Initiation à « R » pour les épidémiologistes débutants - Analyser des données en 30 minutes…

Bernard Branger, Janvier 2026, NANTES

branger44@gmail.cm – Tél 06 32 70 33 80

Site Epi-Σûre : https://www.episure.fr/

1. Télécharger le logiciel « R »

R est **gratuit** et disponible sur tout type d'ordinateur. Aller dans votre navigateur à l'adresse suivante : http://cran.r-project.org/

Si votre ordinateur est un PC cliquez sur :

1) « Download R for Windows », puis « base »
3) « R version 4.5.2 » *a priori pour Windows 64 bits (sinon voir dans Windows/ Système/ Type du Système 32 ou 64 bits)*

Si votre ordinateur est un Mac cliquez sur :

4) "Download R" for "MacOs X"
5) R 4.5.1 "Great Square Root" released on 2025/06/13

Une fois le téléchargement terminé, lancez l'application en cliquant sur « suivant » à chaque fois.

Le **Logiciel "R"®** fonctionne avec une **ligne de commandes** écrites à la main. Il peut importer et utiliser de nombreux « packages » couvrant tous les champs de l'épidémiologie et de la statistique comme *epitools* ou *epiDisplay*. Il nécessite un long apprentissage et semble rebutant dans les débuts en raison des commandes peu intuitives (il est fait, à la base, par des matheux, pas des épidémiologistes). Cependant, on peut s'aider de « **R Studio** » (à télécharger, voir point 2), ou « **R Commander** » (voir supplément) avec menus et clics possibles.

Il existe de nombreuses aides sur internet ou avec l'intelligence artificielle. Il « peut tout faire » en particulier pour l'analyse multivariée et les graphes+++. Possibilité d'import de tout type de fichier. Voir les sites https://www.r-project.org/ ou https://r.developpez.com/tutoriels/r/introduction/.

Note sur les mots employés dans R et RStudio

Bien distinguer les variables qualitatives (factor) des variables quantitatives (int) ++

- Un fichier ou une base ou un tableau de données s'appelle une ***data frame*** *(df dans les tuto) ou* ***tibble***
- Une variable en général s'appelle ***vector***
- Une variable qualitative (ou catégorielle ou codée) s'appelle ***factor***
- Une variable quantitative est dite ***int*** (pour *integer*, chiffres entiers) ou ***num*** (pour *numeric)* ou ***dbl*** (chiffres avec décimales),
- Une variable écrite en texte est dite ***character***,
- Une variable logique booléenne (oui/non) est dite ***logi*** (pour logique qui se nomme TRUE ou FALSE).
- Les importations des chiffres/nombres se font généralement en numérique ; des commandes existent pour pouvoir les transformer en *factor* si besoin.
- L'ensemble des fichiers d'une analyse épidémiologique (pouvant comprendre des fichiers de données en *.xlsx ou en *.RData ou *.rda, ou des fichiers de script en *.R) s'appelle un ***project*** disponible dans un dossier unique.

2. Télécharger "Rstudio"

Il faut aller à l'adresse suivante : Site https://posit.co/download/rstudio-desktop/ . Cliquez sur « donwload now », puis « download Rstudio deskstop»; la version utilisée est (2025.09.1 Build 418). Choisissez la version à installer "Mac" ou "PC" ou la version choisie de "Linux". Une fois le téléchargement terminé, lancez l'application en cliquant sur « suivant » à chaque fois.

