2021-2022

M2 Expertise Sciences des Populations

ANALYSE DES DONNÉES LONGITUDINALES

Arno Muller

arno.muller@ined.fr

(Ined - Service Méthodes Statistiques)

Dernière séance

- Des méthodes
  - *Descriptives*
  - *Régression logistique*
  - *Classification des trajectoires*
  - *Modèle de durée*
- Des données
  - *Prospectives*
  - *Rétrospectives*
    - Fiche AGEVEN

Rappel de R

- Sensible à la casse des caractères
- Opérateur d'affectation <- ou =
- Opérateur commentaire #
- Contenu d'un objet édité en tapant son nom
- Une valeur manquante est représentée par **NA**

Les types d’objets

- Un **vecteur**, est un ensemble de données de même mode.
- Un **facteur** est la représentation d'une variable catégorielle.
- Une **matrice** est un tableau à 2 dimensions, tous les éléments étant de même mode
- Un **array** est une généralisation d'une matrice à des tableaux de dimension supérieure à 2
- Un **data frame** est un ensemble de vecteurs ou facteurs, de même longueur, mais de modes différents.
- Une **liste** est un ensemble pouvant contenir n'importe quel type d'objet, y compris des objets de type liste.
- Un **ts** est un ensemble de séries temporelles.

R : Les modes d'objets

- Logique: logical : TRUE or FALSE
- Numérique: numeric, integer
- Facteur: factor
- Caractère: character

Exemples d’opérations sur les vecteurs

- Opérateurs arithmétiques: +,-,*,/
- Opérateurs mathématiques: sqrt, log
- Opérateurs logiques: TRUE, FALSE, & pour « et », | pour « ou »
- Opérateurs statistiques: mean, sum, var

Les data frames

- Caractéristiques:
- Colonnes toutes de même longueur
- Classe particulière de liste
- Stockage idéal pour tableau individus*variables
- Analogue à une table SAS
- Identification des lignes et colonnes
- Fonctions nrow et ncol, rownames et col.names

Utilisation d'un data frame

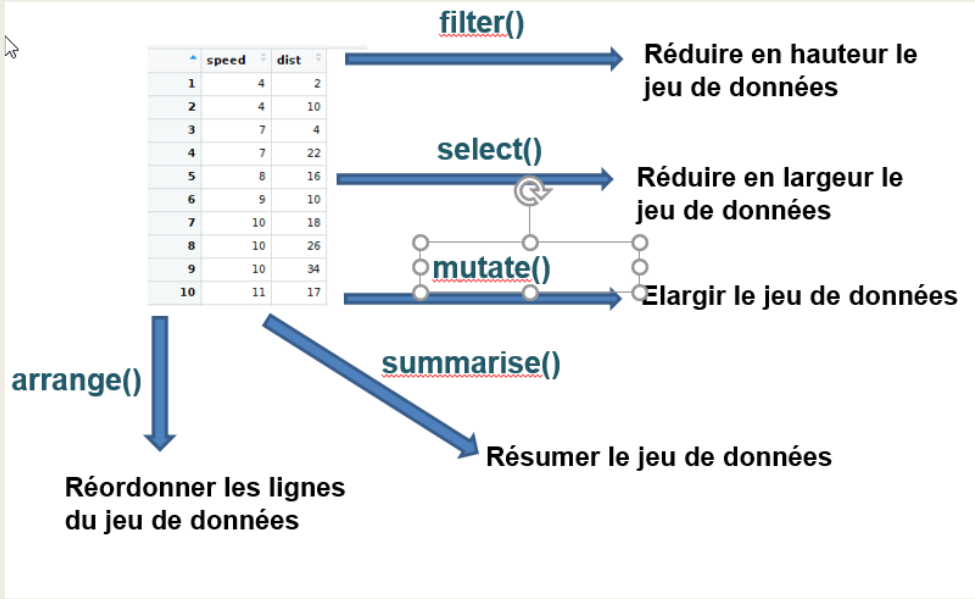
- Accès aux colonnes: *data_frame*$*nom_colonne*
- Sélection de lignes: *data_frame*[*nomcolonne==condition*]
- Sélection d’observations: subset*(data_frame,condition,select=(colonnes))*
- Simplifié par dplyr du tidyverse

