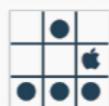


SEMIN'R

INTRODUCTION A L'ANALYSE DE DONNEES

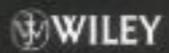
CHRISTOPHE LALANNE



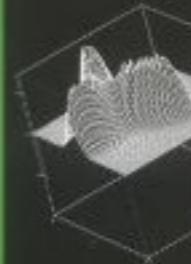
www.aliquote.org

“Let’s not kid ourselves: the most widely used piece of software for statistics is Excel.” — Brian Ripley (2002)

Copyrighted Material



**NUMERICAL ISSUES
IN STATISTICAL
COMPUTING FOR
THE SOCIAL
SCIENTIST**



Micah Altman

Jeff Gill

Michael P. McDonald

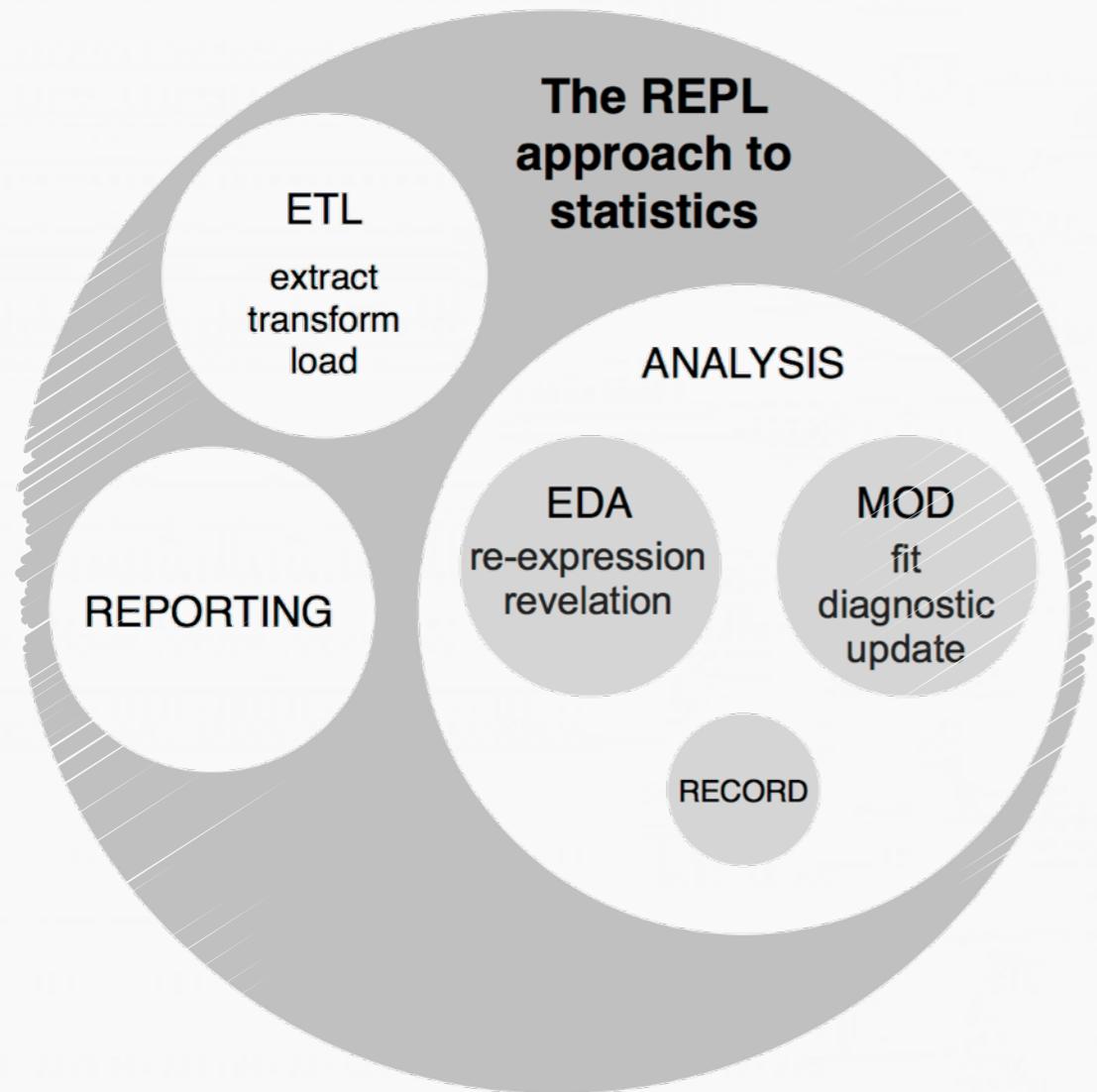
Wiley Series in Probability and Statistics

Copyrighted Material

WWW.
INC. AVAILABLE

ANALYSE EXPLORATOIRE DE DONNEES

UNE APPROCHE INTERACTIVE

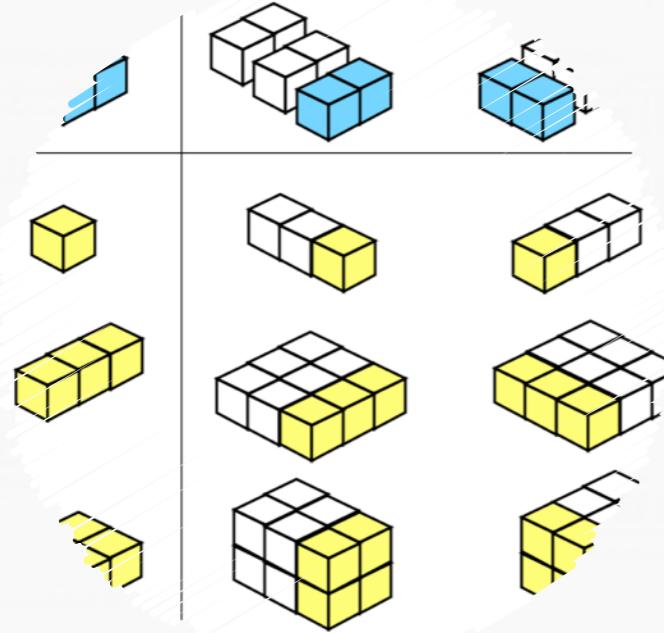


READ-EVAL-PRINT-LOOP

The plural of anecdote is (not) data,
[http://blog.revolutionanalytics.com/
2011/04/the-plural-of-anecdote-is-data-
after-all.html](http://blog.revolutionanalytics.com/2011/04/the-plural-of-anecdote-is-data-after-all.html)