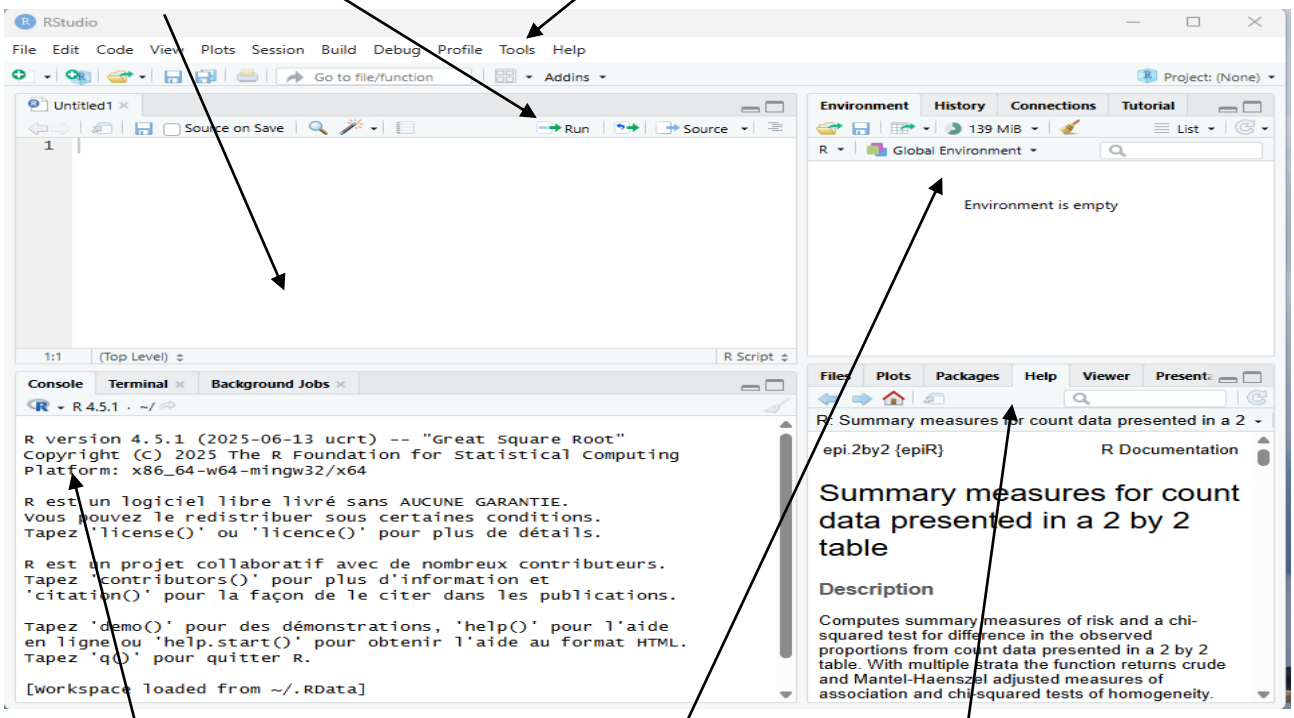
R Studio est un logiciel qui simplifie l'utilisation de « R ». Il existe aussi « R Commander » qui se dédouane d'une ligne de commande et permet d'utiliser la souris et de cliquer. Il n'y pas de statistiques avancées et il faut revenir à un moment à « R Studio ». Voir le document "Supplément" sur le site www.episure.fr

3. Prise en main

Ordonne les 4 fenêtres : *Tools / Show All Panes*

a. Ouvrir « R Studio »

- **Editeur de texte** (fenêtre d'édition ou de **script**) : on peut écrire les commandes et on les lance avec *Run.* On peut sauver le programme pour le re-lancer plus tard (format de fichier en *.R). On peut aussi afficher le tableur de données. **S'il n'est pas visible, cliquer *Files/ New Files / R Script, ou importer un fichier déjà constitué en *.R.***



Console R : **Ecrire directement des commandes en commençant par >, (et faire ↵),** et les résultats s'affichent par incrémentation. On peut sauver les résultats.

- **Fichiers et variables en cours, et historique des commandes**
- **Gestion** des graphiques et des « packages » (programmes clés en main, voir infra), ainsi que des aides sur les commandes (Help) (voir infra), et l'affichage des graphes (Plots).
  - ⇨ ***Touche Ctrl + L (Windows) ou Cmd+Option+ L (Mac) pour effacer le contenu de la console***
  - ⇨ ***Commande remove () ou rm () pour supprimer des éléments***

b. Importer un fichier de données (*.txt, *.csv, *.xls(x), ou tout autre format….

= Copier les fichiers d'Excel en *.xlsx (à dézipper) du site Internet de Episure : https://www.episure.fr/formations, cliquer sur ***Fichiers sur les déclenchements à télécharger (fichier ZIP)*** vers un endroit, sur votre ordi, que vous retrouverez facilement (Bureau par exemple) et le **dézipper** dans un endroit également facilement trouvable.