Package tidyverse

- Composantes
  - *ggplot2() : Traitements graphiques*
  - *tidyr() : Mise en forme de données*
  - *dplyr() : Gestion de données*
  - *readr() : Importation de données ASCII*
  - *stringr() : Traitement chaînes de caractères*
  - *forcats() : Gestion des facteurs*
  - *readxl() : Importation de données Excel*
  - *purrr()*
  - *tibble()*
  - *lubridate()* : Traitement des dates*

Dplyr : Fonctions de base : 5 verbes

- **select**(): Sélection de colonnes(variables)
  - ***filter****(): Sélection de lignes(observations)*
  - ***arrange****(): Réordonnancement des lignes du dataframe*
  - ***mutate****(): Ajout de colonnes(création de variables)*
  - ***summarise****(): Production de statistiques résumées*



Fonction de base : select()

- select() permet de définir une liste de variables à sélectionner dans le data frame résultant
  - *Liste de variables séparées par des “,” ou « : »*
  - *Facilités d’écriture:*
    - starts_with()
    - ends_with
    - contains()
    - matches()
    - one_of()
    - everything()

Fonction de base : filter()

- filter() permet de définir une liste d'observations sélectionnées en regard de conditions
- Exemples:
  - *B=filter(A,age>30)*
  - *B= filter(A, age >= 20 & age <= 30)*
  - *B=filter(A, qualif %in% c("Ouvrier specialise", "Ouvrier qualifie"), age >= 20 & age <= 30)*

Fonction de base : arrange()

- arrange() permet de réordonner les observations en fonction de clés de tris.
- Remarques :
  - *Une option desc permet de trier par valeurs décroissantes*
  - *Il est possible d'indiquer plusieurs clés d'ordonnancement*
- Exemples:
  - *B=arrange(A,sexe,age)*
  - *B=arrange(A,desc(age))*

Fonction de base : summarise()

- summarise() permet de produire des statistiques à un niveau agrégé . Il faut indiquer le(s) critère(s) d'agrégation et la(les) statistique(s) à calculer.
- Syntaxe:
  - *Une instruction group_by est indispensable pour spécifier en amont les critères d'aggregation*
  - *Quelques exemples de function : count, mean, sum, max*
  - *Il est possible de préciser des conditions dans les calculs(exemple sum(x>10)*
  - *Le résultat peut être stocké au niveau individuel(mutate) ou agrégé(summarise)*
  - *Une option na.rm=T permet d'éliminer les observations pour lesquelles figure une valeur manquante.*
  - *Une instruction ungroup() peut être necessaire pour réinitialiser le calcul de variables individuelles.*