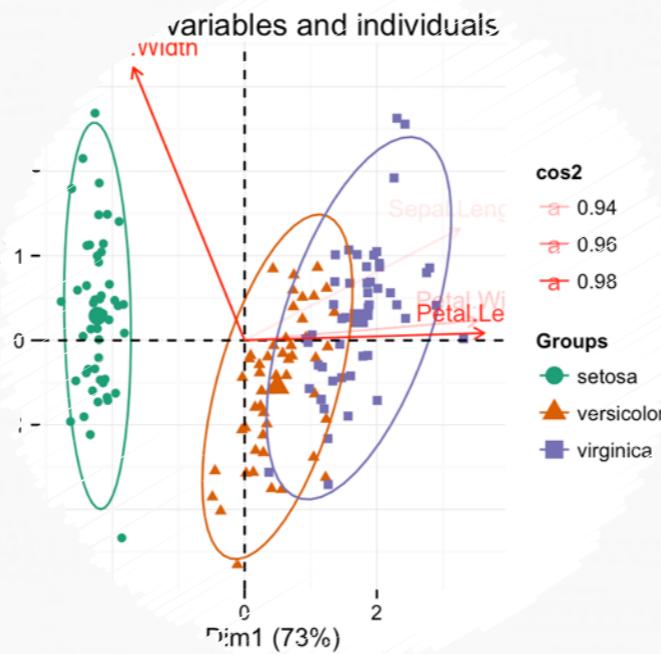
EDA ⇌ MOD

UNE APPROCHE INTERACTIVE



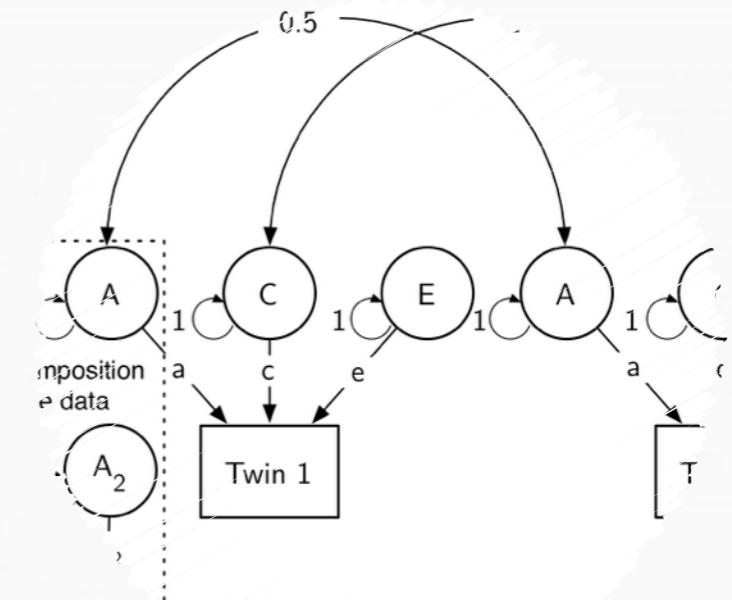
ETL

Extraction de données,
fusion de sources,
nettoyage d'une base de
données, quality control,
recodage de variables,
dictionnaire de données



EDA

Résumés graphiques et
numériques, réduction de
dimension, transformation
de variables, analyse
exploratoire et interactive
des données



MOD

Estimation des
paramètres d'un modèle
statistique, analyse des
résidus, prédiction
ponctuelle et par
intervalles

REPRODUCTIBILITE



70%



30%

PREPARATION DES DONNEES

Nettoyage des données brutes,
recodage des données catégorielles,
création de variables auxiliaires, résumés
numériques univariés, **représentations
graphiques**, représentations graphiques
(encore), agrégation de données,
imputation, etc.

MODELISATION

Modèle statistique, transformation et ré-
expression des variables, qualité
d'ajustement, analyse des résidus, etc.

REPORTING

Peng, R.D. Reproducible research and
biostatistics. *Biostatistics*, 10(3): 405–408, 2009

EDA, TUKEY (1977)

RESISTANCE

Utilisation de techniques peu sensibles aux écarts d'ajustement locaux par rapport au modèle ou aux perturbations (breakdown point, winsorized mean, trimmed mean, etc.)

RESIDUALS

$\text{data} = \text{fit} + \text{residuals}$
Analyse des résidus du modèle afin d'identifier des sources potentielles de déviation ou des observations influentes (outliers ou non)

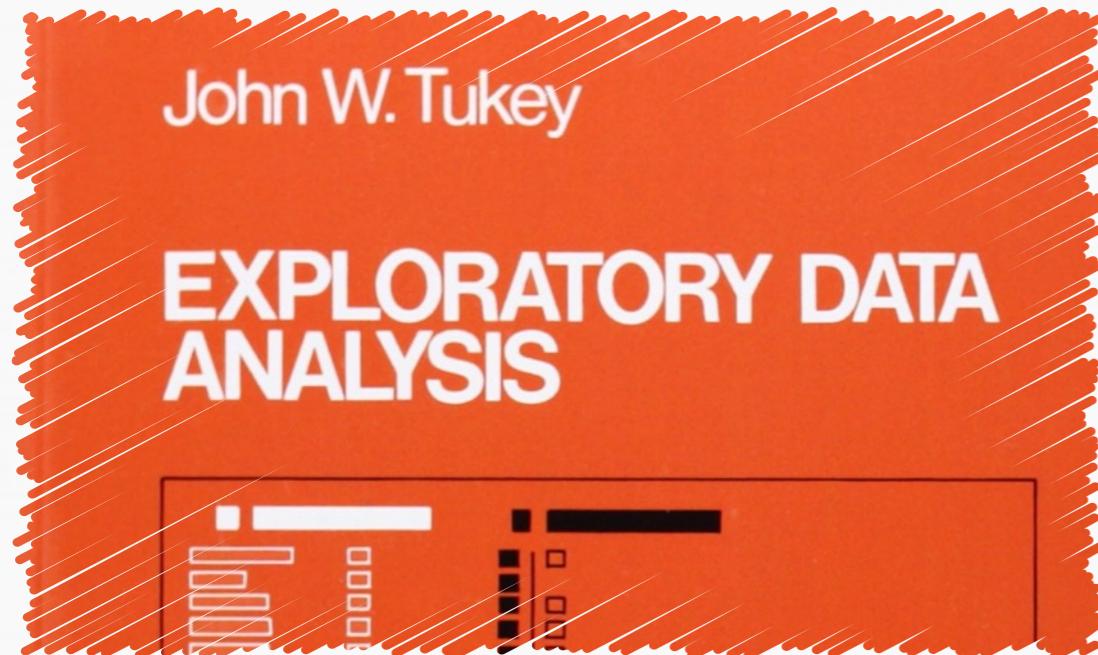
RE-EXPRESSION

Transformation des données (arcsin, logit, z-scores, fréquences relatives, etc.) dans le but de faciliter l'analyse ou l'interprétation des résultats

REVELATION

Utilisation de graphiques « informatifs » (stem-and-leaf plot of effect sizes, dot plot, box plot, etc.) et recherche de patterns inattendus