= Pour importer un fichier dans dans RStudio : Onglet ***"File" / Import Dataset / From Excel*** (on peut aussi choisir d'autres formats de fichier tels que fichiers en *.txt ou SPSS, ou Stata ou SAS…). On peut ouvrir aussi **un fichier déjà sauvé en *. rda ou *. RData sur l'ordinateur.**

= Une fenêtre s'affiche : **Importer le fichier** « **declenche_2025.xlsx** » du lieu où vous l'avez mis, puis s'assurer que la case est cochée : ☒ *First Row as Names* pour garder les noms de variables du fichier primaire. **Voir le document supplémentaire sur les imports avec une ligne de commande.** *Ne rien mettre dans la case en face de "NA".*

c. Les fichiers d'exemple : trois fichiers *.xlsx à choisir selon le cas

Ce sont de vraies données concernant des **déclenchements d'accouchements pour des motifs maternels ou fœtaux**. Certains accouchements se terminent par "voie basse" (VB) (voies naturelles, dites souvent « succès ») ou par césarienne (CS) (dite souvent « échec »). Ce ne sont que des singletons.

L'objectif de l'étude est de déterminer les facteurs associés ou causaux qui conduisent à la césarienne. Il y a 903 observations et 59 variables. Les résultats ont été publiés (en français). Voir *Branger B, Dochez V, Gervier S, Winer N. [Cesarean after labor induction: Risk factors and prediction score]. Césarienne après déclenchement : facteurs de risques et score de prédiction. Gynecol Obstet Fertil Senol. 2018 May;46(5):458-465. doi: 10.1016/j.gofs.2018.03.008.*

d. Voir si les données sont bien importées ("fouille des données", "data mining")

Dans la console : commandes à utiliser pour vérifier : `dim, head, str,`

```
> head (declenche_2025)
# A tibble: 6 × 59
  fiche issue dateacc              hacc  macc hmacc geste parite
  <dbl>  <dbl> <dttm>              <dbl> <dbl> <dbl> <dbl>  <dbl>
1   641      3 2010-11-23 00:00:00    18    31 18.5     NA      1
2   125      3 2010-02-18 00:00:00    18    16 18.3     NA      1
3   171      3 2010-03-06 00:00:00    21    37 21.6     NA      1
4    13      2 2010-01-07 00:00:00     2    40  2.67     1      1
5   239      3 2010-04-02 00:00:00    16    17 16.3     NA      2
6   219      3 2010-03-24 00:00:00    21    22 21.4     NA      1
# ℹ 51 more variables: agemat <dbl>, agsem <dbl>, agj <dbl>,
#   ag <dbl>, modedecl <dbl>, motifdecl <dbl>, motiffoet <dbl>,
#   motifmat <dbl>, motifobste <dbl>, motifautre <chr>,
#   rpde12 <dbl>, modeacc <dbl>, instrum <dbl>, causecesar <dbl>,
#   pn <dbl>, sexe <dbl>, a1 <dbl>, a5 <dbl>, pha <dbl>,
#   transfertn <dbl>, comment1 <chr>, comment2 <chr>, orig <dbl>,
```

```
> dim (declenche_2025)
[1] 903  59
```

Le fichier contient 903 lignes/ observations et 59 variables (colonnes)

Questionnaire de recueil des informations sur les déclenchements

N° Fiche ……. N° Accouchement …….

Issue : 1. Cas (CS) ☐ 2. VB témoins ☐ 3. VB cohorte (cahier acc) ☐

Date accouchement (dd/mm/yyyy) …… Heure (hh/mn) ……

Date naissance mère ……. Age (par différence ou directement) ……..

Date prévue d'accouchement …… Age gestationnel (par différence ou directement) : ……..SA + ……. Jours

Nombre de fœtus : …… Présentation : 1. Céphalique ☐ 2. Siège ☐

Accouchement

Mode de déclenchement : 1. Maturation ☐ 2. Déclenchement ☐ 3. Les deux ☐

Motif de déclenchement : 1. Grossesse prolongée ☐ 2. Fœtal ☐ 3. Maternel ☐
4. Obstétrical ☐ 5. Plusieurs motifs ☐ 6. Convenance ☐

Motif fœtal : 1. ARCF ☐ 2. Patho foetale ☐ 3. Gémellaire ☐ 4. Baisse des MAF ☐
5. RCIU/PAG ☐ 6. Macrosomie ☐ 7. Cassure courbe croissance ☐ 8. MFIU ☐

Motif maternel : 1. HTA ☐ 2. Pathologie ☐ 3. Thrombopénie ☐ 4. Fenêtre thérapeutique ☐ 5. Métro ☐

Motif obstétrical : 1. RPM ☐ 2. Cholestase ☐ 3. Diabète ☐ 4. Allo-imm ☐ 5. Chorio-amniotite ☐
6. Oligamios ☐ 7. Hydramnios ☐ 8. Doppler anomalies ☐