Fonction de base : group_by

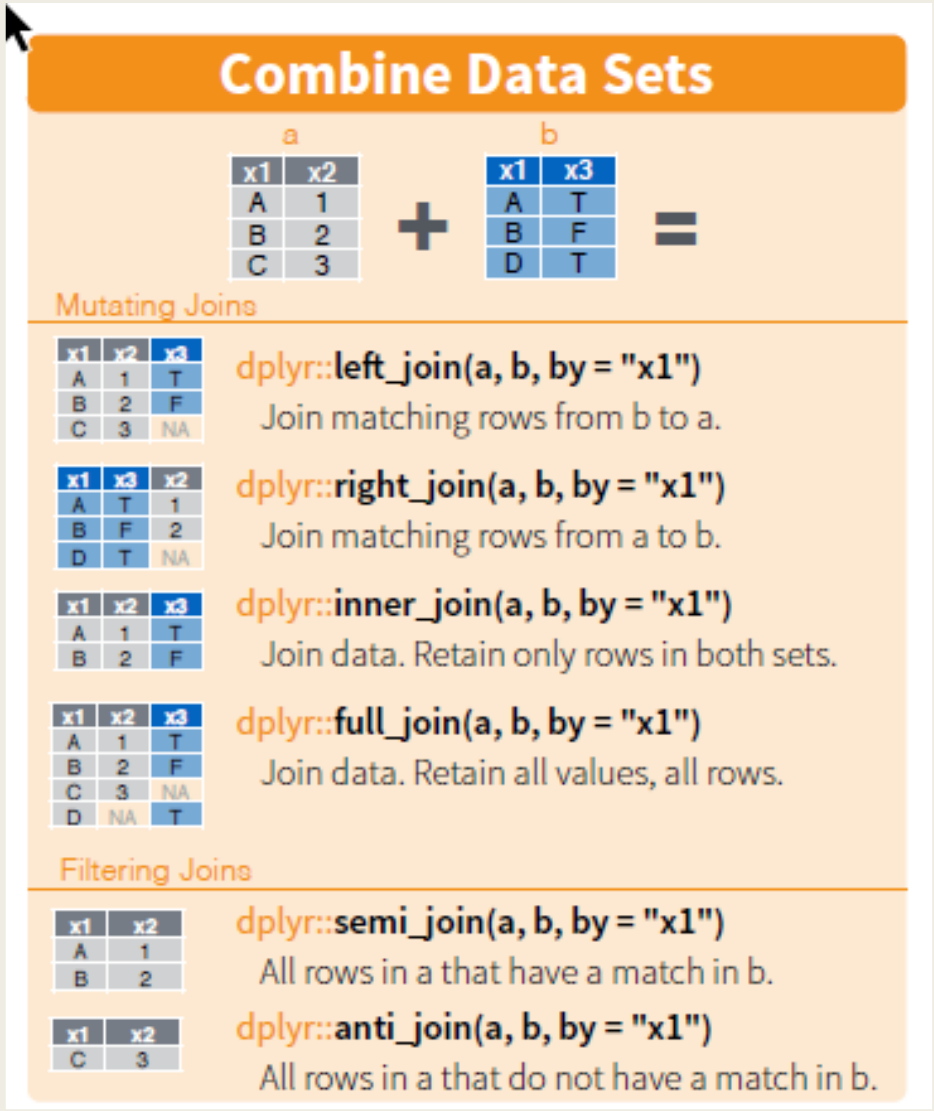
- La fonction group_by s'utilise avec les %>% et permet de définir des groupes de lignes à partir des valeurs d'une ou plusieurs colonnes
- Exemples

```
# creation de variable avec group by:
a2 = a1 %>%
        group_by(sexe) %>%
        mutate(mean_age= mean(age, na.rm = TRUE)) %>%
        select(id,sexe,age,mean_age)
# group_by peut aussi être utile avec filter :
test = a1 %>%
        group_by(sexe) %>%
        filter(age == max(age, na.rm = TRUE))
```

Fonction de base : mutate()

- mutate() permet de créer une variable individuelle en fonction de variables existantes
- # Exemples mutate
  - *b= mutate(a, age_cl = ifelse(age>40,"1","2"))*
  - *b= mutate(a*
  - b= mutate(a, csp_rg= ifelse(as.numeric(qualif) %in% c(1,2), "Ouvrier", "Autre")),*cadre=(qualif=="Cadre")).*

Différents types de jointures



Stockage des données

Stockage des données (1)

- Le stockage des données peut prendre plusieurs formes :
  - *Format large*
  - *Format semi-long*
  - *Format long*

Stockage des données (2) Format large

- Format des fichiers
  - *Format individus « large » : une ligne par individu*

Exemple d'un fichier unions, enquête en 1986

Id	debut1	fin1	Cause rupture1	debut2	fin2	Cause Rupture2
A	1979	1982	Décès cit	1985	.	.
B	1983	1984	Séparation	.	.	.

  - *Inconvénients:*
    Peut générer beaucoup de vecteurs colonnes avec de nombreuses valeurs manquantes

Stockage des données (3) : Format semi-long

- ***Format individus-évènement « semi-long » :*** *une ligne par individu-évènement*

Id	séquence	debut	fin	Cause rupture
A	1	1979	1982	Décès cjt
A	2	1985	.	.
B	1	1983	1984	Séparation

Remarque :

Ce format est souvent celui qui est privilégié dans les mises à disposition de données biographiques.

Stockage des données (4) : Format long

- ***Format individus-période « long » :*** *une ligne par individu- période*

Format notamment pour les analyses en temps discret.