POUR ALLER PLUS LOIN



REFERENCES

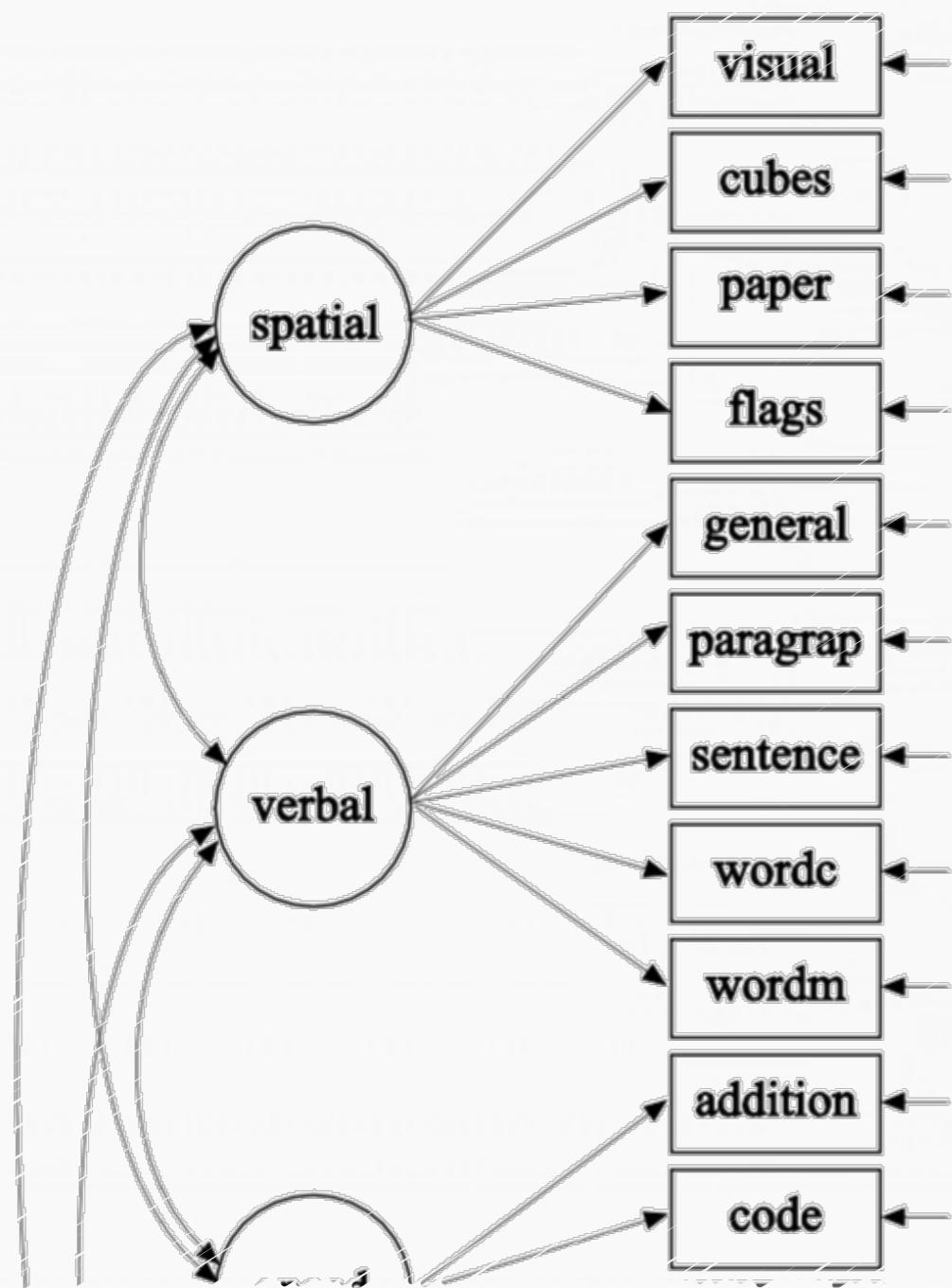
- Hoaglin, D., Mosteller, F. and Tukey, J.W. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, 1985.
- Tufte, E.R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphic Press, 1983.
- Cleveland, W.S. *Visualizing Data*. Summit, NJ: Hobart Press, 1993.

LOGICIELS

- Datadesk (<http://www.datadesk.com>)
- LispStat
- Ggobi (<http://www.ggobi.org>)
- R/cranvas (<http://cranvas.org>)
- R/ggvis (<http://ggvis.rstudio.com>)
- SAS JMP (<http://www.jmp.com>)

MANIPULATION DE DONNEES AVEC R

DONNEES D'ILLUSTRATION



HOLZINGER & SWINEFORD

301 enfants de deux écoles auxquels on a administré 26 tests permettant d'évaluer les compétences suivantes : spatiales, verbales, vitesse de raisonnement, mémoire, mathématiques.

Holzinger, K. J. and Swineford, F. A. A study in factor analysis: The stability of a bi-factor solution. *Supplementary Education Monographs*, 48. University of Chicago, 1939.



lavaan::HolzingerSwineford1939
MBESS::HS.data

LOAD

tests spatiaux

	id	sex	ageyr	agemo	school	grade	visual	cubes	paper	paragrap	sentence	wordm	addition	counting	straight
1	1	1	13	1	Pasteur	7	3.223	7.75	0.375	2.333	5.75	1.286	3.39	5.75	6.36
2	2	2	13	7	Pasteur	7	5.333	5.25	2.125	1.667	3.00	1.286	3.78	6.25	7.92
3	3	2	13	1	Pasteur	7	4.500	5.25	1.875	1.000	1.75	0.429	3.26	3.90	4.42
4	4	1	13	2	Pasteur	7	5.333	7.75	3.000	2.667	4.50	2.429	3.00	5.30	4.86
5	5	2	12	2	Pasteur	7	4.833	4.75	0.875	2.667	4.00	2.571	3.70	6.30	5.92
6	6	2	14	1	Pasteur	7	5.333	5.00	2.250	1.000	3.00	0.857	4.35	6.65	7.50
7	7	1	12	1	Pasteur	7	2.833	6.00	1.000	3.333	6.00	2.857	4.70	6.20	4.86
8	8	2	12	2	Pasteur	7	5.667	6.25	1.875	3.667	4.25	1.286	3.39	5.15	3.67
9	9	2	13	0	Pasteur	7	4.500	5.75	1.500	2.667	5.75	2.714	4.52	4.65	7.36
10	11	2	12	5	Pasteur	7	3.500	5.25	0.750	2.667	5.00	2.571	4.13	4.55	4.36
11	12	1	12	2	Pasteur	7	3.667	5.75	2.000	2.000	3.50	1.571	3.74	5.70	4.31
12	13	1	12	11	Pasteur	7	5.833	6.00	2.875	2.667	4.50	2.714	3.70	5.15	4.14
13	14	2	12	7	Pasteur	7	5.667	4.50	4.125	2.667	4.00	2.286	5.87	5.20	5.86
14	15	2	12	8	Pasteur	7	6.000	5.50	1.750	4.667	4.00	1.571	5.13	4.70	4.44
15	16	1	12	6	Pasteur	7	5.833	5.75	3.625	5.000	5.50	3.000	4.00	4.35	5.86
16	17	2	12	1	Pasteur	7	4.667	4.75	2.375	2.667	4.25	0.714	4.09	3.80	5.14
17	18	2	14	11	Pasteur	7	4.333	4.75	1.500	2.000	4.00	1.286	3.70	6.65	5.25

```

1 data(HolzingerSwineford1939, package="lavaan")
2 HS <- HolzingerSwineford1939
3 names(HS)[7:15] <- c("visual", "cubes", "paper",
4                           "paragrap", "sentence", "wordm",
5                           "addition", "counting", "straight")

