
  - *dichotomique, binaire* : Homme/Femme, Oui/Non, ...
  - *polytomique* : Activité, Avec qui ?, Où ?, ...
- ordinale

  Souhaitiez-vous cette grossesse : a) à ce moment, b) plus tard, c) non
- quantitative (métrique) de type intervalle

  Température, A quelle heure de la journée ?, ...
- quantitative (métrique) de type ratio

  Âge, Depuis combien de temps ?, ...

SPSS distingue : nominal, ordinal, scale (=métrique)

Définition d'une variable : (éditeur de données, page 6)

Nom (Name) obligatoire, 8 caractères au maximum

Type : numérique (par défaut), date, monétaire, texte

Width : nombre maximal de caractères des valeurs (8 par défaut)

Decimals : nombre de décimales (par défaut 2 ou 0 selon type)

Label : description longue de la variable

Values : description des valeurs (vivement conseillé pour variables nominales)

Columns : Largeur affichée de la colonne (8 par défaut)

Align (Alignement) : Left (à gauche), center (centré), right (à droite)

Measure : Nominal, ordinal, scale (par défaut)

3.2 Codage de données

3.2.1 Codage de données nominales

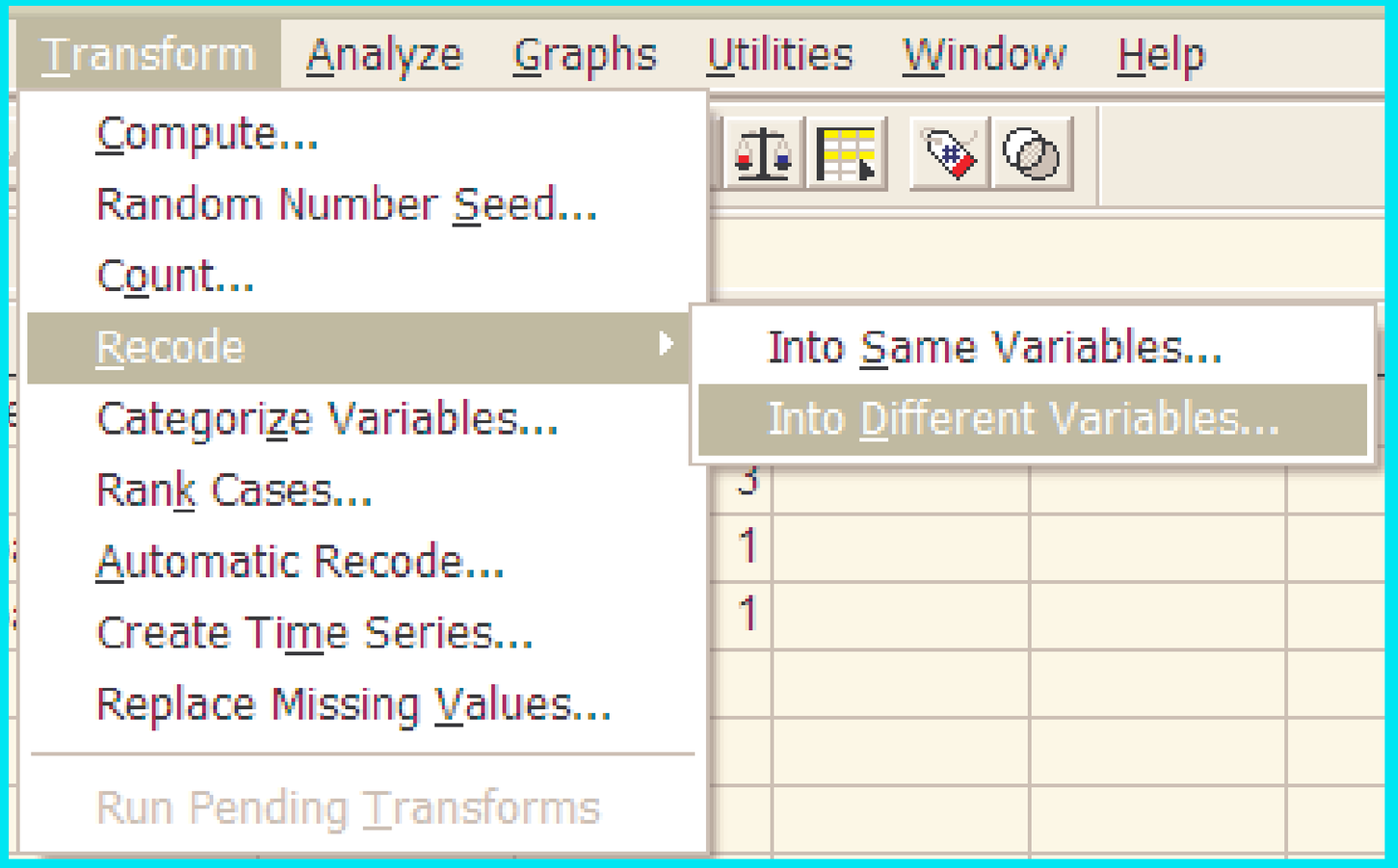
Présentation de données nominales :

État civil

texte (string)	code	étiquette (label)
célibataire	1	célibataire
marié	2	marié
divorcé	3	divorcé
veuf	4	veuf
	9	NR

Attention : variable codée numériquement $\not\Rightarrow$ quantitative.

Les logiciels comme SPSS requièrent souvent des données codées numériquement.



Voir Dialogue «recode» transparent 8

Autorecode : recodage automatique (selon ordre alphabétique)
Met les anciennes valeurs en étiquettes (value labels).

Syntaxe de recodages

```
GET /FILE= 'exemple.sav'.

AUTORECODE
  VARIABLES=ec_txt  /INTO ec_reco
  /PRINT.

RECODE
  ec_txt
  ('celibataire'=1)  ('marie'=2)  ('divorce'=3)  ('veuf'=4)  INTO  ec_rec .
VALUE LABEL ec_rec 1 'celibataire' 2 'marie' 3 'divorce' 4 'veuf' .
VARIABLE LABEL ec_rec 'état-civil (ec_rec)'.

FREQUENCIES
  VARIABLES= anaiss ec_txt ec_reco ec_rec
  /ORDER=  ANALYSIS.
```

3.2.2 Transformation et création de variables par calcul

COMPUTE *nllevar = expression*
IF (*condition*) *nllevar = expression*

expression :

- opération arithmétique entre variables $\{+, -, *, /, **\}$
  exemple : `compute duree = fin - debut.`
- fonction de variables (ABS(),RND(),SUM(),MEAN(),MAX(),MED(),...)
  exemple : `compute max_dur = max(duree1,duree2,duree3).`

condition : expression logique (`var1 OP var2`) = $\begin{cases} \texttt{true (vrai)} \\ \texttt{false (faux)} \end{cases}$

relations (`true` si vérifiée, `false` sinon)

EQ ou =	égal à
NE ou ~= ou <>	pas égal à
LT ou <	plus petit que
LE ou <=	plus petit ou égal à
GT ou >	plus grand que
GE ou >=	plus grand ou égal à

négation : NOT ou ~

var = 0 $\Rightarrow$ if var $\Leftrightarrow$ if false $\Rightarrow$ `var > 1` $\Rightarrow$ `NOT(var) = false`

Exemples :

```
if (agefin < agedeb) erreur = 1 .
if  missing(annais) age = year - annais.
```

Opérateur logique : AND et OR

AND ou & `(cond1 & cond2)` true si cond1 et cond2 sont vrais

OR ou | `(cond1 | cond2)` true si cond1 ou cond2 est vrai

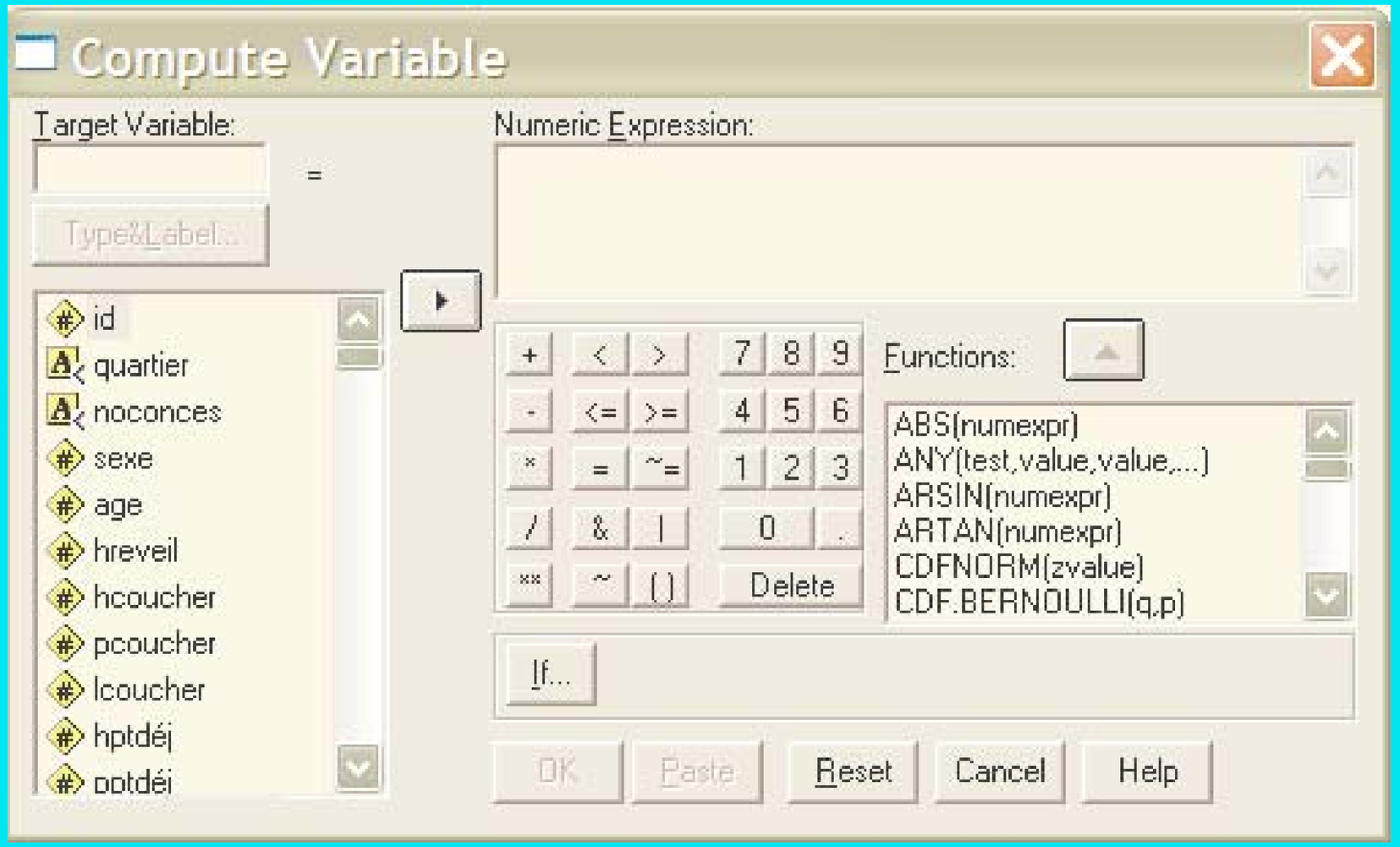
AND	true	false	missing
true	true	false	missing
false	false	false	false

OR	true	false	missing
true	true	true	true
false	true	false	missing

Exemple :

```
compute filter = 0.
if ((age>22) and (sexe=1)) filter = 1.
```

Menu : Transform/Compute...