Motifs autres ……. RPDE > 12 h : 1. Oui ☐ 2. Non ☐

Mode accouchement : 1. Voie basse ☐ 2. Instrumental ☐ 3. Césarienne ☐

Instrument : 1. Forceps ☐ 2. Ventouse ☐ 3. Spatule ☐ 4. Plusieurs ☐

Cause césarienne : 1. Itérative ☐ 2. Fœtale ☐ 3. Maternelle ☐
4. Obstétricale ☐ 5. Plusieurs causes ☐ 6. Convenance ☐

Nouveau-Né

Poids de naissance …….. Sexe : 1. Garçon ☐ 2. Fille ☐

Apgar 1 minute …… Apgar 5 minutes … pH artériel …… pH veineux …….

Compléments dans les dossiers

Utérus cicatriciel : 1. Oui ☐ 2. Non ☐

Poids mère …….. Taille mère …… IMC ……..

Prise de poids pendant la grossesse …… Bishop initial ……

RPDE au début du déclenchement : 1. Oui ☐ 2. Non ☐

Couleur du LA : 1. Clair ☐ 2. Teinté ☐ 3. Méconial ☐ 4. Sanglant ☐ 5. Clair puis teinté ☐

Surveillance de 2è ligne : 1. Toco interne ☐ 2. pH au scalp ☐ 3. ECG scalp ☐ 4. Toco + autre ☐

Dilatation au moment césarienne ……

Cause carienne en cours de travail : 1. ARCF ☐ 2. Stagnation ☐ 3. Echec ☐ 4. Non engagement ☐
5. ARCF + stagnation 6. Autre ☐

Date déclenche …….. Heure déclenche ……

(calcul de délai déclenche = heure acc - heure déclenche (avec correction en cas de franchissement de minuit)

Méthode déclenchement codage de 1 à 14 (pas mis ici)

On peut voir aussi le format des données. Toutes les variables sont sous la forme de « num » pour numériques, même, par exemple, la variables « sexe » (du bébé) avec 1 = Garçon et 2 = Fille (chiffres entiers), ou 40.3 et 41.1 pour les semaines d'aménorrhée (nombre avec une décimale). Seuls les formats de Date sont conservés (correspondant à la colonne des dates validés dans le fichier Excel), et il y a une variable en texte (en caractère) ($ motifautre: chr)

Les « NA » sont des cellules vides sous Excel : soit par défaut de notification, soit parce que sans objet.

Liste et format des variables :

> str (declenche_2025)	# Explications

tibble [903 × 59] (S3: tbl_df/tbl/data.frame)
$ fiche : num [1:903] 641 125 171 13 239 219 779
$ issue : num [1:903] 3 3 3 2 3 3 1 1 2 3 ...
$ dateacc : POSIXct[1:903], format: "2010-11-23" ...
$ hacc : num [1:903] 18 18 21 2 16 21 7 19 0 18 ..
$ macc : num [1:903] 31 16 37 40 17 22 26 4 30 30
$ hmacc : num [1:903] 18.52 18.27 21.62 2.67 16.28
$ geste : num [1:903] NA NA NA 1 NA NA 1 1 1 NA ...
$ parite : num [1:903] 1 1 1 1 2 1 1 1 1 1 ...
$ agemat : num [1:903] 19 19 17 18 19 19 18 19 19 19
$ agsem : num [1:903] 40 40 41 41 41 41 41 41 40 41
$ agj : num [1:903] 2 NA 1 1 5 NA 2 4 2 1 ...
$ ag : num [1:903] 40.3 NA 41.1 41.1 41.7 ...
$ modedecl : num [1:903] 1 2 2 2 3 2 2 2 2 2 ...
$ motifdecl : num [1:903] 4 2 1 1 1 1 3 1 2 1 ...
$ motiffoet : num [1:903] NA 7 NA NA NA NA NA NA 4 NA
$ motifmat : num [1:903] NA NA NA NA NA NA 1 NA NA NA
$ motifobste: num [1:903] 6 NA NA NA NA NA NA NA NA NA
$ motifautre: chr [1:903] NA NA NA NA ...
$ rpde12 : num [1:903] 2 2 2 2 2 2 1 2 2 2 ...
$ modeacc : num [1:903] 3 1 2 1 1 1 3 3 1 1 ...
$ instrum : num [1:903] NA NA 4 NA NA NA NA NA NA NA
$ causecesar: num [1:903] 5 NA NA NA NA NA 4 2 NA NA .
$ pn : num [1:903] 2985 2890 3665 3690 3475 ...
$ sexe : num [1:903] 2 2 1 1 2 1 1 1 1 2 ...
……..
	Numéro de ficher
	Césarienne / Voie basse
	Date d'accouchement
	Heure (entière) accouche
	Minute accouchement
	H décimale avec h+(m/60)
	Gestité
	Parité
	Age de la mère (années)
	Age gesta (sem entières)
	Idem + x jours / semaine
	Age gesta en décimales
	Mode de déclenchement
	Motif principal codé
	Motifs fœtaux
	Motifs maternels
	Motifs obstétricaux
	Autres motifs en clair
	Rupture poche eaux > 12 h
	Mode accouche (VB, I, CS)
	Forceps
	Cause césarienne
	Poids de naissance
	Sexe nouveau-né
	……..