Id	séquence	Année	Cause rupture
A	1	1979	.
A	1	1980	.
A	1	1981	.
A	1	1982	Décès conjoint
A	2	1985	.
A	2	1986	.
B	1	1983	.
B	1	1984	Séparation

Exemple Format semi-long (5.1)

- Enquête « Biographies et entourage »

Base « caractéristiques individuelles »

VIEWTABLE: TMP1.tego

	Identifiant questionniare	prénom d ego	sexe d ego	Date de naissance	Département de naissance	Commune ou pays de naissance	Pays ou DOM-TOM de naissance	Numéro INSEE de la commune de naissance	Nationalité actuelle en clair
1	101	ANDREE	2	06/19/1938	93	LIVRY-GARGAN		46	FRANCAISE
2	102	JEANINE	2	06/11/1934	37	TOURS		261	FRANCAISE
3	103	MANUEL	1	08/20/1942	99	NR	PORTUGAL	99139	PORTUGAISE
4	104	LEON	1	01/13/1933	93	BONDY		10	FRANCAISE
5	105	FRANCOIS	1	12/27/1932	99	ALGER	ALGERIE	99352	FRANCAISE
6	106	EVELYNE	2	11/21/1950	99	NR	ALGERIE	99352	FRANCAISE
7	107	MICHEL	1	05/23/1949	75	PARIS-20E__ARRONDISSEMENT		120	FRANCAISE
8	108	JEANNINE	2	05/21/1948	94	PERREUX-SUR-MARNE		58	FRANCAISE
9	109	BEATRICE	2	06/09/1949	59	LOUVROIL		365	FRANCAISE
10	110	THANH CUA	1	03/16/1941	99	TRAVINH	VIET NAM	99243	FRANCAISE
11	111	MAXIME	1	07/31/1950	77	LAGNY-SUR-MARNE		243	FRANCAISE
12	112	JACQUELINE	2	09/25/1934	54	SAINT-MAX		482	FRANCAISE
13	113	YVETTE	2	09/09/1937	19	CORNIL		61	FRANCAISE
14	114	ZOFIA	2	06/11/1935	99	EMILOWNA	POLOGNE	99122	POLONAISE
15	115	ANTONIO	1	09/19/1932	99	SEVILLE	ESPAGNE	99134	ESPAGNOL
16	116	JEAN PIERRE	1	04/18/1930	75	PARIS-12E__ARRONDISSEMENT		112	FRANCAISE
17	117	JOSETTE	2	04/20/1939	75	PARIS- 6E__ARRONDISSEMENT		106	FRANCAISE
18	118	RADA	2	12/18/1945	99	ZAGREB	YOUGOSLAVIE	99121	CROATE
19	119	JACQUELINE	2	03/23/1933	92	CLICHY		24	FRANCAISE
20	120	CLAUDE	1	09/11/1942	83	TOULON		137	FRANCAISE
21	121	MARIE-NOELLE	2	07/06/1944	21	SEMUR-EN-AUXOIS		603	FRANCAISE
22	122	ROGER	1	12/03/1935	62	ESQUERDES		309	FRANCAISE
23	123	DANIEL	1	06/12/1948	75	PARIS-14E__ARRONDISSEMENT		114	FRANCAISE
24	124	JEAN-CLAUDE	1	08/31/1936	92	NEUILLY-SUR-SEINE		51	FRANCAISE
25	125	GHISLAINE	2	01/20/1944	60	BRETEUIL		104	FRANCAISE
26	126	JOCELYNE	2	06/28/1949	28	BOULLAY-LES-DEUX-EGLISES		53	FRANCAISE
27	127	MARIE-JOSE	2	10/31/1949	76	MONT-SAINT-AIGNAN		451	FRANCAISE

Exemple format semi-long (5.2)