```

DATA FRAME

colnames (names)

HS



	id	sex	ageyr	agemo	school	grade	visual	cubes	paper	paragrap	sentence	wordm	addition	counting	straight
1	1	1	13	1	Pasteur	7	3.333	7.75	0.375	2.333	5.75	1.286	3.39	5.75	6.36
2	2	2	13	7	Pasteur	7	5.333	5.25	2.125	1.667	3.00	1.286	3.78	6.25	7.92
3	3	2	13	1	Pasteur	7	4.500	5.25	1.875	1.000	1.75	0.429	3.26	3.90	4.42
4	4	1	13	2	Pasteur	7	5.333	7.75	3.000	2.667	4.50	2.429	3.00	5.30	4.86
5	5	2	12	12	Pasteur	7	4.833	4.75	0.875	2.667	4.00	2.571	3.70	6.30	5.92
6	6	2	14	1	Pasteur	7	5.333	5.00	2.250	1.000	3.00	0.857	4.35	6.65	7.50
7	7	1	12	1	Pasteur	7	2.833	6.00	1.000	3.333	6.00	2.857	4.70	6.20	4.86
8	8	2	12	2	Pasteur	7	5.667	6.25	1.875	3.667	4.25	1.286	3.39	5.15	3.67
9	9	2	13	0	Pasteur	7	4.500	5.75	1.500	2.667	5.75	2.714	4.52	4.65	7.36
10	11	2	12	5	Pasteur	7	3.500	5.25	0.750	2.667	5.00	2.571	4.13	4.55	4.36
11	12	1	12	2	Pasteur	7	3.667	5.75	2.000	2.000	3.50	1.571	3.74	5.70	4.31
12	13	1	12	11	Pasteur	7	5.833	6.00	2.875	2.667	4.50	2.714	3.70	5.15	4.14
13	14	2	12	7	Pasteur	7	5.667	4.50	4.125	2.667	4.00	2.286	5.87	5.20	5.86
14	15	2	12	8	Pasteur	7	6.000	5.50	1.750	4.667	4.00	1.571	5.13	4.70	4.44
15	16	1	12	6	Pasteur	7	5.833	5.75	3.625	5.000	5.50	3.000	4.00	4.35	5.86
16	17	2	12	1	Pasteur	7	4.667	4.75	2.375	2.667	4.25	0.714	4.09	3.80	5.14
17	18	2	14	11	Pasteur	7	4.333	4.75	1.500	2.000	4.00	1.286	3.70	6.65	5.25

rownames



1 HS[5, 3]
2 HS[5, "ageyr"]



1 HS[10, 7:9]
2 HS[10, c("visual", "cubes", "paper")]

INSPECT

The image shows two side-by-side R console windows. The left window displays the output of the `str(HS)` command, which provides a detailed structure of the dataset `HS`. The right window displays the output of the `summary(HS[,7:ncol(HS)])` command, which provides a numerical summary for the variables from the 7th to the last column of the dataset.

Console ~/Desktop/SEMinR/

```
> str(HS)
'data.frame': 301 obs. of 15 variables:
 $ id      : int 1 2 3 4 5 6 7 8 9 11 ...
 $ sex     : int 1 2 2 1 2 2 1 2 2 2 ...
 $ ageyr   : int 13 13 13 13 12 14 12 12 13 12 ...
 $ agemo   : int 1 7 1 2 2 1 1 2 0 5 ...
 $ school  : Factor w/ 2 levels "Grant-White",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ grade   : int 7 7 7 7 7 7 7 7 7 7 ...
 $ visual  : num 3.33 5.33 4.5 5.33 4.83 ...
 $ cubes   : num 7.75 5.25 5.25 7.75 4.75 5 6 6.25 5.75 5.25 ...
 $ paper   : num 0.375 2.125 1.875 3 0.875 ...
 $ paragrap: num 2.33 1.67 1 2.67 2.67 ...
 $ sentence: num 5.75 3 1.75 4.5 4 3 6 4.25 5.75 5 ...
 $ wordm   : num 1.286 1.286 0.429 2.429 2.571 ...
 $ addition: num 3.39 3.78 3.26 3 3.7 ...
 $ counting: num 5.75 6.25 3.9 5.3 6.3 6.65 6.2 5.15 4.65 4.55 ...
 $ straight: num 6.36 7.92 4.42 4.86 5.92 ...
```

Console ~/Desktop/SEMinR/

```
> summary(HS[,7:ncol(HS)])
    visual      cubes      paper      paragrap      sentence
Min.   :0.67   Min.   :2.25   Min.   :0.25   Min.   :0.00   Min.   :1.00
1st Qu.:4.17  1st Qu.:5.25  1st Qu.:1.38  1st Qu.:2.33  1st Qu.:3.50
Median :5.00   Median :6.00   Median :2.12   Median :3.00   Median :4.50
Mean   :4.94   Mean   :6.09   Mean   :2.25   Mean   :3.06   Mean   :4.34
3rd Qu.:5.67  3rd Qu.:6.75  3rd Qu.:3.12  3rd Qu.:3.67  3rd Qu.:5.25
Max.   :8.50   Max.   :9.25   Max.   :4.50   Max.   :6.33   Max.   :7.00
    wordm      addition      counting      straight
Min.   :0.14   Min.   :1.30   Min.   :3.05   Min.   :2.78
1st Qu.:1.43  1st Qu.:3.48  1st Qu.:4.85  1st Qu.:4.75
Median :2.00   Median :4.09   Median :5.50   Median :5.42
Mean   :2.19   Mean   :4.19   Mean   :5.53   Mean   :5.37
3rd Qu.:2.71  3rd Qu.:4.91  3rd Qu.:6.10  3rd Qu.:6.08
Max.   :6.14   Max.   :7.43   Max.   :10.00  Max.   :9.25
```

Deux commandes essentielles :

`str()` : mode de représentation des variables et aperçu des observations

`summary()` : résumé numérique (5-point versus tableau d'effectif)