3.3 Tests de cohérence de données

Il s'agit de repérer

- Les valeurs interdites de variables
  - variable discrète : si valeur $\notin$ liste des valeurs autorisées
    exemple : sexe (1=homme,2=femme) $\Rightarrow$ repérer valeurs de sexe autre que 1 ou 2.
  - variable continue : si valeur $<$ minimum ou $>$ maximum
    exemple : age ($[15;30]$) $\Rightarrow$ repérer cas avec age<15 ou age>30.
- Les valeurs d'une variable incompatibles avec valeur prise par une autre
  exemple : état matrimonial = célibataire et âge du mari = 25.

```
*** charger les données 'demo_amiegal.sav'.
get file='demo_amiegal.sav' .

*** sauvegarder sous autre nom ('demo1.sav').
save outfile='demo1.sav' .

** vérifier égalité entre                    .
* numéro, sex,  profper,  etude,    age    .
* numero, sexe, travpere, instruct, catage.

compute sel = 0.   /* valeur par défaut */
if travpere <> profper sel = 1.
if sexe <> sex sel = 2.
if instruct <> etude sel = 3.
if catage <> age sel = 4.
if numéro <> numero sel= 9.
```

```
value label
    sel 0 'toutes les variables égales'
        1 'travpere <> profper'
        2 'sexe <> sex'
        3 'instruct <> etude'
        4 'catage <> age'
        9 'numéro <> numéro'.

frequencies sel.

** filtrer et lister les erreurs.
compute filtre_ = 0.
if sel > 0 filtre_ = 1.
filter by filtre_.

list numéro numero travpere profper
   sexe sex instruct etude catage age.

filter off.
```

3.4 Filtre et sélection de variables

Pour travailler sur sous-ensemble des cas, deux possibilités :

FILTER BY var rend inactifs les cas non sélectionnés (pur lesquels var=0) sans les supprimer de la base de données courante.

Exemple : Distribution des femmes selon état matrimonial

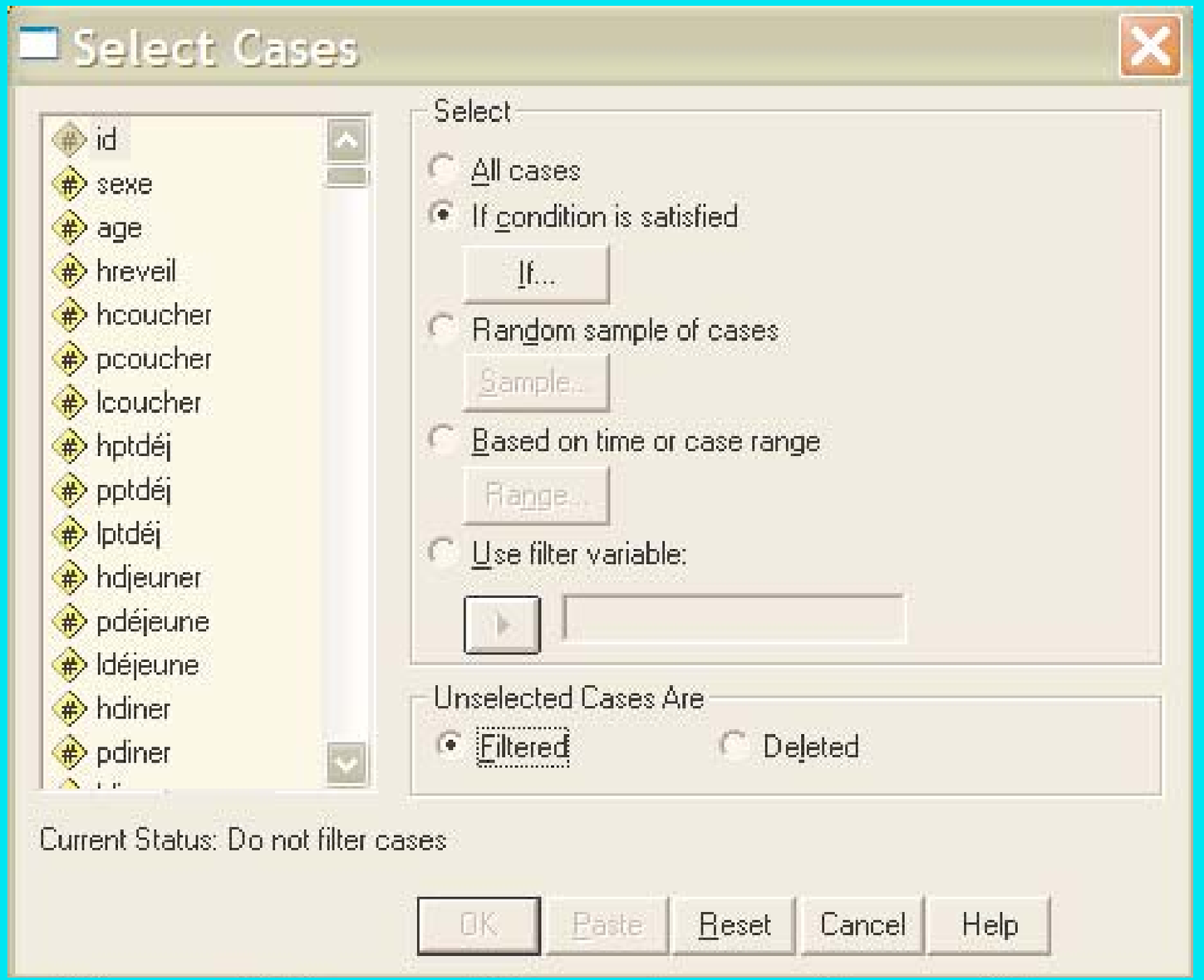
```
compute filtre = 0.
if sexe=2 filtre=1.
filter by filtre.
frequencies variable=etatmat.
filter off.
```

SELECT IF cond supprime de la base courante les cas ne vérifiant pas la condition.

Exemple : définir un fichier avec les femmes seulement

```
select if sexe=2.
save outfile='fichier_femmes.sav'.
```


Data
Transform
Analyz
Define Dates...
Insert Variable
Insert Cases
Go to Case...
Sort Cases...
Transpose...
Restructure...
Merge Files
Aggregate...
Orthogonal Design
Split File...
Select Cases...
Weight Cases...


Select Cases
id
sexe
age
hreveil
hcoucher
pcoucher
lcoucher
hptdéj
pptdéj
lptdéj
hdjeuner
pdéjeune
ldéjeune
hdiner
pdiner
Select
All cases
If condition is satisfied
If...
Random sample of cases
Sample...
Based on time or case range
Range...
Use filter variable:
Unselected Cases Are
Filtered
Deleted
Current Status: Do not filter cases
OK
Paste
Reset
Cancel
Help

3.5 Données manquantes

- Donnée manquante système (si pas de valeur entrée)
- Donnée manquante utilisateur (si = valeur définie par utilisateur)

Définir valeurs des données manquantes

```
MISSING VALUES var1 (7,8,9).
MISSING VALUES var1 ().    /* supprime les valeurs manquantes
```

déclarées.

```
MISSING VALUES all (999).  /* déclare valeur pour toutes les variables.
```

Fonctions de valeurs manquantes (utiliser dans *expression* ou *condition*)

`VALUE(var1)`	ignore le statut "missing".
`SYSMIS(var1)`	$= 1$ si valeur manquante système, 0 sinon
`MISSING(var1)`	$= 1$ si valeur manquante système ou utilisateur, 0 sinon.

Gestion des données manquantes

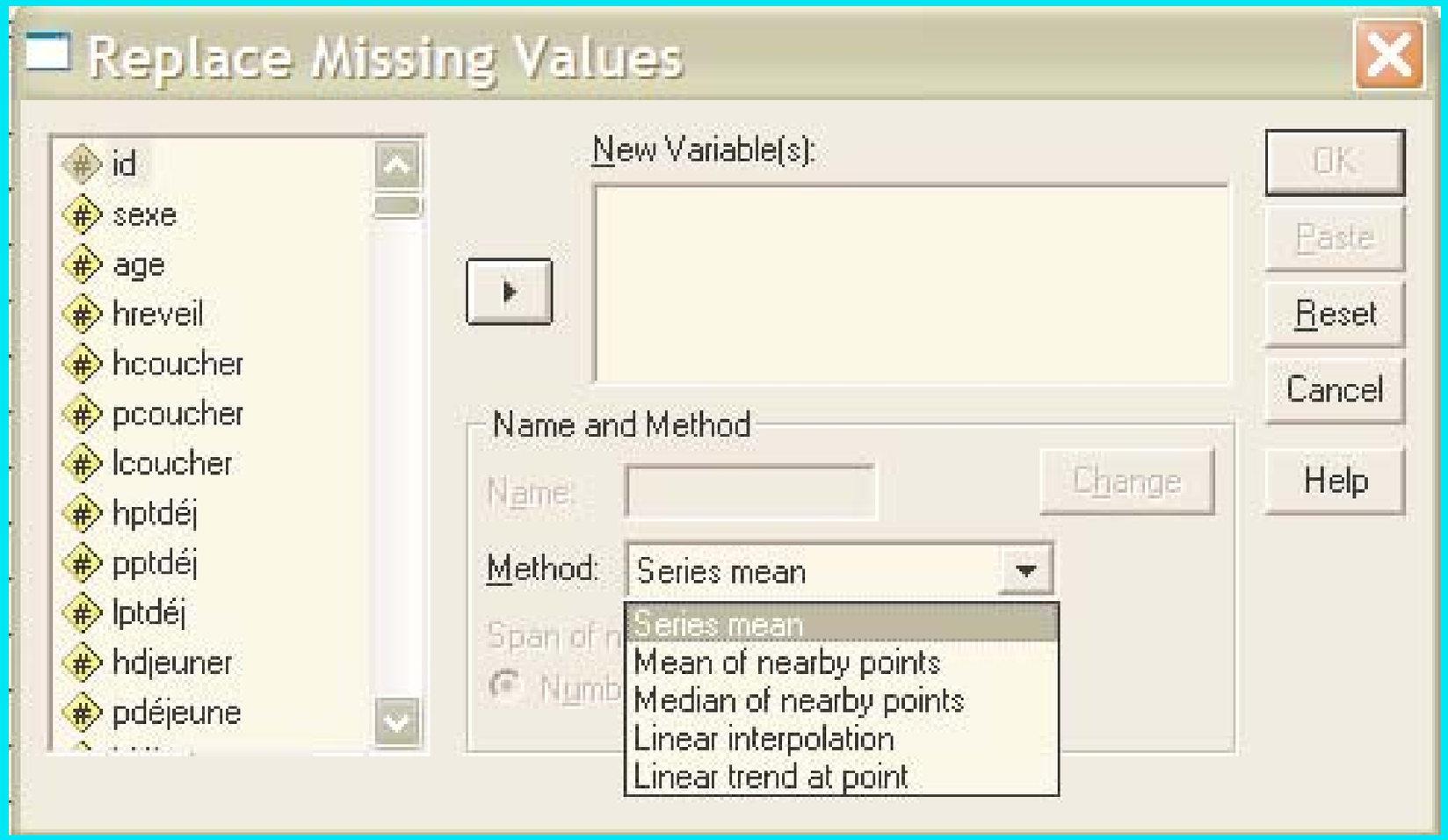
- Supprimer les cas avec données manquantes
  - listwise : cas avec valeur manquante dans une variable de la liste.
  - pairwise : lorsque le cas intervient effectivement dans un calcul.

- Imputer une valeur de substitution
  - Imputation fondée sur la seule variable (colonne)
    exemples : moyenne, moyenne des cas voisins, médiane, (mode), ...

    `RMV /age_1 = smean(age).` (voir aussi dialogue)

    Inconvénient : réduit la dispersion.
  - Imputation fondée sur valeurs prises par d'autres variables (lignes).
    (Plus complexe, non décrit ici.).

Menu : Transform/Replace Missing Values...