- *Base biographique « logements »*

	Identifiant questionnaire	Age en début de période	Code des événements familiaux	Etape	Département	Liste de communes ou pays ou DOM-TOM	INSEEL3	Type de logement (appartement, maison, ...)	Nombre de pièces dans le logement	Confort sanitaire	Détenteur du statut
1	101	0		1	93	LIVRY-GARGAN	46	21	3	1	P M
2	101	18	M1	2	93	LIVRY-GARGAN	46	22	3	0	2
3	101	23		2M	93	LIVRY-GARGAN	46	22	3	4	2
4	101	49	DCC1	2M	93	LIVRY-GARGAN	46	22	3	4	1
5	102	0		1	37	TOURS	261	12	99	99	P M
6	102	5		2	37	TOURS	261	22	4	1	P M
7	102	7		3T	.		.	.	.	.	
8	102	7		3	37	TOURS	261	12	99	1	P M
9	102	10	NF3	4	75	PARIS-18E_ARRONDISSEMENT	118	41	2	0	P M
10	102	22	M1	5	93	BOBIGNY	8	22	1	1	1 2
11	102	26		6	93	BOBIGNY	8	21	4	4	1 2
12	102	37		7	93	LIVRY-GARGAN	46	21	3	4	1 2
13	103	0		1	99	PORTUGAL	99139	22	2	0	P M
14	103	20		2T	.		.	.	.	.	
15	103	20		2	92	NANTERRE	50	43	1	88	1
16	103	22		3	93	DRANCY	29	43	1	88	1
17	103	24	M1	4	93	LIVRY-GARGAN	46	22	2	2	1
18	103	27		5	93	LIVRY-GARGAN	46	21	3	4	1 2

Exemple format semi-long (5.3)

- Enquête « MAFE » : *Base « caractéristique individuelles »*

ident	q1	q1a	statu_mig	year	age_survey
E1	Man	1972	Migrant	2008	37
E10	Man	1966	Migrant	2008	43
E100	Man	1972	Migrant	2008	37
E101	Woman	1977	Migrant	2008	32
E102	Woman	1966	Migrant	2008	43
E103	Woman	1978	Migrant	2008	31
E104	Woman	1958	Migrant	2008	51
E105	Man	1968	Migrant	2008	41
E106	Man	1961	Migrant	2008	48
E107	Woman	1965	Migrant	2008	44
E108	Man	1972	Migrant	2008	37
E109	Woman	1966	Migrant	2008	43
E11	Man	1979	Migrant	2008	30
E110	Man	1966	Migrant	2008	43
E111	Woman	1983	Migrant	2008	26
E112	Man	1972	Migrant	2008	37
E113	Man	1977	Migrant	2008	32
E114	Man	1964	Migrant	2008	45
E115	Woman	1983	Migrant	2008	26
E116	Man	1951	Migrant	2008	58
E117	Man	1963	Migrant	2008	46
E118	Woman	1965	Migrant	2008	44
E119	Woman	1968	Migrant	2008	41
E12	Woman	1977	Migrant	2008	32
E120	Woman	1973	Migrant	2008	36

Exemple format semi-long (5.4)

- ***Base biographique « logements »***

Remarque:

- Certaines informations de la base caractéristiques individuelles ont été ajoutées à la base biographique.

ident	num_log	q301d	q301f	q302	q303	age_survey	q1a
E1	1	1972	1975	SENEGAL	Namanieque	37	1972
E1	2	1975	2001	SENEGAL	Madina Aly	37	1972
E1	3	2001	2007	SPAIN	Santa Maria De Palautordera	37	1972
E1	4	2007	.	SPAIN	Santa Maria De Palautordera	37	1972
E10	1	1966	1996	SENEGAL	Anambe	43	1966
E10	2	1996	1997	SPAIN	Pineda De Mar	43	1966
E10	3	1997	1999	SPAIN	Granollers	43	1966
E10	4	1999	2006	SPAIN	Figueres	43	1966
E10	5	2006	.	SPAIN	Figueres	43	1966
E100	1	1972	2004	SENEGAL	Dakar	37	1972
E100	2	2004	2007	SENEGAL	Fass / Colobane / Gueule Tapee	37	1972
E100	3	2007	.	SPAIN	Murcia	37	1972
E101	1	1977	1997	SENEGAL	Mandegane	32	1977
E101	2	1997	2006	SENEGAL	Dakar	32	1977
E101	3	2006	2007	SPAIN	Rubi	32	1977
E101	4	2007	.	SPAIN	Rubi	32	1977
E102	1	1966	2005	SENEGAL	Bignona	43	1966
E102	2	2005	.	SPAIN	Mataro	43	1966
E103	1	1978	1992	SENEGAL	Medina Yero	31	1978
E103	2	1992	1995	SPAIN	Calella	31	1978
E103	3	1995	1997	SENEGAL	Medina Yero	31	1978
E103	4	1997	.	SPAIN	Barcelona	31	1978
E104	1	1958	2004	SENEGAL	Dakar	51	1958
E104	2	2004	2007	SPAIN	Salou	51	1958
E104	3	2007	.	SPAIN	Salou	51	1958

Format de fichier semi-long

- Le format de la base d'entrée peut-être en format dit «individus-évènement» en particulier lorsque l'évènement n'est pas intrinsèquement unique, ce qui est souvent le cas dans les sciences sociales. On peut appelé ce format « semi-long ».