RECODE

1 = M

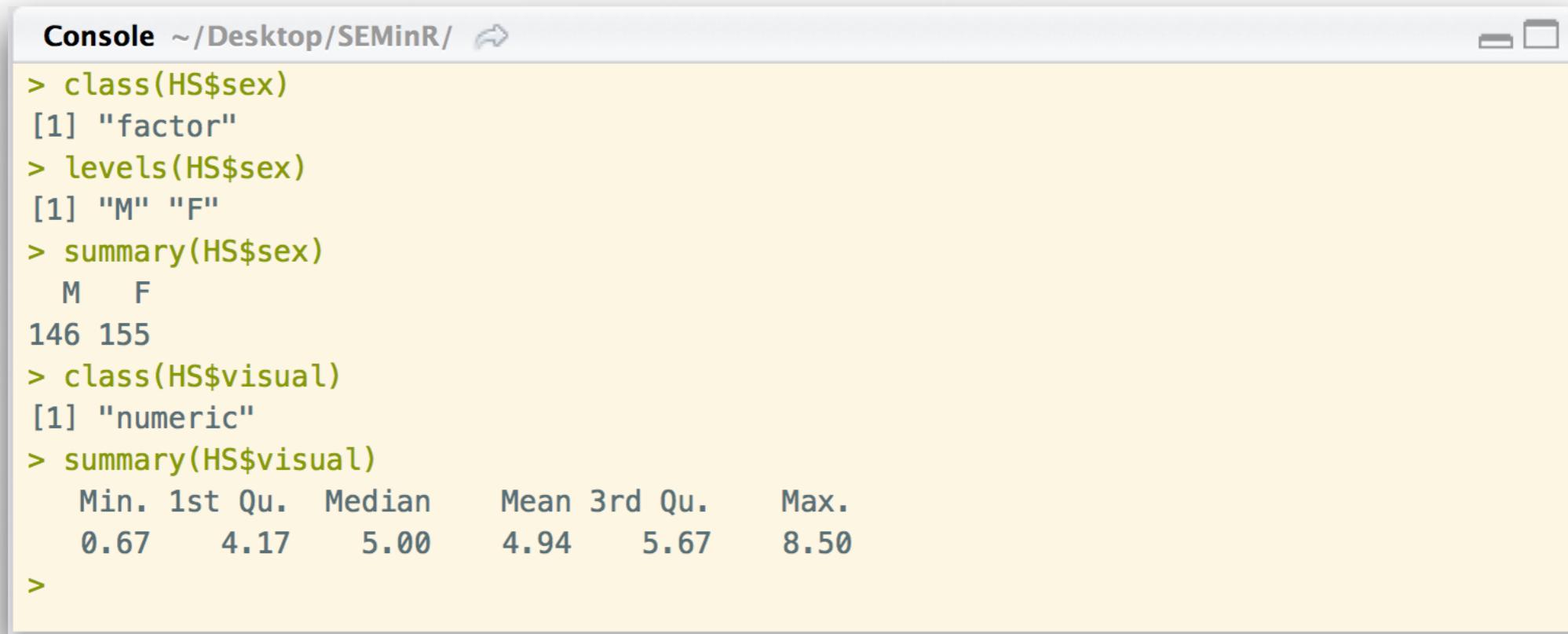
1	1	M	13	1	Pasteur	7	3.333	7.75	0.375	2.333	5.75	1.286	3.39	5.75	6.36
2	2	F	13	7	Pasteur	7	5.333	5.25	2.125	1.667	3.00	1.286	3.78	6.25	7.92
3	3	F	13	1	Pasteur	7	4.500	5.25	1.875	1.000	1.75	0.429	3.26	3.90	4.42
4	4	M	13	2	Pasteur	7	5.333	7.75	3.000	2.667	4.50	2.429	3.00	5.30	4.86
5	5	F	12	2	Pasteur	7	4.833	4.75	0.875	2.667	4.00	2.571	3.70	6.30	5.92
6	6	F	14	1	Pasteur	7	5.333	5.00	2.250	1.000	3.00	0.857	4.35	6.65	7.50
7	7	M	12	1	Pasteur	7	2.833	6.00	1.000	3.333	6.00	2.857	4.70	6.20	4.86
8	8	F	12	2	Pasteur	7	5.667	6.25	1.875	3.667	4.25	1.286	3.39	5.15	3.67
9	9	F	13	0	Pasteur	7	4.500	5.75	1.500	2.667	5.75	2.714	4.52	4.65	7.36
10	11	F	12	5	Pasteur	7	3.500	5.25	0.750	2.667	5.00	2.571	4.13	4.55	4.36
11	12	M	12	2	Pasteur	7	3.667	5.75	2.000	2.000	3.50	1.571	3.74	5.70	4.31
12	13	M	12	11	Pasteur	7	5.833	6.00	2.875	2.667	4.50	2.714	3.70	5.15	4.14
13	14	F	12	7	Pasteur	7	5.667	4.50	4.125	2.667	4.00	2.286	5.87	5.20	5.86
14	15	F	12	8	Pasteur	7	6.000	5.50	1.750	4.667	4.00	1.571	5.13	4.70	4.44
15	16	M	12	6	Pasteur	7	5.833	5.75	3.625	5.000	5.50	3.000	4.00	4.35	5.86
16	17	F	12	1	Pasteur	7	4.667	4.75	2.375	2.667	4.25	0.714	4.09	3.80	5.14
17	18	F	14	11	Pasteur	7	4.333	4.75	1.500	2.000	4.00	1.286	3.70	6.65	5.25

```

1 HS <- within(HS, {
2   id <- factor(id)
3   sex <- factor(sex, levels = c(1, 2), labels = c("M", "F"))
4   grade <- factor(grade)
5 })

```

CHIFFRES ET LETTRES



The screenshot shows an R console window with the following content:

```
Console ~/Desktop/SEMinR/ 
> class(HS$sex)
[1] "factor"
> levels(HS$sex)
[1] "M" "F"
> summary(HS$sex)
   M   F
146 155
> class(HS$visual)
[1] "numeric"
> summary(HS$visual)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.
   0.67    4.17   5.00    4.94    5.67    8.50
>
```

Deux types d'objets sous R : les **nombres** (entiers et flottants) et les **facteurs**.

Autre type : chaîne de **caractères** (character).

LES FACTEURS



The screenshot shows an R console window with the following output:

```
Console ~/Desktop/SEMinR/ ↵
> head(HS$sex)
[1] M F F M F F
Levels: M F
> head(as.numeric(HS$sex))
[1] 1 2 2 1 2 2
>
```

(...) et les **facteurs**. Pour le meilleur et pour le moins pratique parfois...