3.6 Exportation et importation de fichiers

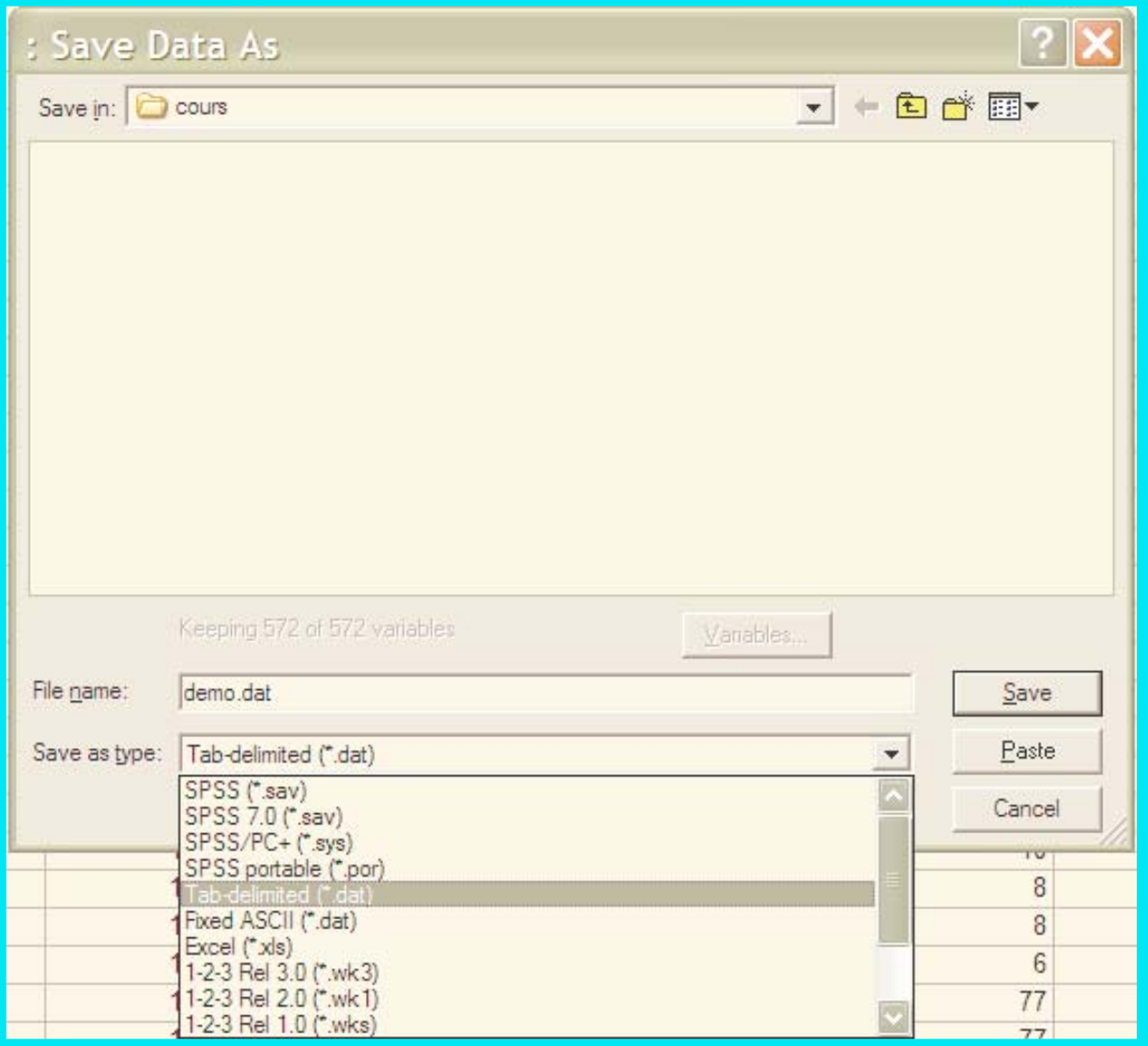
Sauvegarde

Menu : File/Save As...

```
SAVE OUTFILE='demo.sav'
  /DROP ec_rec.
```

Exportation

```
SAVE TRANSLATE
  OUTFILE='demo.dat'
  /TYPE=TAB
  /MAP
  /REPLACE
  /FIELDNAMES.
```



Ouverture fichier

Menu : File/Open/Data...

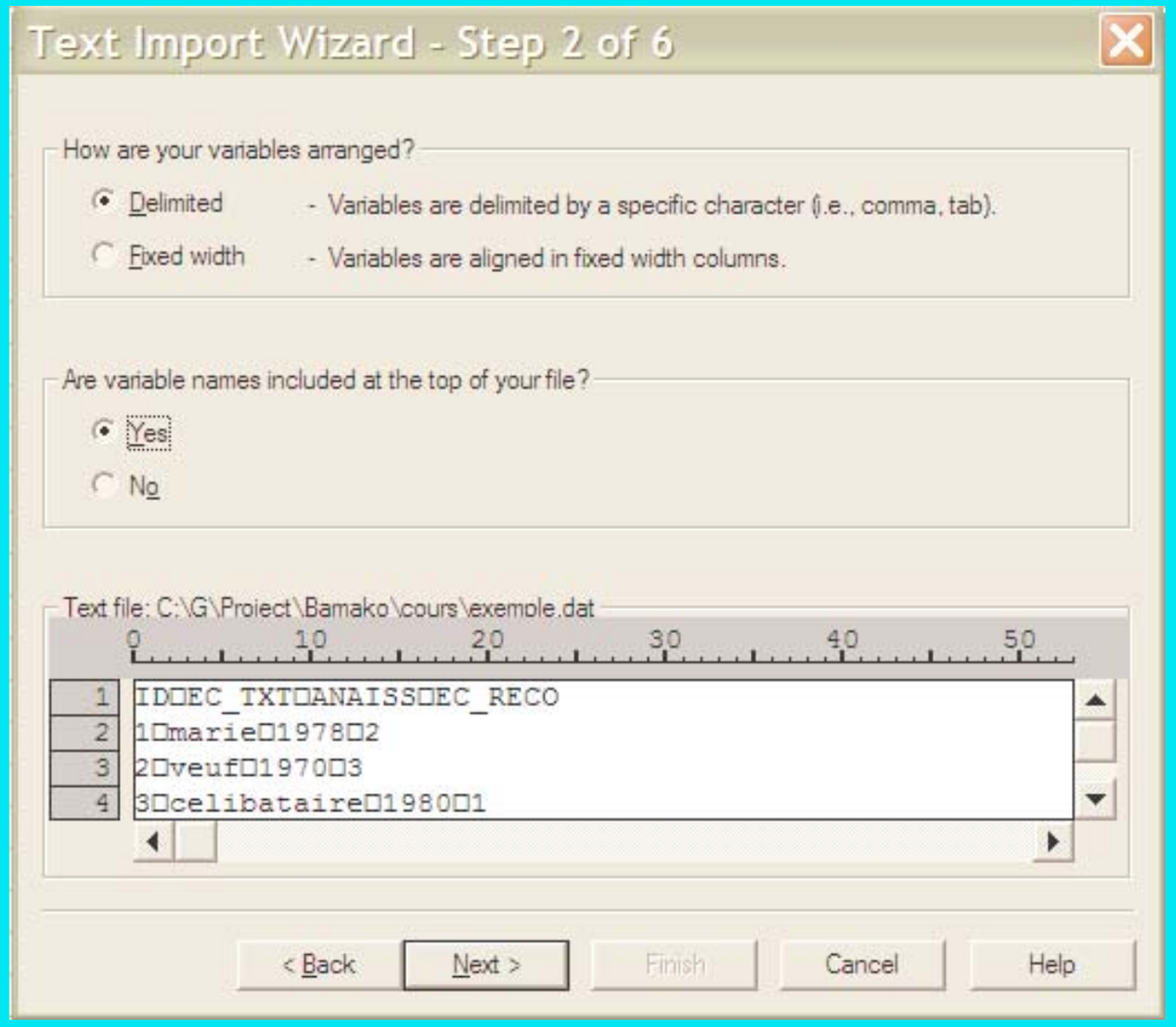
```
GET FILE='demo.sav'.
```

Importation

```
GET DATA
```

Syntaxe lourde.

Utiliser de préférence le dialogue File/Open/Data...



Format d'échange de fichier

Si l'autre logiciel lit/écrit le format SPSS (.sav), utiliser de préférence ce format (qui optimise la taille et préserve toutes les informations : étiquettes, format des variables, ...)

Sinon, les formats les plus courants d'échange sont :
- tab-delimited (TAB) : colonnes séparées par des tabulateurs.
- champs fixe (ASCII) : valeurs alignées en colonnes.
- Excel (XLS) : lit/crée un fichier Excel

l'option /fieldnames met les noms de variables dans la première ligne.

!!! toutes étiquettes de variable et de valeurs sont perdues !!!

A la lecture, il faut aussi préciser si le fichier contient les noms de variables en première ligne.

Le type des variables est déterminé selon la valeur qui est en première ligne. Exemple : 4.5 $\Rightarrow$ numérique, célibataire $\Rightarrow$ string.

3.7 Agrégation et fusion de fichiers

Agrégation : regrouper les cas selon une ou plusieurs variables (break)

Exemple : selon sexe, classe d'âge, quartier.

+ calculer pour chaque groupe des valeurs synthétiques des variables retenues : SUM, MEAN, MEDIAN, MIN, MAX, N (nbre cas), ...

Fusion : ajouter

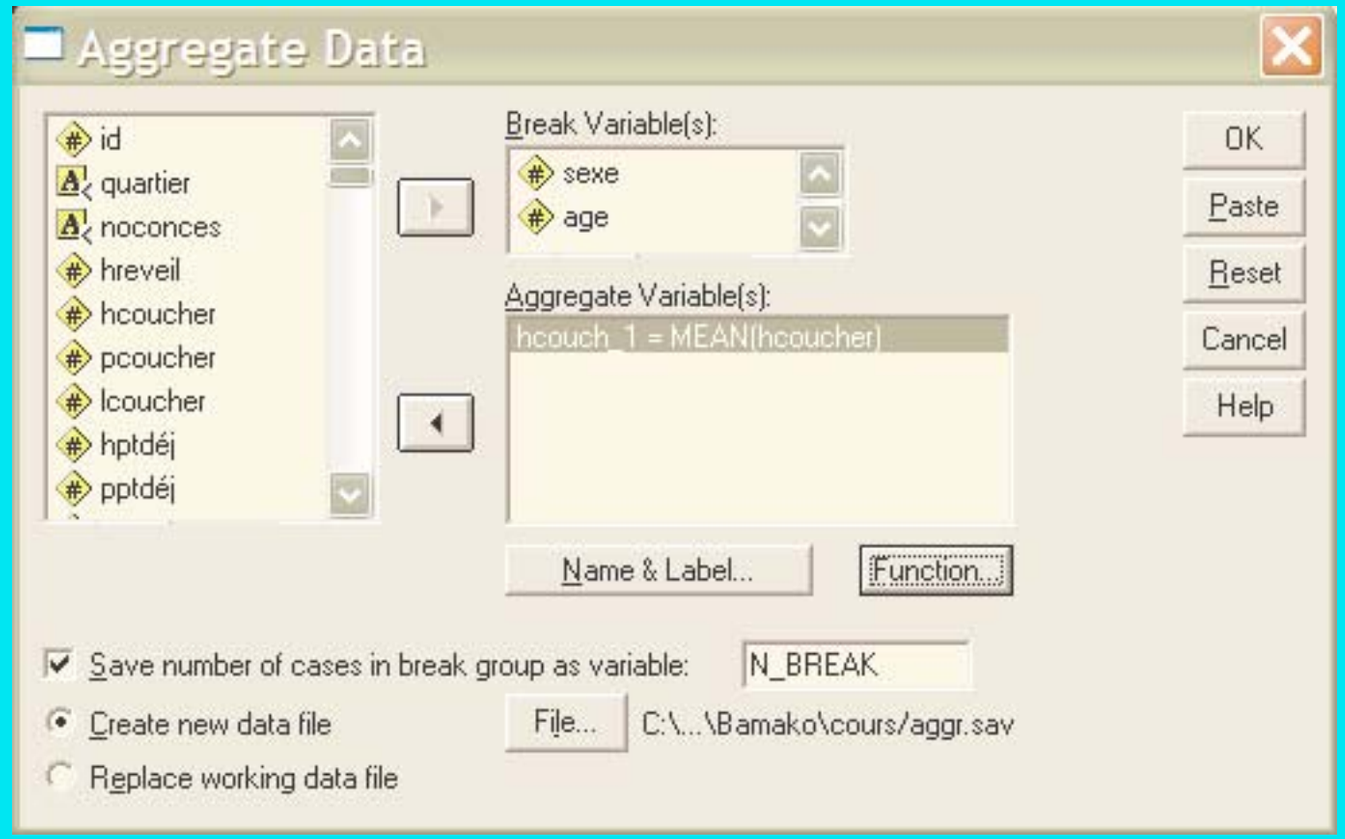
- les cas (lignes) d'un fichier à un autre (ADD FILES)
- les variables (colonnes) d'un fichier à un autre (MATCH FILES)
- mixte, mettre à jour un fichier avec données (cas et variables) d'un autre (UPDATE)

Les deux derniers requièrent un identificateur (key variables) unique pour chaque cas et commun aux deux fichiers.