Voir dans le document supplémentaire le questionnaire avec ses codages

Question importante : en cas de réponse binaire *ne pas coder "Yes" et "No", ni "Y" ou "N" car ce sont des lettres. Préférer deux possibilités : soit "1" pour* oui *et "2" pour* non *: ce sera plus commode pour les tableaux 2 x 2, soit "0" pour* non *et "1" pour* oui *: c'est ennuyeux pour les tableaux (mais rattrapable) et commode pour les analyses multivariées.*

Comment obtenir de l'aide sur "R" ?

- Sur internet avec un moteur de recherche
- Sur le site du CRAN project R : https://cran.r-project.org/
- Avec l'intelligence artificielle : apport indéniable, pas toujours clair, se méfier des erreurs ++
- Chaque commande ou package peut être interrogée avec la commande `> help (command)` ou `> help (package)`. La fenêtre inférieure droite affiche les composantes. Pas toujours clair…
- Voir la liste des commandes dans un package `> ls ("package")` (voir dans le document supp)

4. Comprendre les données

(partie pouvant être sautée si les données sont "solides" ; passer alors à la partie 5)

Note : les commandes sont présentées une à une pour être lancées dans la Console (à la suite de >). On peut aussi copier ou écrire les commandes dans l'éditeur de texte (sans le >) et les lancer groupées ou à la suite avec le petit bouton *Run* : c'est un fichier qui peut être sauvé en *.R pour être sauvé et repris ultérieurement

a. Valeurs manquantes, valeurs aberrantes, valeurs minimales et maximales

Permet d'explorer les variables sous format numérique (ne pas tenir compte des résultats pour les variables qualitatives (codées), ou celles de date ou de caractères (texte).

```
> summary (declenche_2025)
     fiche            issue          dateacc
 Min.   :   1.0   Min.   :1.000   Min.   :2010-01-01 00:00:00
 1st Qu.: 249.5   1st Qu.:2.000   1st Qu.:2010-04-06 00:00:00
 Median : 504.0   Median :3.000   Median :2010-07-06 00:00:00
 Mean   : 504.9   Mean   :2.507   Mean   :2010-07-04 16:00:00
 3rd Qu.: 760.5   3rd Qu.:3.000   3rd Qu.:2010-10-03 00:00:00
 Max.   :1010.0   Max.   :3.000   Max.   :2010-12-31 00:00:00

      hacc            macc           hmacc           geste
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.030   Min.   :1.000
 1st Qu.: 6.50   1st Qu.:14.00   1st Qu.: 6.995   1st Qu.:1.000
 Median :15.00   Median :29.00   Median :15.970   Median :2.000
 Mean   :13.16   Mean   :29.04   Mean   :13.648   Mean   :2.186
 3rd Qu.:19.00   3rd Qu.:44.00   3rd Qu.:19.460   3rd Qu.:3.000
 Max.   :23.00   Max.   :59.00   Max.   :23.930   Max.   :9.000
                                                  NA's   :537
     parite          agemat          agsem            agj
 Min.   :1.000   Min.   :16.00   Min.   :27.00   Min.   :0.000
 1st Qu.:1.000   1st Qu.:26.00   1st Qu.:38.00   1st Qu.:2.000
 Median :1.000   Median :30.00   Median :40.00   Median :3.000
 Mean   :1.827   Mean   :29.98   Mean   :39.27   Mean   :3.237
 3rd Qu.:2.000   3rd Qu.:34.00   3rd Qu.:41.00   3rd Qu.:5.000
 Max.   :8.000   Max.   :46.00   Max.   :42.00   Max.   :9.000
 NA's   :14      NA's   :1       NA's   :16      NA's   :250
```