SUBSET

HS

id	sex	ageyr	agemo	school	grade
1	M	13	1	Pasteur	7
2	F	13	7	Pasteur	7
3	F	13	1	Pasteur	7
4	M	13	2	Pasteur	7
5	F	12	2	Pasteur	7
6	F	14	1	Pasteur	7
7	M	12	1	Pasteur	7
8	F	12	2	Pasteur	7
9	F	13	0	Pasteur	7
11	F	12	5	Pasteur	7
12	M	12	2	Pasteur	7
13	M	12	11	Pasteur	7
14	F	12	7	Pasteur	7
15	F	12	8	Pasteur	7
16	M	12	6	Pasteur	7
17	F	12	1	Pasteur	7
18	F	14	11	Pasteur	7

BASE::SUBSET

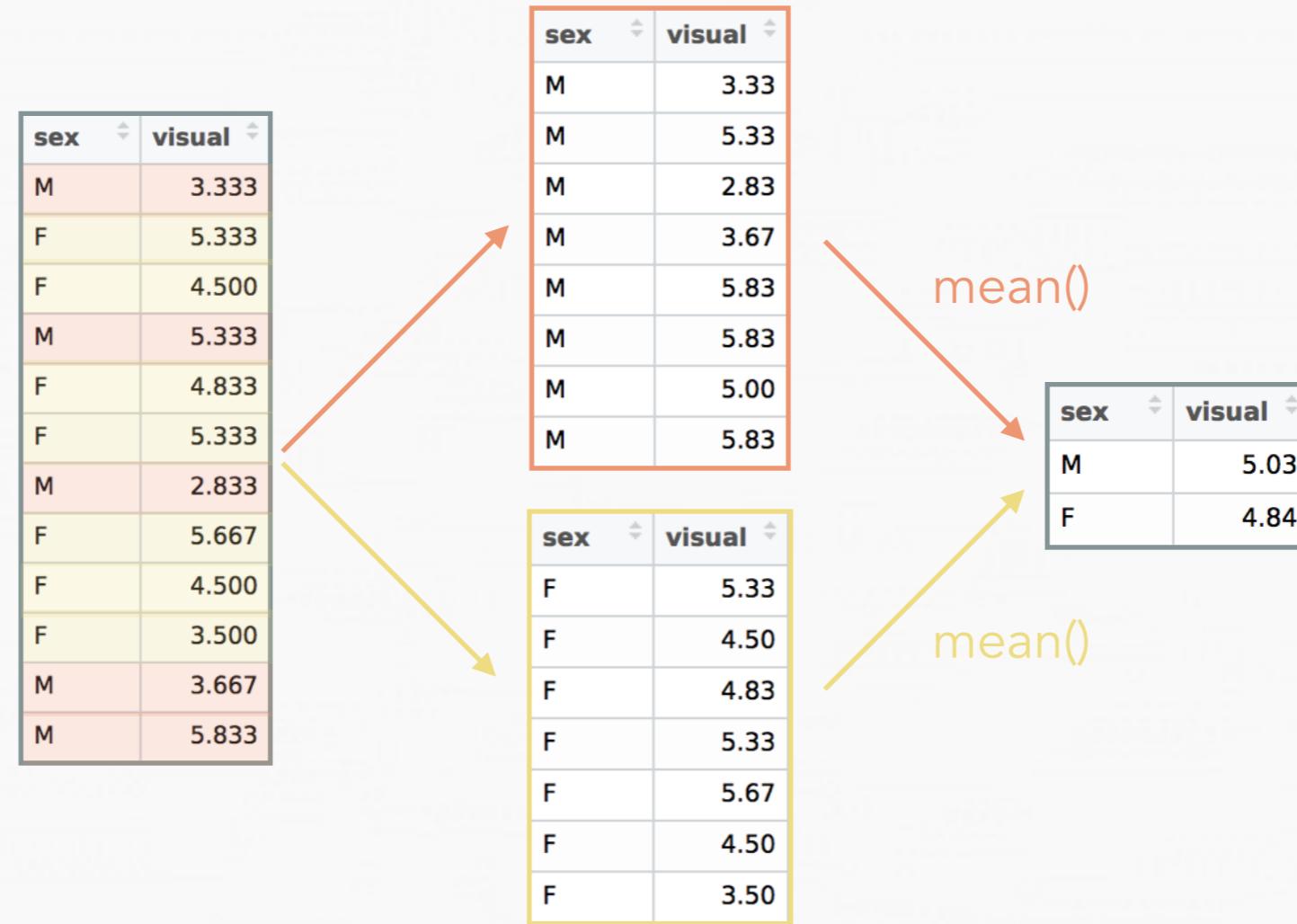
`subset(data, subset, select)`

`subset` sélection des observations (lignes)
sur critère(s) logique(s)

`select` sélection des variables (colonnes)
selon le nom ou l'index

1 `subset(HS, subset = sex == "M")`
2 `subset(HS, select = c(ageyr, agemo))`

SPLIT - APPLY - COMBINE



(...) break up a big problem into manageable pieces, operate on each piece independently and then put all the pieces back together. – Hadley Wickham (2011)

AGGREGATE

HS

sex	ageyr	agemo	school	grade	visual
M	13	1	Pasteur	7	3.333
F	13	7	Pasteur	7	5.333
F	13	1	Pasteur	7	4.500
M	13	2	Pasteur	7	5.333
F	12	2	Pasteur	7	4.833
F	14	1	Pasteur	7	5.333
M	12	1	Pasteur	7	2.833
F	12	2	Pasteur	7	5.667
F	13	0	Pasteur	7	4.500
F	12	5	Pasteur	7	3.500
M	12	2	Pasteur	7	3.667
M	12	11	Pasteur	7	5.833
F	12	7	Pasteur	7	5.667
F	12	8	Pasteur	7	6.000
M	12	6	Pasteur	7	5.833
F	12	1	Pasteur	7	4.667
F	14	11	Pasteur	7	4.333

BASE::AGGREGATE

aggregate(formula, data, function)

formula= variable réponse ~ variable explicative

function= mean, median, sd, summary, etc.

- 1 aggregate(visual ~ sex, data = HS, mean)
- 2 aggregate(visual ~ sex, data = HS, sd)

APPLICATION

```
Console ~/Desktop/SEMinR/ ⌂
> aggregate(visual ~ sex + grade, HS, mean)
  sex grade visual
 1   M    7  4.86
 2   F    7  4.64
 3   M    8  5.22
 4   F    8  5.07
> aggregate(cbind(visual,cubes,paper) ~ sex, HS, mean)
  sex visual cubes paper
 1   M    5.03  6.23  2.46
 2   F    4.84  5.95  2.05
>
```

`visual ~ sex + grade` : 1 variable réponse + 2 variables explicatives

`cbind(visual,cubes,paper) ~ sex` : 3 variables réponses + 1 variable explicative

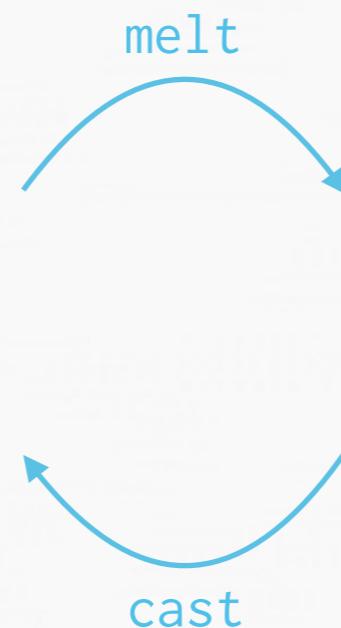
RÉSULTAT = UN DATA FRAME

LES FORMULES R

- ✓ Un moyen simple et élégant d'exprimer la relation entre une ou plusieurs variable(s) réponse(s) et une ou plusieurs variable(s) explicative(s) (Wilkinson & Rogers, 1973 ; Chambers & Hastie, 1992).
- ✓ Idée reprise dans Python (pandas) et Julia (DataFrames).
- ✓ Structuration particulière des données (Wickham, 2014) : package reshape2.