Agrégation

Menu : Data/Aggregate...

```
AGGREGATE
  /OUTFILE='aggr.sav'
  /BREAK=sexe age
  /hcou_moy = MEAN(hcoucher)
  /N_BREAK=N.
get file='aggr.sav'.
list all.
```



L'exemple précédent à partir d'un fichier de 40 cas produit :

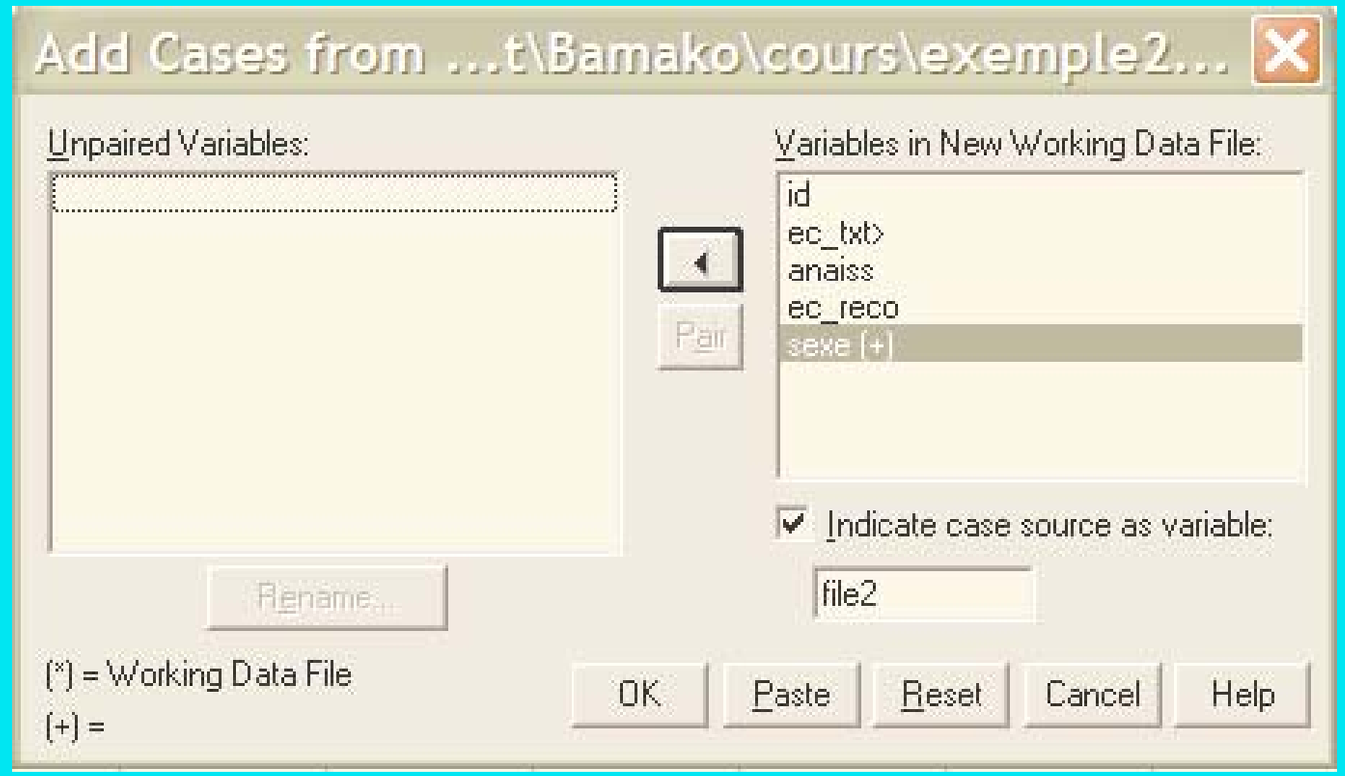
SEXE	AGE	HCOU_MOY	N_BREAK
1	13	22.00	1
1	14	24.00	1
1	15	21.00	2
1	16	21.33	3
1	17	19.50	6
1	18	22.75	4
1	19	22.33	3
1	20	18.20	5
1	21	1.67	3
1	22	22.50	2
1	24	24.00	1
1	25	6.00	1
1	28	23.50	2
1	29	8.33	3
1	30	11.75	4

Fusion

Menu : Data/Merges Files

Add cases

```
ADD FILES /FILE=*
 /FILE='exemple2.sav'
 /IN=file2.
VARIABLE LABELS file2
 'Case de exemple2.sav'.
EXECUTE.
```



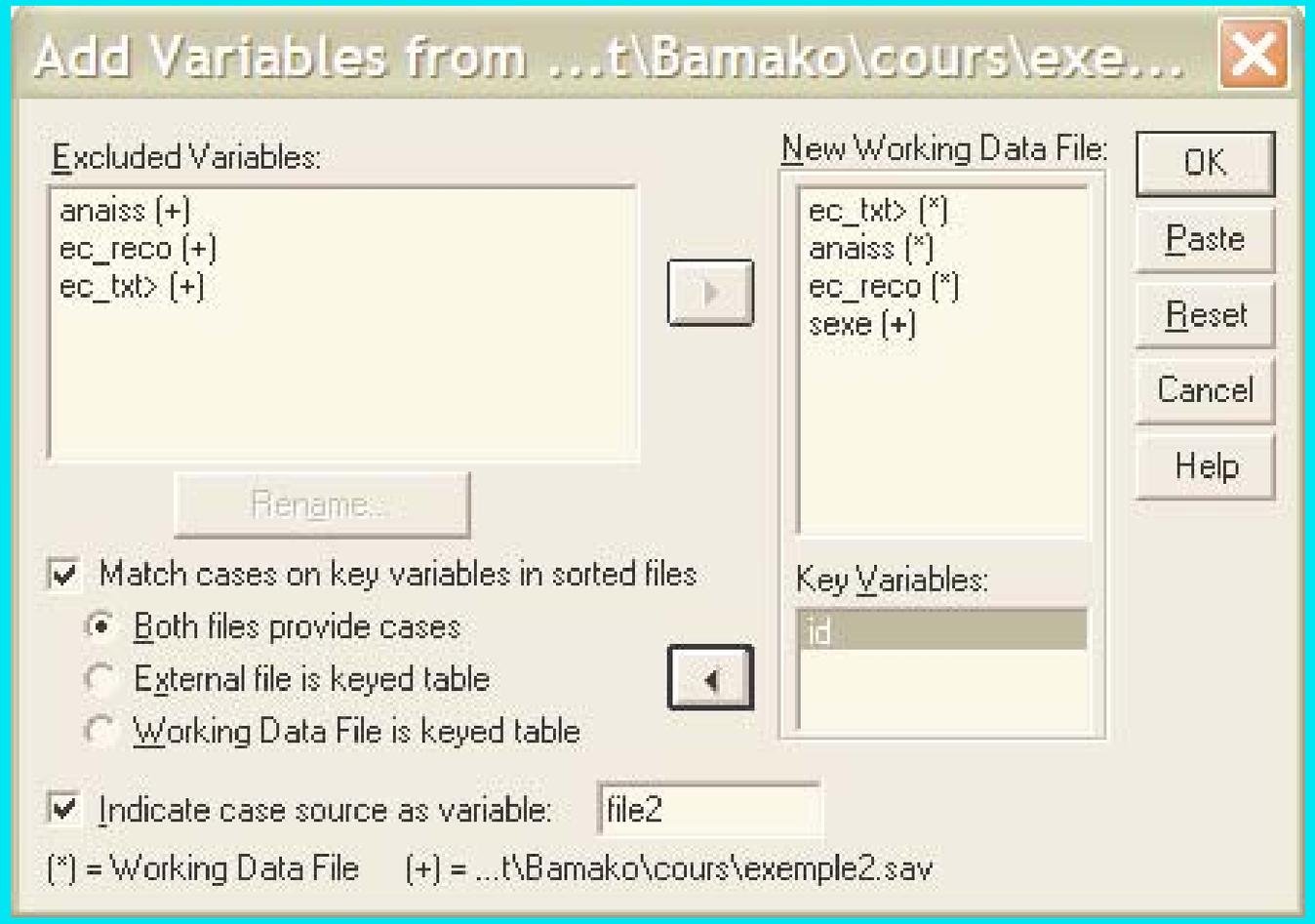
ID	EC_TXT	ANAISS	EC_RECO	SEXE	FILE2
1	marie	1978	2	.	0
2	veuf	1970	3	.	0
3	celibataire	1980	1	.	0
4	celibataire	1988	1	.	0
1	marie	1978	2	1	1
2	veuf	1970	3	1	1
3	celibataire	1980	1	2	1
5	celibataire	1988	1	2	1

Fusion

Menu : Data/Merges Files

Add Variables

```
MATCH FILES /FILE=*
 /FILE='exemple2.sav'
 /RENAME
   (anaiss ec_reco ec_txt
    = d0 d1 d2)
 /IN=file2
 /BY id
 /DROP= d0 d1 d2.
EXECUTE.
```



ID	EC_TXT	ANAISS	EC_RECO	SEXE	FILE2
1	marie	1978	2	1	1
2	veuf	1970	3	1	1
3	celibataire	1980	1	2	1
4	celibataire	1988	1	.	0
5		.	.	2	1

UPDATE (seulement par syntaxe)

```
UPDATE FILE='exemple.sav'
    /IN= updated
    /FILE= 'exemple2.sav'
    /BY id .
EXECUTE.
list all.
```

produit :

ID	EC_TXT	ANAISS	EC_RECO	SEXE	UPDATED
1	marie	1978	2	1	1
2	veuf	1970	3	1	1
3	celibataire	1980	1	2	1
4	celibataire	1988	1	.	1
5	celibataire	1988	1	2	0

4 Analyse statistique descriptive

Données univariées

1. Tableau de distribution
2. Présentations graphiques
3. Indicateurs statistiques
   - positionnement
   - dispersion, asymétrie et aplatissement

Données bivariées

1. Tableau croisé et présentation graphique de données bivariées
2. Association, Indépendance
3. Corrélation et autres mesures d'association

4.1 Tableau de distribution

Variables discrètes sans trop (< 15) de modalités.

$\Rightarrow$ tableau des fréquences généré avec FREQUENCIES

heure pt déjeuner

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	1	2.4	4.3	4.3
	6	1	2.4	4.3	8.7
	7	5	12.2	21.7	30.4
	8	10	24.4	43.5	73.9
	9	2	4.9	8.7	82.6
	10	3	7.3	13.0	95.7
	11	1	2.4	4.3	100.0
	Total	23	56.1	100.0	
Missing	non concerné	18	43.9		
Total		41	100.0		

Variables continues ou avec beaucoup de modalités.

$\Rightarrow$ regrouper les valeurs en classes de valeurs.

Exemple : Regrouper les valeurs de «hcoucher» en 6 classes de longueurs égales.

(voir syntaxe page suivante)

HCOU_CLS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-4	9	22.0	22.0	22.0
	5-8	1	2.4	2.4	24.4
	17-20	1	2.4	2.4	26.8
	21-24	30	73.2	73.2	100.0
	Total	41	100.0	100.0	

```
get file='donnéesbama1.sav'.
** recherche du min et du max .
compute temp=1.
AGGREGATE  /OUTFILE='aggr.sav'
  /BREAK=temp  /hcou_max = MAX(hcoucher)  /hcou_min = MIN(hcoucher)

** ajouter les min et max à chaque cas du fichier original.
MATCH FILES /TABLE='aggr.sav'
   /FILE=*         /* base de données courante */
   /BY temp.
VARIABLE LABELS
   hcou_min 'min(hcoucher)'  hcou_max 'max(hcoucher)'.

** calcul de la classe.
compute delta_h = (hcou_max - hcou_min)/6. /* longueur du pas */
compute hcou_cls=1.                        /* initialisation */
loop #i = 1 to 5.
+    if (hcoucher > hcou_min + #i*delta_h) hcou_cls = (#i+1).
end loop.
value labels hcou_cls
    1 '0-4' 2 '5-8' 3 '9-12' 4 '13-16' 5 '17-20' 6 '21-24'.
frequencies hcou_cls.
```

Quantiles

Découpage en classes d'égale longueur pas satisfaisant pour notre exemple.

Il est préférable de définir les classes en fonctions de quantiles.