Le nom des variables est normalement du type "**declenche_2025$var**" avec le premier mot qui correspond au nom du tableau de données puis la valeur dollar du clavier $, puis le nom de la variable en respectant les majuscules et les minuscules ++ Avec la commande `> attach(declenche_2025)` on peut se contenter de nommer les variables simplement par leur nom +++

```
> attach (declenche_2025)
```

On peut reprendre les lignes précédentes avec la touche ↑ flèche vers le haut

La parité est le nombre d'enfants qu'a eus la mère en comptant celui qui va naître

```
> summary(parite)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   1.000   1.827   2.000   8.000      14
```

"NA's" veut dire "cellules vides" telles qu'elles ont été importées à partir d'Excel. On peut les rendre vraiment vide avec la commande `na.omit`

```
> parite <- na.omit (parite)
> summary(parite)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.827   2.000   8.000
```

On voit que, au lieu d'écrire une commande directement, on peut diriger le résultat vers un nom choisi a priori, puis **de définir ce résultat avec la flèche <-** (il s'agit donc d'une commande qui va dans le sens contraire de l'écriture). Ensuite, mettre seul le nom sur la ligne suivante puis ↵, ou l'explorer avec `> summary (fichier` ou `variable)` l'insérer dans une nouvelle commande.

Autre exemples : taille de la mère, poids de naissance

```
> summary (taille)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.45    1.58    1.63    1.63    1.68    1.79     617
> summary (pn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1070    2940    3322    3274    3641    4720       7
```

b. Transformer les formats des variables

1. Transformer une variable numérique en facteur (variable qualitative ou codée pour décrire et faire des tableaux croisés) : commande `as.factor`

```
> geste <- as.factor(geste)
> str(geste)
 Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 3 1 1 1 1 4 ...
> head (geste)
[1] 1 1 1 1 3 1
Levels: 1 2 3 4 5 6 7 8 9
> summary(geste)
  1   2   3   4   5   6   7   8   9
152 106  51  28  14   8   4   2   1
```

Cette dernière commande montre les codages (et non plus les moyennes) tels qu'il y a 152 observations de codage "1", 106 de codage "2" etc..

2. Transformer une variable facteur en numérique (variable codée pour décrire et faire des tableaux) **:** commande `as.numeric`

```
> geste_num <- as.numeric (geste)
> summary (geste_num)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.186   3.000   9.000
```

3. Transformer les dates

Voir dans le document supplémentaire, cet aspect est souvent compliqué, surtout si lors de la saisie le format n'a pas été maîtrisé.

c. Recoder des variables (remplacer 3 par 2 par exemple)

A manipuler sur des données numériques +++ (avant de transformer en factor)

```
> issue2 <- replace (issue, issue==3,2)
```

Puis repasser aux factor :

```
> issue <- as.factor(issue)
> summary (issue)
  1   2   3
148 149 606
> issue2 <- as.factor(issue2)
> summary(issue2)
  1   2
148 755
```

On voit que le codage "3" a été renommé "2" et que le nouveau codage "2" est la somme de 149 + 606 = 755. Sauver le nouveau fichier avec un autre nom (voir infra).

d. Recoder une variable en changeant le code "2" par "0"

Créer une nouvelle variable (`issue2`) si on ne veut pas casser le codage initial (`issue`).

Là aussi sur des variables en numérique +++ ; passer en `factor` ensuite

```
> issue2 <- replace (issue, issue==2,0)   # avec 2 signes = à la suite
> issue2 <- as.factor (issue2)
> summary(issue2)
```

e. Sélectionner une partie du fichier (choisir des observations ou des lignes)

La variable "`issue`" est codée en 3 modalités (CS, VB témoins, VB cohorte). On veut analyser seulement des observations codées "1" (césarienne) et codées "2" (voie basse à partir des dossiers) pour réaliser une enquête cas-témoins.

```
> decl_cas_tem <- subset (declenche_2025, issue<3)
```

Création d'un nouveau fichier **`decl_cas_tem`** qui contient toutes les variables pour lesquelles issue=1 ou 2. Ensuite quitter le fichier précédent et analyser le nouveau en cours.

f. Sauver les fichiers et les sorties

Sauver ou importer des fichiers

```
> save (decl_cas_tem, file="decl_cas_tem.RData")
```

Préciser le chemin pour le retrouver : voir l'onglet **Terminal** au-dessus de la console