id	sex	visual	cubes	paper
1	M	3.333	7.75	0.375
2	F	5.333	5.25	2.125
3	F	4.500	5.25	1.875
4	M	5.333	7.75	3.000
5	F	4.833	4.75	0.875
6	F	5.333	5.00	2.250
7	M	2.833	6.00	1.000
8	F	5.667	6.25	1.875
9	F	4.500	5.75	1.500
11	F	3.500	5.25	0.750
12	M	3.667	5.75	2.000
13	M	5.833	6.00	2.875
14	F	5.667	4.50	4.125

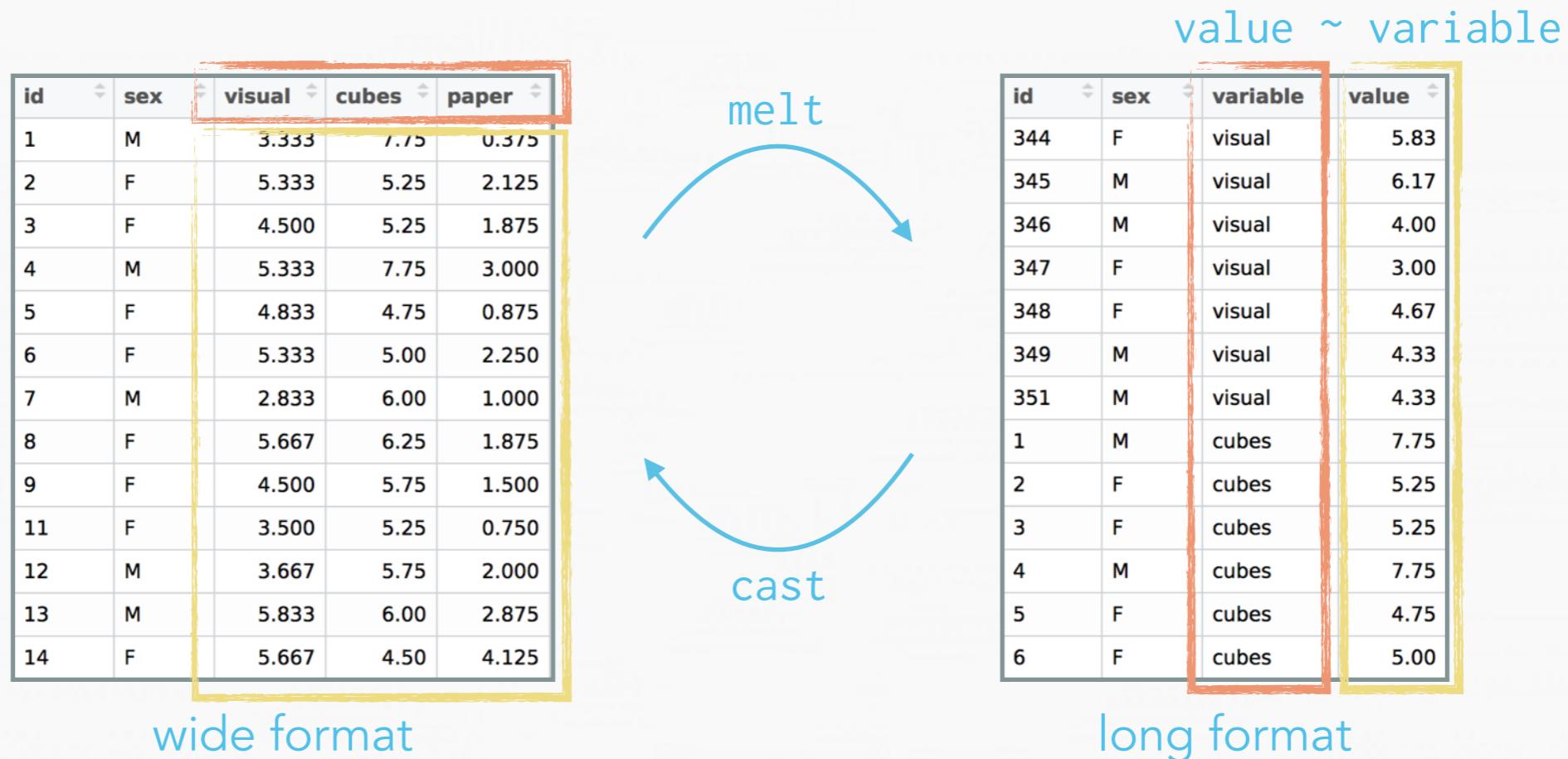
wide format



id	sex	variable	value
344	F	visual	5.83
345	M	visual	6.17
346	M	visual	4.00
347	F	visual	3.00
348	F	visual	4.67
349	M	visual	4.33
351	M	visual	4.33
1	M	cubes	7.75
2	F	cubes	5.25
3	F	cubes	5.25
4	M	cubes	7.75
5	F	cubes	4.75
6	F	cubes	5.00

long format

RESHAPING



```
1 library(reshape2)
2
3 dw <- HS[,c("id","sex","visual","cubes","paper")]
4
5 dl <- melt(dw)                                     ## long
6 dw <- dcast(dl, sex + id ~ variable)             ## wide
```

APPLICATION

```
Console ~/Desktop/SEMinR/ ⌂
> aggregate(cbind(visual,cubes,paper) ~ sex, HS, mean)
  sex visual cubes paper
1   M    5.03  6.23  2.46
2   F    4.84  5.95  2.05
> aggregate(value ~ variable + sex, dl, mean)
  variable sex value
1   visual   M  5.03
2   cubes   M  6.23
3   paper   M  2.46
4   visual   F  4.84
5   cubes   F  5.95
6   paper   F  2.05
> |
```

value ~ variable + ...

APPLY

id	sex	visual	cubes	paper
1	M	3.333	7.75	0.375
2	F	5.333	5.25	2.125
3	F	4.500	5.25	1.875
4	M	5.333	7.75	3.000
5	F	4.833	4.75	0.875
6	F	5.333	5.00	2.250
7	M	2.833	6.00	1.000
8	F	5.667	6.25	1.875
9	F	4.500	5.75	1.500
11	F	3.500	5.25	0.750
12	M	3.667	5.75	2.000
13	M	5.833	6.00	2.875
14	F	5.667	4.50	4.125

BASE::APPLY

apply(data, margin, function)

margin=1 opération par lignes

margin=2 opération par colonnes

function mean, median, sd, summary, etc.

```
1 apply(HS[,c("visual", "cubes", "paper")], 1, sum)
2 apply(HS[,c("visual", "cubes", "paper")], 2, mean)
```

APPLICATION



The screenshot shows an R console window titled "Console ~/Desktop/SEMinR/". The console displays the following R code:

```
> names(HS)
[1] "id"         "sex"        "ageyr"      "agemo"      "school"     "grade"      "visual"
[8] "cubes"      "paper"      "paragrap"   "sentence"   "wordm"      "addition"   "counting"
[15] "straight"
> spatial <- apply(HS[,c("visual","cubes","paper")], 1, sum)
> verbal <- apply(HS[,c("paragrap","sentence","wordm")], 1, sum)
> speed <- apply(HS[,c("addition","counting","straight")], 1, sum)
> |
```

Les variables spatial, verbal et speed sont créées dans l'espace de travail mais n'appartiennent pas au data frame HS.