Exemple :

- 6 classes ayant chacune la même proportion de cas.
- Groupes 1, 2, 5, 6 : 20% chacun, Groupes 3 et 4 : 10% chacun
  $\Rightarrow$ quantiles 20%, 40%, 50%, 60%, 80%

Statistics

heure coucher

N	Valid	41
	Missing	0
Percentiles	20	2.40
	40	21.00
	50	22.00
	60	22.20
	80	23.60

Statistics

heure coucher

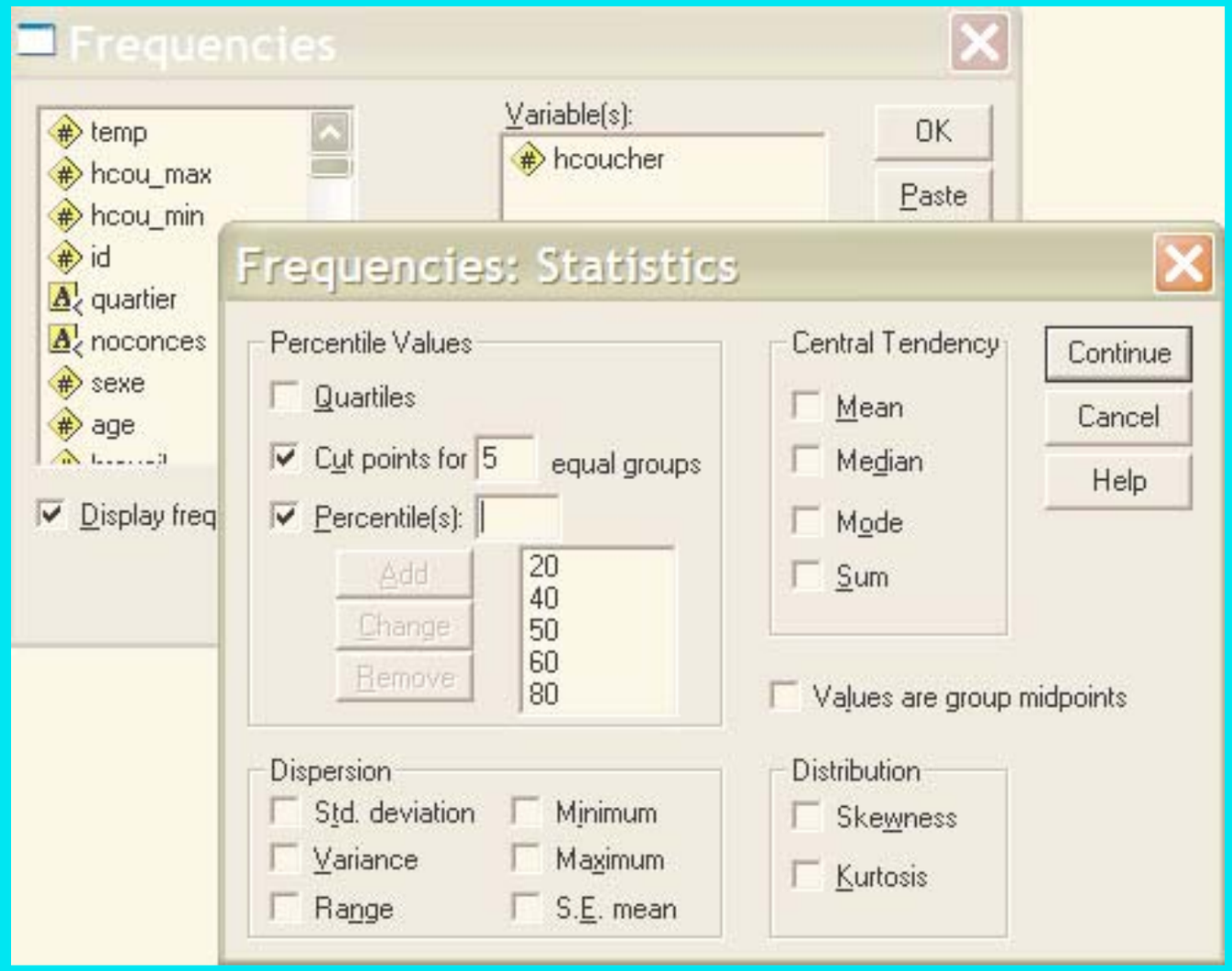
N	Valid	41
	Missing	0
Percentiles	16.66666667	2.00
	33.33333333	21.00
	50	22.00
	66.66666667	23.00
	83.33333333	24.00

percentiles

```
FREQUENCIES
  VARIABLES=hcoucher
  /FORMAT=notable
  /PERCENTILES=
     20 40 50 60 80.
```

ntiles

```
FREQUENCIES
  VARIABLES=hcoucher
  /FORMAT=notable
  /NTILES= 6.
```



4.2 Présentations graphiques

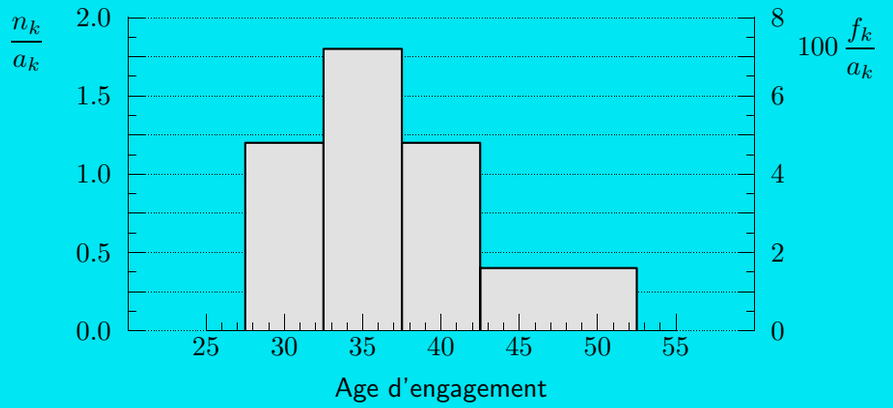
Principe : surfaces proportionnelles aux grandeurs illustrées.

Barres : hauteurs proportionnelles aux fréquences si bases égales.

Histogramme (diagramme en barres pour données quantitatives)

– Partition des valeurs en classes disjointes
  $\Rightarrow$ les barres se touchent

– Bases des barres représentent amplitude $\Rightarrow$ si amplitudes pas toutes égales, adapter hauteurs pour avoir surfaces proportionnelles.



classe d'âge	amplitude	nombre
27.5 – 32.5	5	6
32.5 – 37.5	5	9
37.5 – 42.5	5	6
42.5 – 52.5	10	4

4.3 Indicateurs statistiques

n données : $x_1, x_2, \ldots, x_n$,
exemple : 20, 25, 25, 30, 50

Tendance centrale :
mode, médiane, moyenne

Mode : valeur la plus fréquente
Exemple : mode = 25

Médiane :

$$50\% \text{ des } x_i \leq \text{med} \quad \text{et} \quad 50\% \text{ des } x_i \geq \text{med}$$

Exemple : med = 25

Moyenne :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Exemple : $\bar{x} = 150/5 = 30$

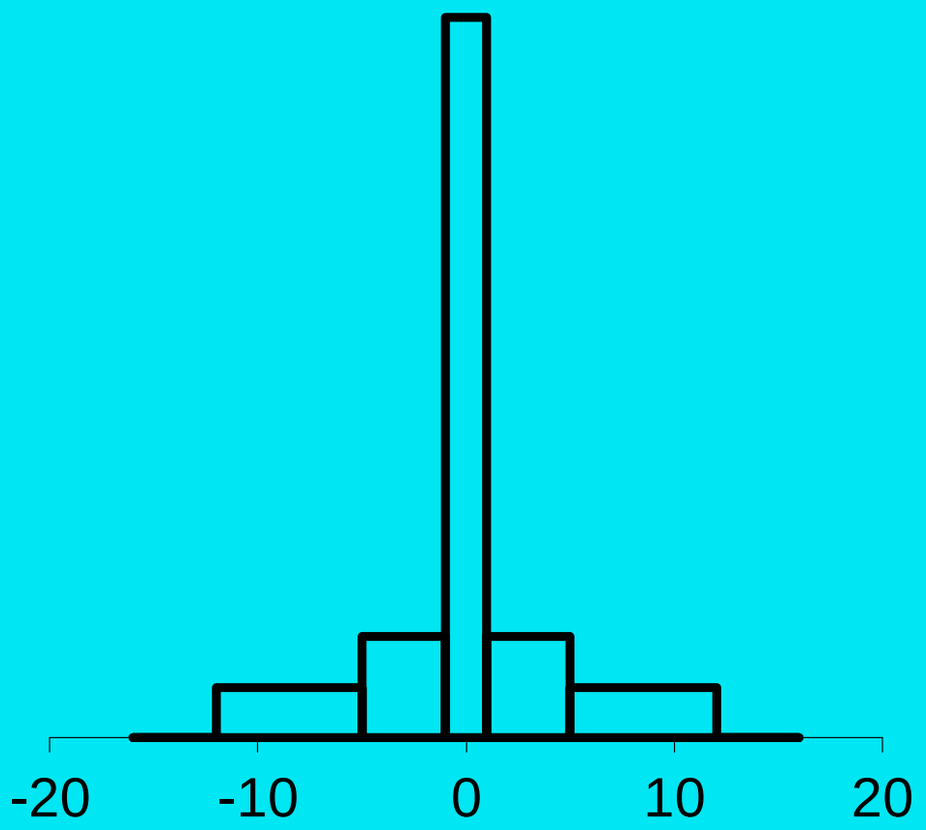
Moyenne pondérée :

$$\bar{x} = \sum_{i=1}^{n} w_i x_i \qquad \text{avec} \sum_i w_i = 1$$

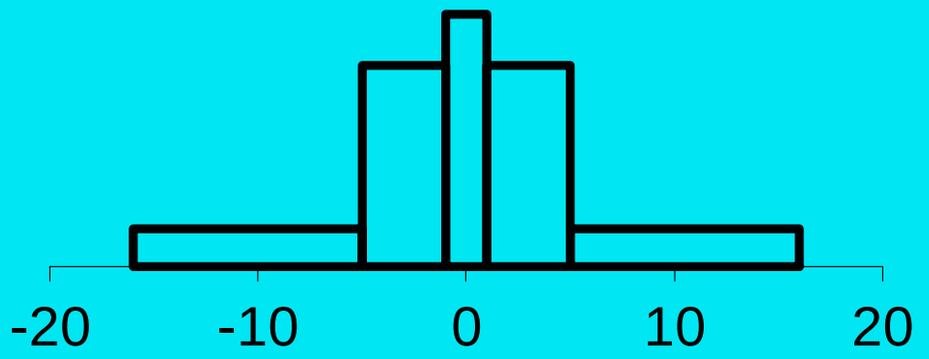
Exemple : $w_1 = w_5 = \frac{1}{8}$ et $w_2 = w_3 = w_4 = \frac{2}{8}$