Pour le charger pour une étude ultérieure (barre supérieure : `File / Open File ou`

```
> load("~/decl_cas_tem.RData")
```

Sauver éditeur de texte (commandes à la suite)

Cliquer sur la disquette en haut de la fenêtre de l'éditeur de texte

Sauver les commandes et les résultats

Mettre en surbrillance les résultats voulus et copier/coller dans un fichier de texte (Word ou autre)

g. Pour quitter

Quitter un fichier ou pour en ouvrir un autre

```
> detach (fichier_en_cours)
```

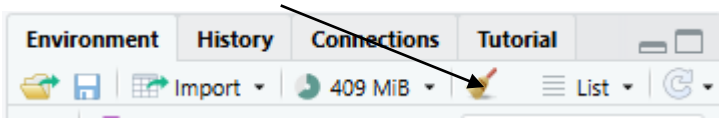
Supprimer un fichier en cours

```
> remove (fichier)
```
OU `> rm (fichier)`

Vider "Environment"

Commande `rm(list = ls())`

OU Cliquer sur le balai OU Ctrl-Shift-F10 (Windows) OU Cmd-Shift-F10 (Mac)



Quitter RStudio

Sauve par défaut le travail en cours et l'aspect de l'écran.

Pour ne pas sauver ce qui a été fait (par défaut) :

Onglets *Tools / Global Options /* décocher *Restore…* .

Pour quitter "R" : Onglet *File / Quit Session : choisir Save ou Don't Save*

5. Décrire les données

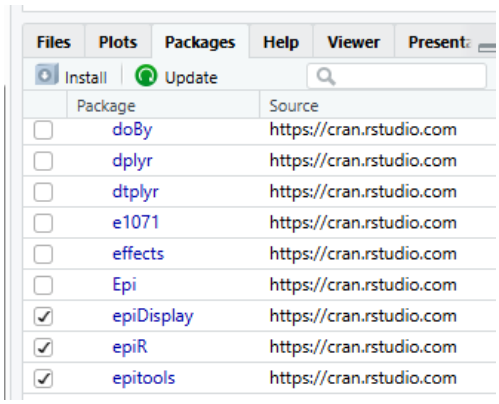
Le logiciel "R" de base n'a pas été fait pour les études épidémiologiques. Il faut télécharger – au fil de l'avancée de l'analyse – des "packages" ou "paquets" qui font fonctionner des commandes pour simplifier le travail. On va prendre pour commencer trois packages : *epitools*, *epiDisplay* (orthographe à respecter avec un "D" majuscule) et *epiR*, il en existe une trentaine dans le domaine épidémiologique (voir https://cran.r-project.org/web/views/Epidemiology.html) !

a. Installer des packages (avec des guillemets) et les activer (sans les guillemets)

```
> install.packages ("epitools")
> install.packages ("epiDisplay")
> install.packages ("epiR")

> library (epitools)
> library (epiDisplay)
> library (epiR)
```

S'assurer que, dans l'onglet *Packages* (en bas à droite), les packages en question sont cochés ++



b. Téléchargez le fichier de l'enquête cas-témoins

Dans Windows ou Mac

- Copier le dossier à partir du site www.episure.fr/ Formations/ "declenchement.zip"
- Dézipper le dossier
- Mettre à part le fichier "**decl_temoins.xlsx**"

Dans RStudio

- Onglet *File / Import Dataset* : importer le fichier "`decl_temoins.xslx`"
- Commande `attach(decl_temoins)`

c. Décrire les variables qualitatives (codées ou catégorielles ou "factor)

La commande « table » seule est assez succincte (pas en tableau, pas de fréquences)

```
> table (issue)
```

```
issue
  1   2
147 148
```

Le codage 1 correspond aux césariennes (CS) à la suite du déclenchement, 2 aux voies basses témoins (VBT) tirées au sort parmi toutes les voies basses pour un examen de dossier plus complet. Le fait que ce ne soit pas le même nombre n'est pas gênant.