- Dans ce cadre, chaque évènement aura un numéro de séquence.
- Si on analyse un évènement selon un rang d'occurence (première union, premier enfant…) il faudra sélectionner cet évènement au préalable.
- Généralement, la durée doit-être calculée, la base présente un point d'entrée et un point de sortie de la séquence.

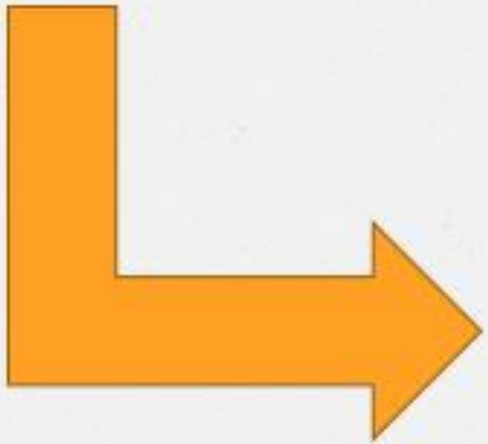
Format « individus – périodes », exemple résidentiel					
id	*n*	*debut*	*fin*	*fin_obs*	*Statut*
A	1	0	5	30	LOC
A	2	15	17	30	LOC
A	3	20	.	30	PRO
B	.	0	.	30	LOC

Transformation format « large » vers un format « long »

- *Pour créer une base en format LONG :*
  - Etape DATA nécessaire pour changer de niveau d'observation
  - Nécessité de créer des variables du moment
  - Nécessité de créer un identifiant à 2 composantes :
    - *Identifiant individuel*
    - *Identifiant temporel*

Passage d'un format large à un format long

id	*t*	*e*	*X*	Z_1	Z_2	Z_3	Z_4	Z_5
A	5	1	1	0	0	1	1	1
B	2	0	0	1	0	.	.	.



Format individus – évènements (durée) « long »

id	*t*	*e*	*X*	*Z*
A	1	0	1	0
A	2	0	1	0
A	3	0	1	1
A	4	0	1	1
A	5	1	1	0
B	1	0	0	1
B	2	0	0	0

Passage d'un format « semi-long » à un format « long »

- **Les trajectoires sont en général continues, ce qui implique par exemple que** le point de départ dans l'occupation d'un logement correspond au point de départ dans la localisation de la trajectoire précédente.

- Exemple

Format « individus – période »				
id	*n*	*debut*	*fin*	*t*
A	1	0	2	0
A	1	0	2	1
A	1	0	2	**2**
A	2	2	4	**2**
A	2	2	4	3
A	2	2	4	4

Format « individus – évènement »			
id	*n*	*debut*	*fin*
A	1	0	2
A	2	2	4

La question des évènements simultanés

- Ce peut être par exemple le cas de durées en couple, dans le cas d'unions polygames. Un homme, à un instant t, peut avoir plusieurs unions.
- Il va falloir « élargir » la base. Le critère du changement de format est donné par le numéro de la séquence.

Format « individus – évènements »

id	*n*	*debut*	*fin*	x
A	1	0	2	0
A	2	1	4	1

Format « individus – période »

id	*n*	*debut*	*fin*	*t*	x
A	1	0	2	0	0
A	1	0	2	**1**	0
A	2	1	4	**1**	1
A	1	0	2	**2**	0
A	2	1	4	**2**	1
A	2	1	4	3	1
A	2	1	4	4	1

Format « individus – périodes » corrigé

id	*t*	x_1	x_2		*X*
A	**0**	0	.		0
A	**1**	0	1		1
A	**2**	0	1		1
A	**3**	.	1		1
A	**4**	.	1		1