$\Rightarrow \bar{x} = 230/8 = 28,75$

Dispersion : variance et écart type



dispersion plus faible



dispersion plus forte

Variance : moyenne des carrés des écarts à la moyenne

$$\mathrm{var}(x) = \sum_{i=1}^{n} w_i (x_i - \bar{x})^2$$

Écart type :

$$\text{écart type}(x) = s_x = \sqrt{\mathrm{var}(x)}$$

Écart interquartile : $q_3 - q_1$

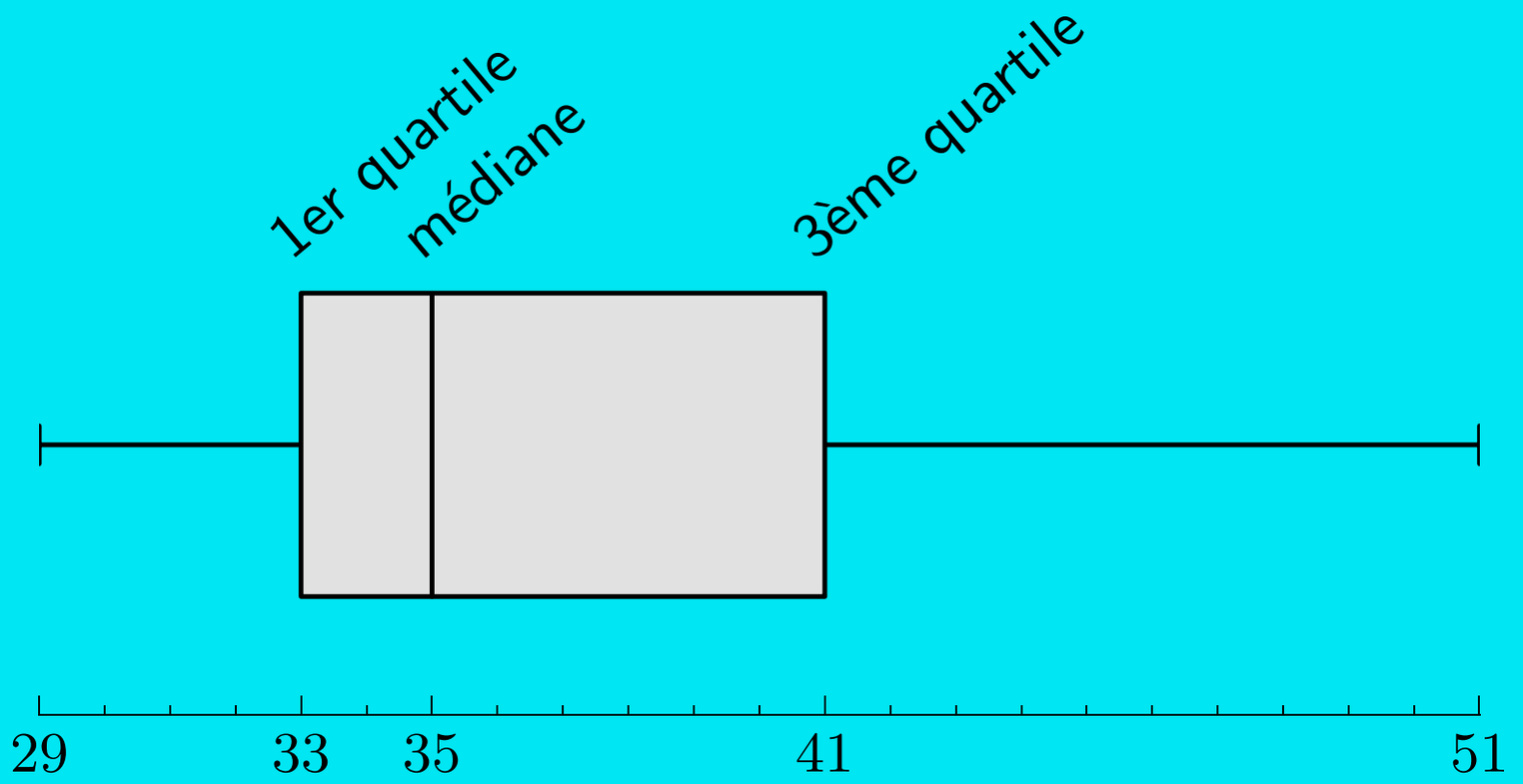
q_1 : 1er quartile

25% des $x_i \leq q_1$ et 75% des $x_i \geq q_1$

q_3 : 3ème quartile

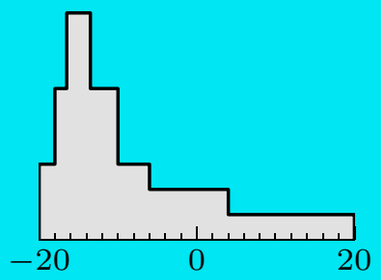
75% des $x_i \leq q_3$ et 25% des $x_i \geq q_3$

Boxplot

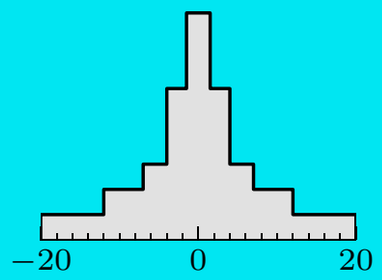


Voir procédure EXPLORE (menu/descriptives/explore)

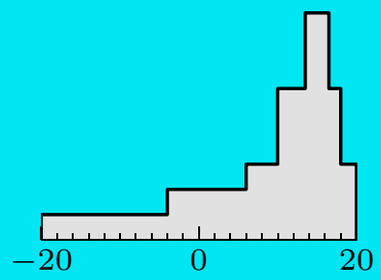
Asymétrie et aplatissement (kurtose)



asymétrie positive



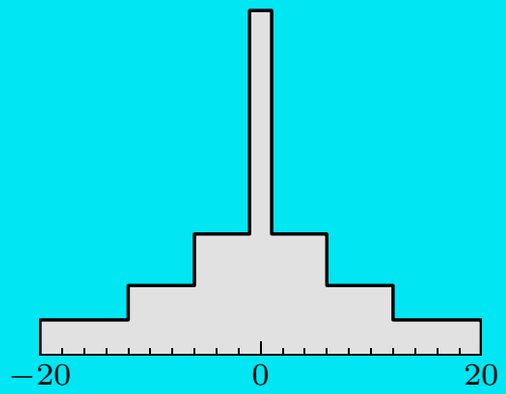
symétrie



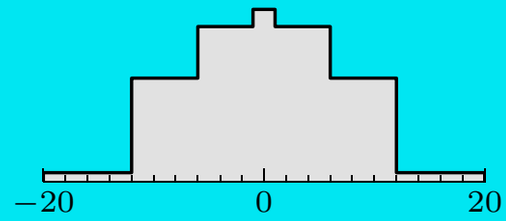
asymétrie négative

$\lambda > 0$ étalement à droite

$\lambda < 0$ étalement à gauche



forme en pic



forme aplatie

kurt grand : pic et grosses queues

kurt petit : aplatissement

4.4 Tableau croisé et présentation graphique de données bivariées

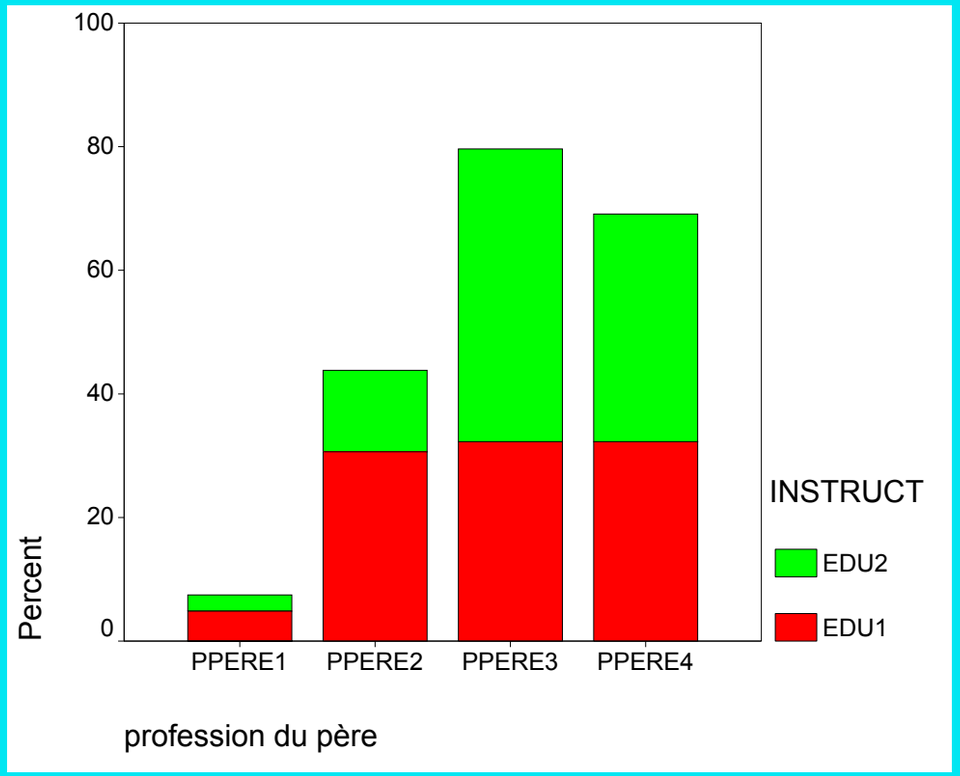
menu : Analysis/Descriptives/Crosstabs...

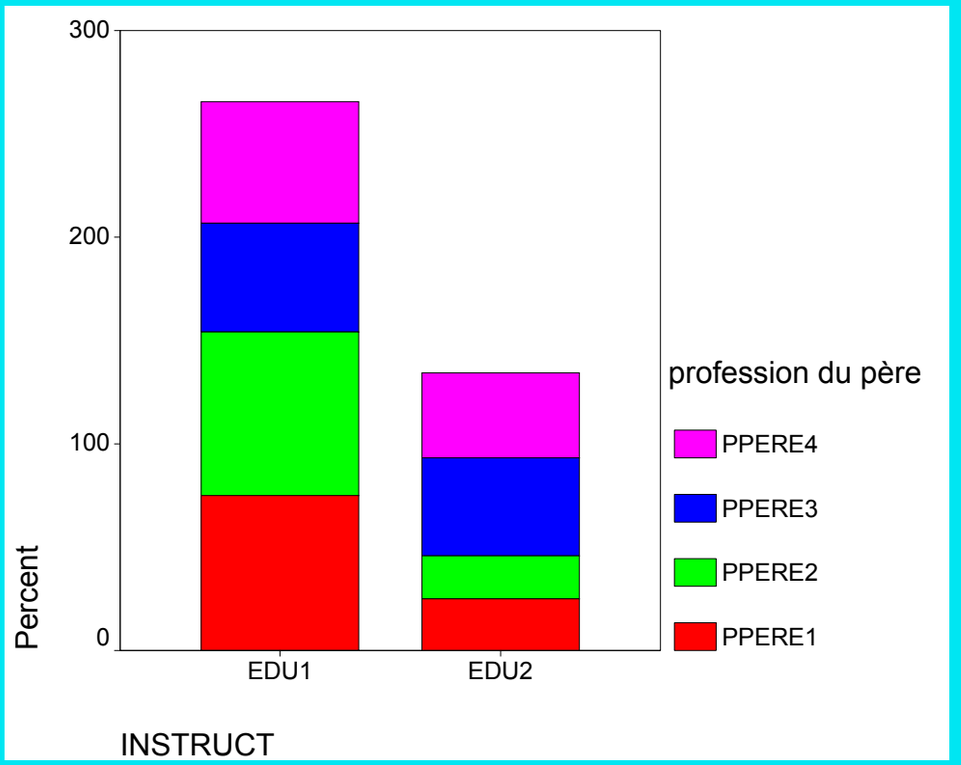
INSTRUCT * profession du père Crosstabulation

			profession du père				
			PPERE1	PPERE2	PPERE3	PPERE4	Total
INSTRUCT	EDU1	Count	3	19	20	20	62
		Expected Count	2.5	14.9	23.6	21.1	62.0
		% within INSTRUCT	4.8%	30.6%	32.3%	32.3%	100.0%
		% within profession du père	75.0%	79.2%	52.6%	58.8%	62.0%
	EDU2	Count	1	5	18	14	38
		Expected Count	1.5	9.1	14.4	12.9	38.0
		% within INSTRUCT	2.6%	13.2%	47.4%	36.8%	100.0%
		% within profession du père	25.0%	20.8%	47.4%	41.2%	38.0%
Total		Count	4	24	38	34	100
		Expected Count	4.0	24.0	38.0	34.0	100.0
		% within INSTRUCT	4.0%	24.0%	38.0%	34.0%	100.0%
		% within profession du père	100.0%	100.0%	100.0%	100.0%	100.0%