APPLICATION

```
Console ~/Desktop/SEMinR/ ⌂
[1] "id"        "sex"       "ageyr"     "agemo"     "school"    "grade"     "visual"
[8] "cubes"     "paper"     "paragrap"   "sentence"   "wordm"     "addition"  "counting"
[15] "straight"
> spatial <- apply(HS[,c("visual","cubes","paper")], 1, sum)
> verbal <- apply(HS[,c("paragrap","sentence","wordm")], 1, sum)
> speed <- apply(HS[,c("addition","counting","straight")], 1, sum)
> HS$spatial <- apply(HS[,c("visual","cubes","paper")], 1, sum)
> HS$verbal <- apply(HS[,c("paragrap","sentence","wordm")], 1, sum)
> HS$speed <- apply(HS[,c("addition","counting","straight")], 1, sum)
> apply(HS[,c("spatial","verbal","speed")], 2, mean)
spatial  verbal   speed
  13.27    9.59   15.09
>
```

En intégrant ces variables auxiliaires au data frame, on préserve la cohérence des données.

TESTS STATISTIQUES

PROCÉDURES DE TEST

TEST DE STUDENT

```
t.test(visual ~ sex, data=HS, var.equal=TRUE)
```

Test de Student avec l'option var.equal = TRUE,
test de Welch par défaut.

Alternative non paramétrique : wilcox.test().

ANALYSE DE VARIANCE

```
summary(aov(visual ~ sex, data=HS))
```

ANOVA avec sommes de carrés de type II. Voir
également drop1() et car::Anova().

Alternative non paramétrique : kruskal.test().

TEST DE CORRELATION

```
cor.test(~ visual+cubes, data=HS, method="pear")
```

Test de nullité du coefficient de corrélation de
Pearson (défaut) ou de Spearman
(method="spear").

TEST DE PROPORTION

```
summary(xtabs(~ sex+grade, data=HS))  
prop.test(xtabs(~ sex+grade, data=HS),  
          correct=FALSE)
```

Test du χ^2 de Pearson et test de proportion
utilisant l'approximation par la loi normale.
Alternatives : fisher.test() et binom.test().

APPLICATION

```
Console ~/Desktop/SEMinR/ 
> aggregate(visual ~ sex, data = HS, sd)
  sex visual
 1   1   1.12
 2   2   1.21
> t.test(visual ~ sex, data = HS, var.equal = TRUE)

  Two Sample t-test

data: visual by sex
t = 1, df = 300, p-value = 0.2
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.0755  0.4535
sample estimates:
mean in group 1 mean in group 2
      5.03        4.84
aggregate(visual ~ sex, HS, mean)
> |
```

Conditions d'application du test : normalité, égalité des variances parentes, indépendance (cas des échantillons appariés, paired = TRUE).

APPLICATION

```
Console ~/Desktop/SEMinR/ 
> xtabs(~ sex + grade, data = HS)
  grade
sex 7 8
  1 74 71
  2 83 72
> prop.table(xtabs(~ sex + grade, data = HS), margin = 2)
  grade
sex    7     8
  1 0.471 0.497      74/(74+83)
  2 0.529 0.503
> summary(xtabs(~ sex + grade, data = HS))
Call: xtabs(formula = ~sex + grade, data = HS)
Number of cases in table: 300
Number of factors: 2
Test for independence of all factors:
  Chisq = 0.19, df = 1, p-value = 0.7
> |
```

Conditions d'application du test : effectifs théoriques « pas trop petits ».

Voir chisq.test() (\$expected et \$residuals).

GESTION DES SCRIPTS R

load.r

```
1 ## load.r
2 ## Chargement des données Holzinger & Swineford (1939)
3 ##
4
5 options(digits = 3)
6
7 ## importation des données
8 data(HolzingerSwineford1939, package="lavaan")
9 HS <- HolzingerSwineford1939
10 names(HS)[7:15] <- c("visual", "cubes", "paper",
11                         "paragrap", "sentence", "wordm",
12                         "addition", "counting", "straight")
13
14 head(HS)
15
16 ## indexation d'observations
17 HS[5, 3]
18 HS[5, "ageyr"]
19
20 HS[10, 7:9]
21 HS[10,c("visual", "cubes", "paper")]
22
23 ## recodage des variables catégorielles
24 HS <- within(HS, {
25   id <- factor(id)
26   sex <- factor(sex, levels = c(1, 2), labels = c("M", "F"))
27   grade <- factor(grade)
28 })
```

GESTION DES SCRIPTS R

aggregate.r

```
1 ## aggregate.r
2 ## Statistiques descriptives
3 ##
4 ## load.r : chargement des données
5
6 ## sélection de sous-ensemble d'observations
7 subset(HS, subset = sex == "M")
8 subset(HS, select = c(ageyr, agemo))
9
10 ## calcul de moyennes/ety conditionnels
11 aggregate(visual ~ sex, data = HS, mean)
12 aggregate(visual ~ sex, data = HS, sd)
13
14 aggregate(visual ~ sex + grade, HS, mean)
15 aggregate(cbind(visual,cubes,paper) ~ sex, HS, mean)
16
17 ## passage du format large au format long
18 library(reshape2)
19
20 dw <- HS[,c("id","sex","visual","cubes","paper")]
21
22 dl <- melt(dw)                                     ## long
23 dw <- dcast(dl, sex + id ~ variable)             ## wide
24
25 ## opérations par lignes/colonnes
26 apply(HS[,c("visual","cubes","paper")], 1, sum)
27 apply(HS[,c("visual","cubes","paper")], 2, mean)
```

REFERENCES

1. Peng, R.D. Reproducible Research And Biostatistics. *Biostatistics*, 10(3): 405–408, 2009.
2. Tufte, E.R. *Exploratory Data Analysis*. Addison-Wesley, 1977.
3. Hoaglin, D., Mosteller, F. and Tukey, J.W. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, 1985.
4. Tufte, E.R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphic Press, 1983.
5. Cleveland, W.S. *Visualizing Data*. Summit, NJ: Hobart Press, 1993.
6. Holzinger, K. J. and Swineford, F. A. A study in factor analysis: The stability of a bi-factor solution. *Supplementary Education Monographs*, 48. University of Chicago, 1939.
7. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40: 1–29, 2011.
8. Wickham, H. Tidy Data. *Journal of Statistical Software*, 59: 1–23, 2014.
9. Wilkinson, G. and Rogers, C. Symbolic Description Of Factorial Models For Analysis Of Variance. *Applied Statistics*, 22: 392–399, 1973.
10. Chambers, J. and Hastie, T. *Statistical Models in S*. Wadsworth & Brooks, 1992.