Graphiques

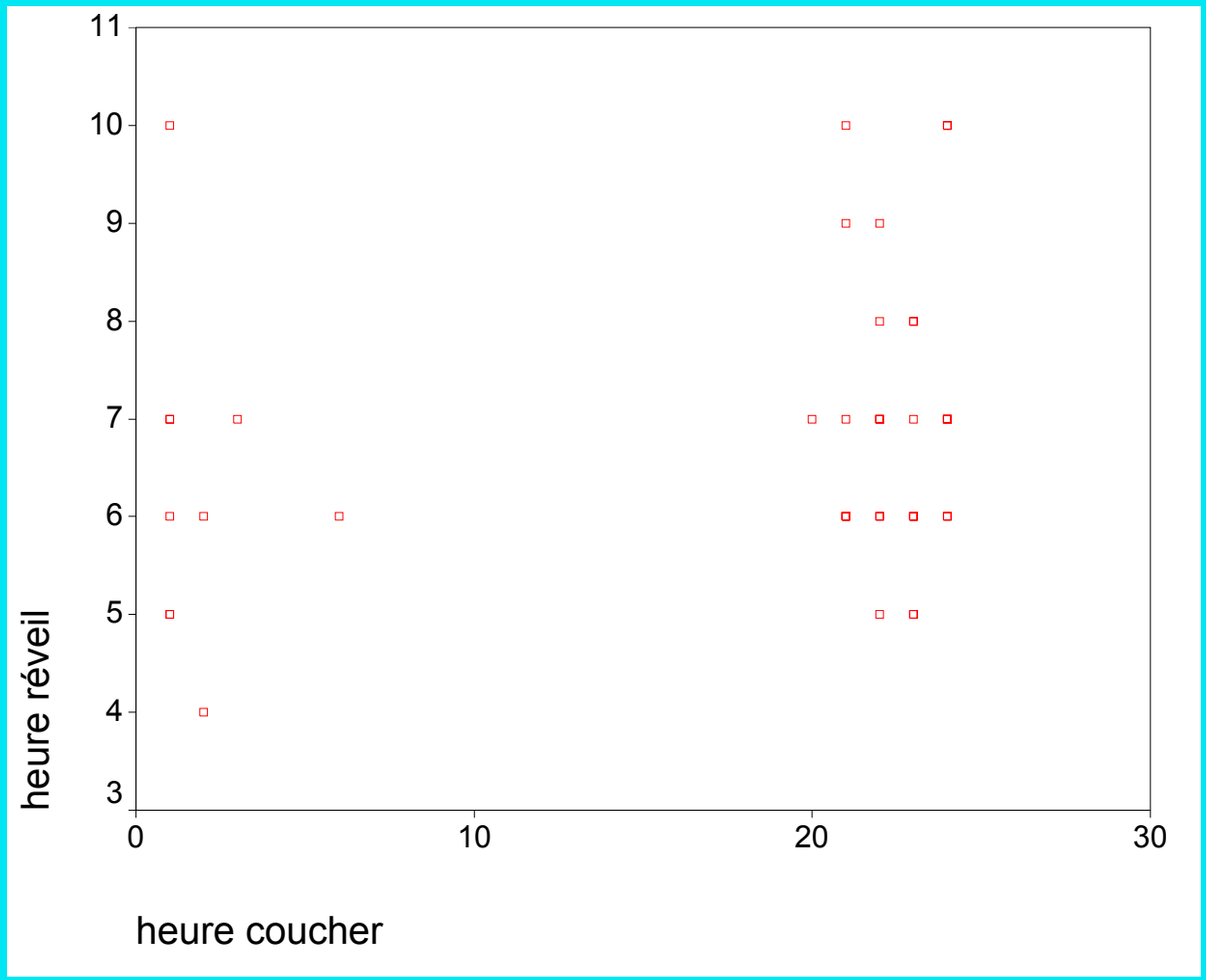
menu : Graphs/Bars





Données quantitatives : diagramme de dispersion (scatterplot)

menu : Graphs/Scatter



4.5 Association, Indépendance

Indépendance

Si distributions lignes semblables $\leftrightarrow$ Si distributions colonnes semblables

	B1	B2	total
A1	10	30	40
A2	20	60	80
total	30	90	120

Association Si distribution ligne dépend de la ligne.

Association parfaite

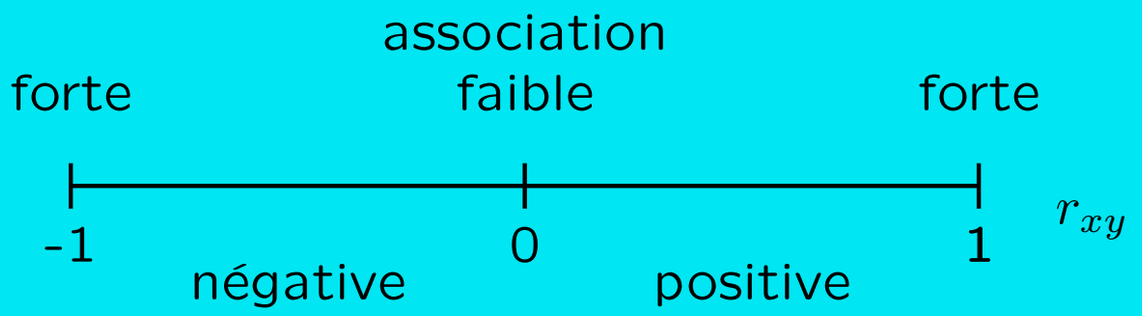
Si un seul élément non nul par colonne (ou ligne)

	B1	B2	total
A1	40	0	40
A2	0	80	80
total	40	80	120

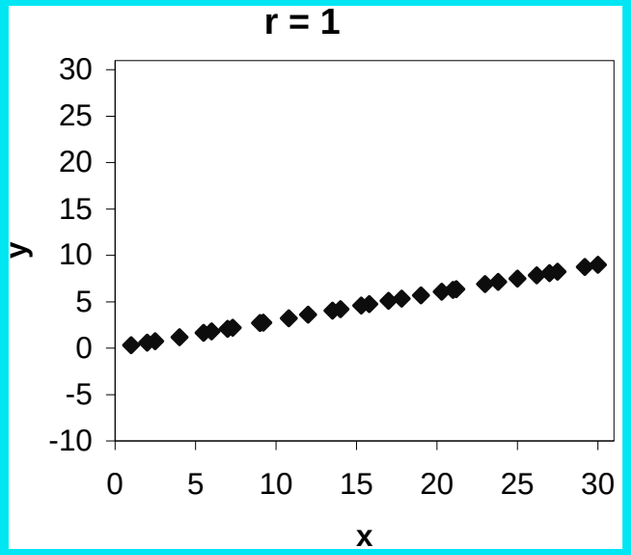
4.6 Corrélation et autres mesures d'association

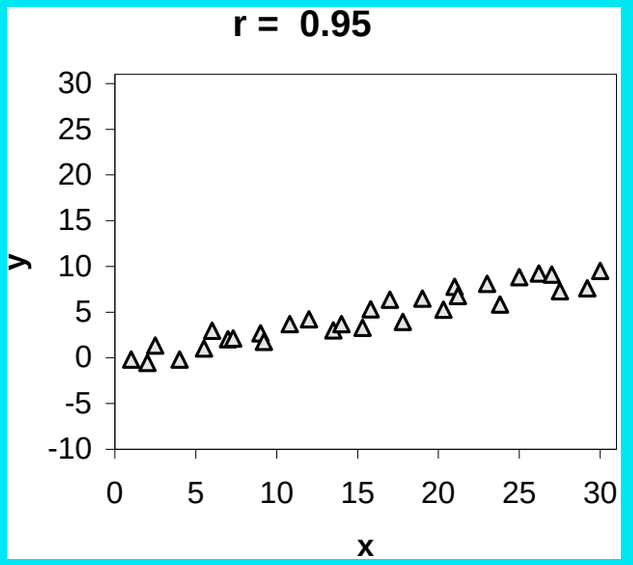
Coefficient de corrélation linéaire r_{xy}

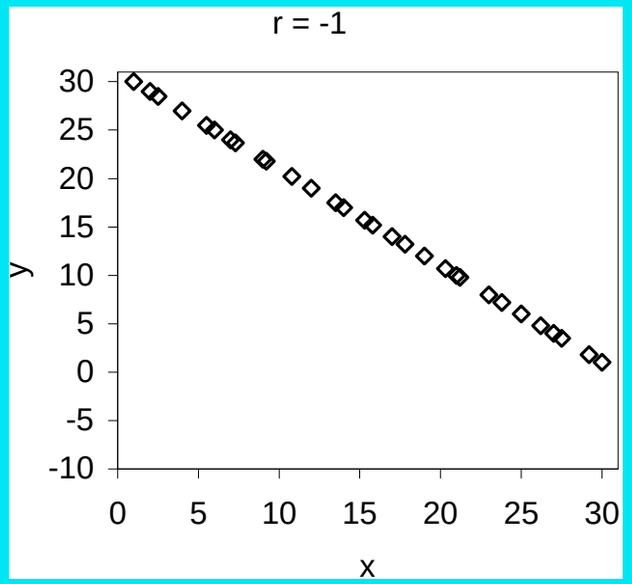
Mesure la tendance des points à s'aligner le long le d'une droite, c'est-à-dire l'allongement du nuage de points

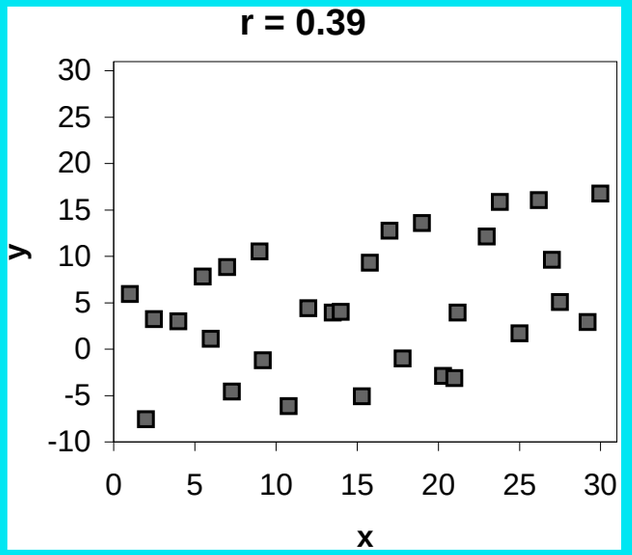


Corrélation : exemples









Khi-2 de Pearson et v de Cramer

(Tableau croisé)

Khi-2 de Pearson : distance entre distribution conjointe et distribution indépendante.

Khi-2 $= 0 \Leftrightarrow$ toutes les distributions ligne (colonne) identiques.

Dépend du nombre d'observations et de la dimension du tableau

v de Cramer : forme standardisée du Khi-2 de Pearson

v Cramer : exemples

taille	revenu [20,30[	[30,50[	[50,60]	total
1 ou 2	0	50	0	50
3 ou 4	25	0	25	50
total	25	50	50	100

$\Rightarrow$ corrélation $= 0$, v-Cramer $= 1$.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.850[a]	3	.183
Likelihood Ratio	5.107	3	.164
Linear-by-Linear Association	2.212	1	.137
N of Valid Cases	100		

a. 2 cells (25.0%) have expected count less than 5. The minimum expected count is 1.52.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.220	.183
	Cramer's V	.220	.183
N of Valid Cases		100	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

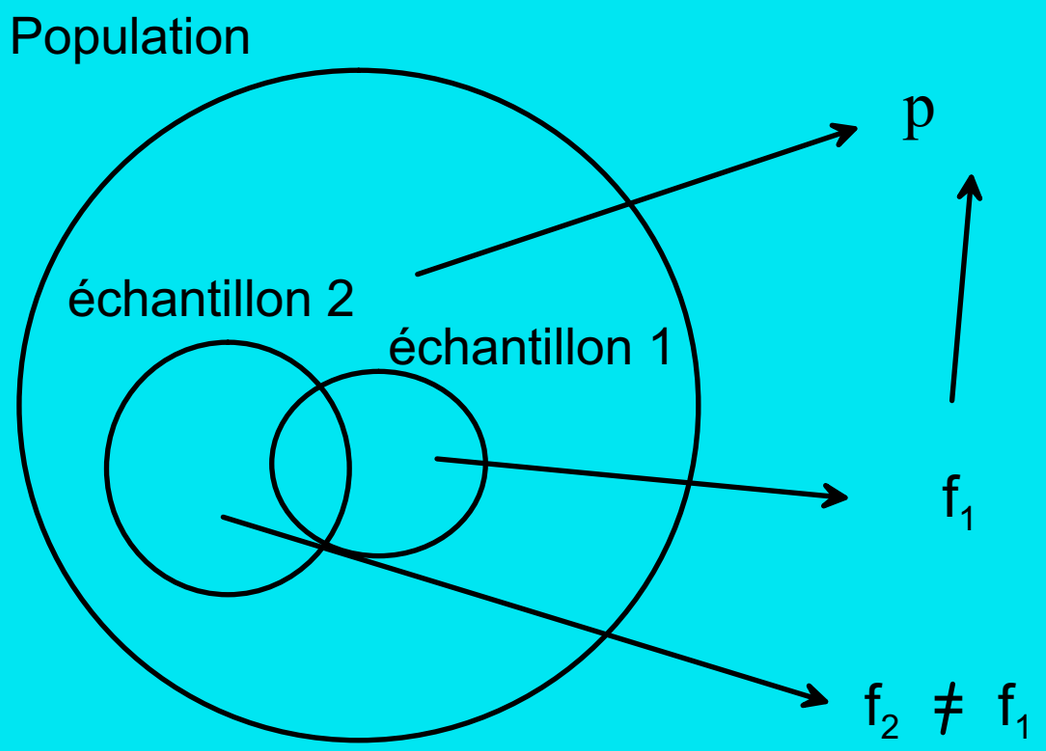
5 Éléments de statistique inférentielle

1. Estimation ponctuelle: biais, variance, erreur
2. Estimation par intervalle: marge d'erreur
3. Principe du test statistique d'hypothèse

Statistique inférentielle

Quelle information l'observation d'un échantillon donne-t-il sur le tout ?

$\Rightarrow$ Evaluation de la confiance, marge d'erreur,...



Aléa de l'échantillonnage

Echantillon aléatoire de taille n : $(X_1, X_2, \ldots, X_n)$

Le résultat x_i obtenu au $i$ème tirage diffère d'un échantillon à l'autre.

Toute fonction (moyenne, variance, proportion, ...) de l'échantillon aléatoire est aléatoire.

Estimation : *quantification*, à partir de l'échantillon, de la valeur d'une caractéristique numérique de la population (moyenne μ, variance σ^2, proportion p, corrélation ρ, ...).

Test statistique : *validation* empirique d'une hypothèse.

L'estimation ou la conclusion du test varie d'un échantillon à l'autre.

Fiabilité de l'estimation ou du test ?

5.1 Estimation ponctuelle : biais, variance, erreur

Absence de biais :
espérance de l'estimateur (estimation moyenne)
= vraie valeur du paramètre.

Efficacité :
faible dispersion des valeurs d'un échantillon à l'autre

Erreur quadratique moyenne : EQM $= \mathrm{var} + \text{biais}^2$ petit

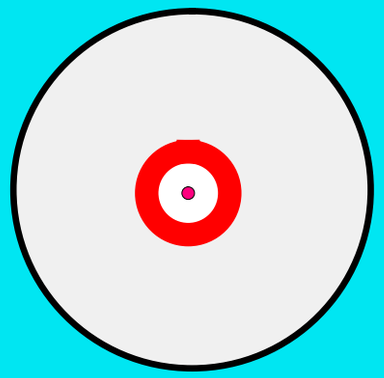
Exemple : $\bar{X}$ estimateur de μ

Si tirage au hasard (même prob pour chacun) et indépendants

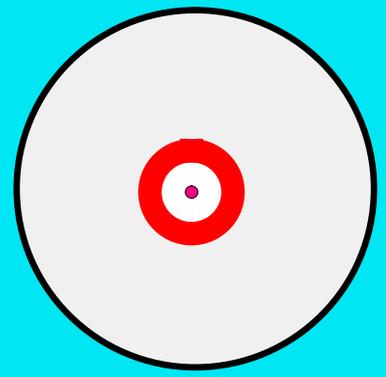
$$E(\bar{X}) = \mu \quad \text{et} \quad \mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$$

D'autant plus efficace que n est grand.

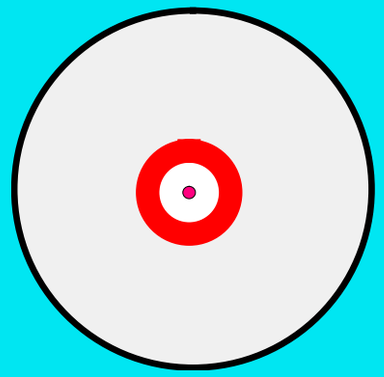
Estimation : analogie avec tir sur une cible



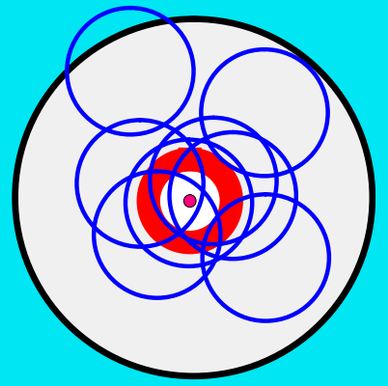
Estimation non biaisée



Estimation biaisée



Estimation plus efficace



Estimation par intervalle

5.2 Estimation par intervalle : marge d'erreur

Degré de confiance : probabilité d'obtenir a priori un intervalle qui comprend la vraie valeur.

La longueur de l'intervalle croît avec

- le degré de confiance
- la dispersion de l'estimateur (erreur standard $\hat{\sigma}_{\bar{X}}$)

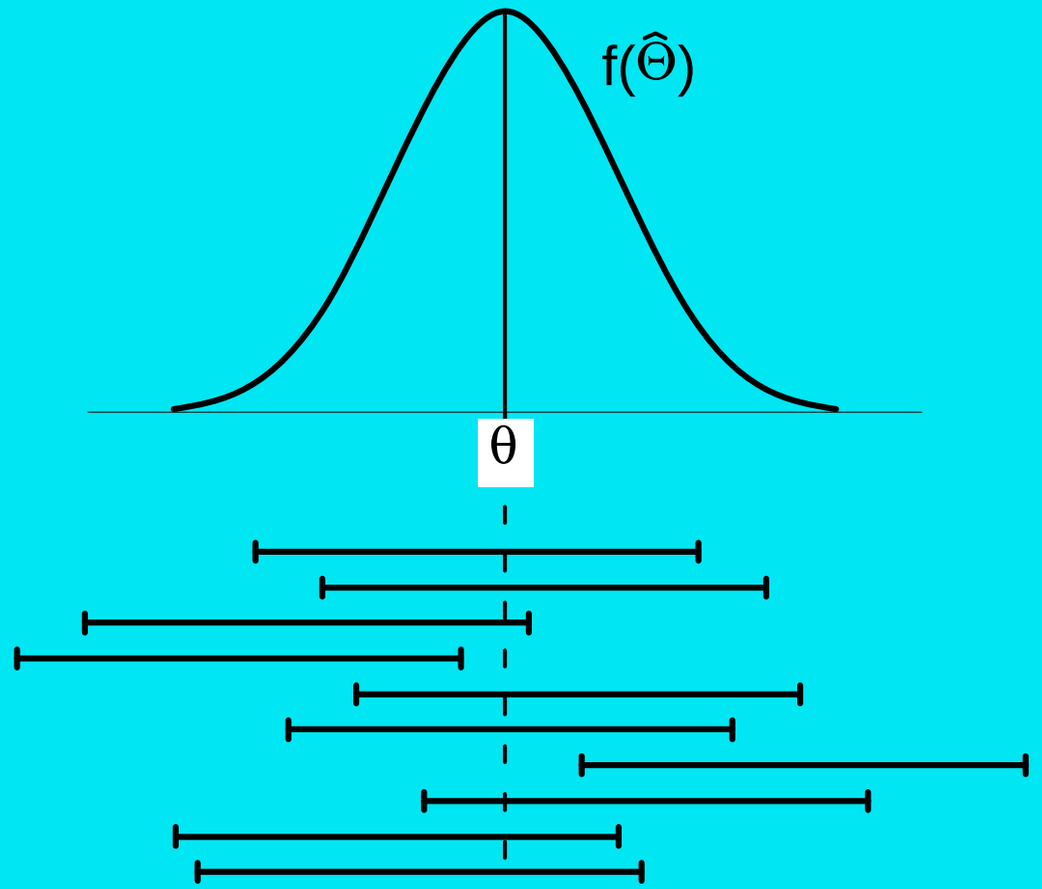
Exemple : intervalle pour la moyenne μ

$$\mu = \bar{x} \pm z_{1-\alpha/2}\hat{\sigma}_{\bar{X}}$$

où $z_{1-\alpha/2}$ est un seuil critique qui augmente avec le degré de confiance $1-\alpha$.

Confiance $1-\alpha = 95\% \Rightarrow z_{1-\alpha/2} \simeq 2$

Cas d'un intervalle centré sur $\hat{\Theta}$



10 intervalles (10 échantillons)

8 intervalles recouvrent l'estimé θ

Marge d'erreur

- erreur d'échantillonnage
  - biais
  - erreur aléatoire
- erreur d'observation
- erreur d'interprétation

En général, la marge d'erreur concerne l'erreur d'échantillonnage, qui est, si le biais est nul, l'erreur aléatoire.

Marge d'erreur $\simeq 2 \times$ erreur standard

Pour l'estimation d'une proportion p, c'est approximativement

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Par exemple, si l'on estime qu'une proportion est de 30% avec un échantillon de $n = 100$ personnes,

$$\begin{aligned} \text{marge d'erreur} &\simeq 2\sqrt{\frac{0,3 \cdot 0,7}{100}} \\ &= 2\sqrt{0,0021} = 2 \cdot 0,046 \\ &= 0,092 \end{aligned}$$

$\Rightarrow$ proportion $= 30\% \pm 9,2\%$

5.3 Principe du test statistique d'hypothèse

On rejette l'hypothèse s'il est peu vraisemblable d'obtenir l'échantillon observé par tirage au hasard dans une population vérifiant l'hypothèse.

Test de H_0 contre H_1

Règle de décision fondée sur une statistique Q quantité (p.ex. Khi-2 pour test de l'indépendance) dépendant de l'échantillon.

Région critique (rejet de H_0) : ensemble des valeurs de Q peu probables (¡ 5%) lorsque H_0 vrai.

Risque d'erreur

Deux risques d'erreur

	Décision	
Etat de la nature	H_0	H_1
H_0		α
H_1	β	

On contrôle α : risque de première espèce.

β : risque de seconde espèce.

$1-\beta$: puissance du test.

Plus α est petit, plus β explose.

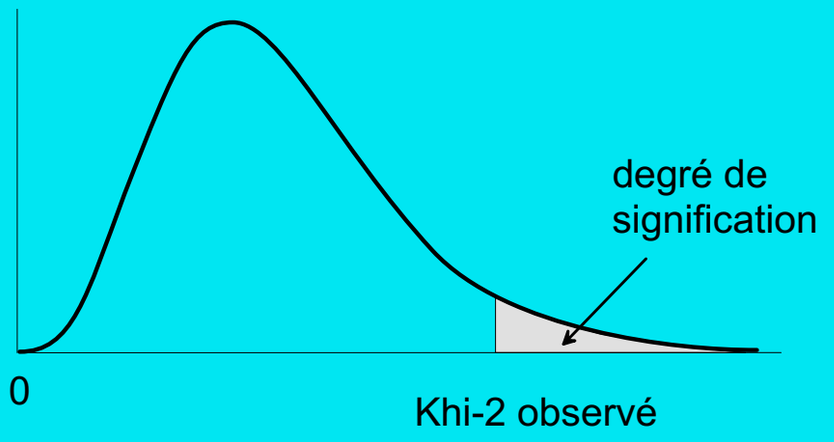
Degré de signification (p-valeur)

Probabilité qu'un échantillon provenant d'une population vérifiant l'hypothèse H_0 (par exemple l'indépendance) donne lieu à une valeur de la statistique de test Q plus extrême que la valeur observée.

$$p(|Q| > |q_0| \mid H_0) \quad q_0 \text{ valeur observée}$$

Degré de signification petit (<5%)
$\Rightarrow$ échantillon exceptionnel sous H_0
$\Rightarrow$ rejet de l'hypothèse H_0.

Exemple : Test d'indépendance,
khi-2 = 8,27, d.l.=3, deg. signif. = 4%



Références

Droesbeke, J.-J. (1992). Eléments de statistique (2ème ed.). Bruxelles: Editions Ellipses.

Ritschard, G. (1989). Introduction à la statistique. Polycopié, Faculté des SES, Genève.

Ritschard, G. (2002). Statistique pour sciences sociales I, transparents du cours. Polycopié, Faculté des SES, Genève.

SPSS Inc. (1992). SPSS Base System Reference Guide, Release 5.0. Chicago, IL: SPSS Inc.

Wonacott, T. H. and R. J. Wonacott (1991). Statistique. Paris: Economica.