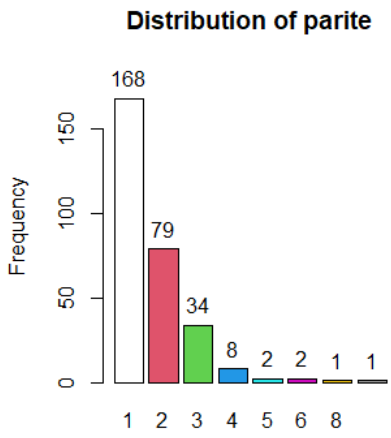
On peut donner un label (une étiquette) à la variable "`issue`"

```
> issue <- factor(issue, levels=c(1,2), labels=c("CS", "VBT"))
> table (issue)  OU summary (issue)
```

```
issue
 CS VBT
147 148
```

Pour améliorer le résultat :

- Avec un graphe en barre associé (fenêtre en bas à droite) : avec l'axe des ordonnées en effectifs (c'est « `Frequency`» dans R qui veut dire « effectifs) ou en % (c'est « `Percent`»)
- Avec ou sans les pourcentages cumulés
- Avec un nombre de décimales précises
- Avec ou sans les manquants



Utiliser la commande "`tab1`" de "*epiDisplay*" (sur une seule ligne à la suite)

```
> tab1 (issue)
issue :
        Frequency Percent Cum. percent
CS            147    49.8         49.8
VBT           148    50.2        100.0
  Total       295   100.0        100.0
```

```
> tab1 (parite,decimal=2, cum.percent=FALSE, graph=TRUE,
missing=TRUE,bar.values=c("frequency"))
parite :
        Frequency   %(NA+)   %(NA-)
1             168    56.95    57.14
2              79    26.78    26.87
3              34    11.53    11.56
4               8     2.71     2.72
5               2     0.68     0.68
6               2     0.68     0.68
8               1     0.34     0.34
<NA>            1     0.34     0.00
  Total       295   100.00   100.00
```

- Montre les % avec ou sans les "NA". Les primipares sont au nombre de 168, soit 57.1 % sans NA
- Le graphique en barre montre les effectifs (fréquence) des différentes catégories
- La mention "`graph=TRUE`" ajoute le graphe (succinct en 1ère utilisation) dans la fenêtre inférieure droite.
- Faire des graphes sous "R" est un métier. Voir tuto à part. Essayer avec package "`ggplot2`"

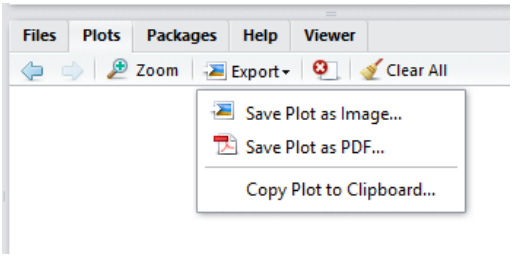
Autre exemple

```
> barplot(table (issue))
> pie(table(parite))
```

Pour préciser des couleurs : rajouter dans les parenthèses

```
> pie(table(parit),col=c("red","blue","green"))
```

Note : On ne fait plus de graphe en camembert dans les articles ("trop d'encre")



Pour sauver le graphe, voir la fenêtre en bas à droite : cliquer sur Export, ou Copy Image

d. Décrire les variables quantitatives

Plusieurs possibilités, mais pas de commande unique :

Une seule variable

```
> summary (pn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1340    2908    3280    3249    3660    4720
```

Toutes les variables :

```
> summary(decl_temoins)
       Résultat non mis
> median(pn)
[1] 3280
> sd (pn)
[1] 613.9723
> options (digits=4) #maîtrise de la taille des nombres et de décimales
> quantile (pn)
  0%  25%  50%  75% 100%
1340 2908 3280 3660 4720
```

Pour avoir une vue d'ensemble avec une seule commande (ouvrir un package rien que pour cela…)

```
> install.packages("psych")
> library(psych)
> describe (pn)
          vars   n   mean  sd    median  trimmed    mad  min  max   range
       X1    1  296 3249 612.93   3275   3278.53   563.39 1340 4720  3380
```

NB : `trimmed` est la moyenne tronquée de 5 % max et 5 % min

`mad` est la médiane des écarts à la médiane (mesure de dispersion comme l'écart-type)

`range` est la différence max – min

Pour créer des classes ou des seuils (il faut 6 seuils pour 5 classes avec un très petit et un très grand) ➔ *Comprendre la forme dans les commandes de = c(… qui définit les valeurs d'un vecteur).*

```
> ageclasse=cut(agemat,breaks=c(0,20,25,30,35,45))
> summary (ageclasse)
 (0,20] (20,25] (25,30] (30,35] (35,45]
     13      45      98      88      51
```

Les parenthèses incluent la valeur dans la classe, les crochets l'excluent. Si on a un doute :

```
> ageclasse = cut(agemat, breaks=c(0,20,25,30,35,45), labels=c("<20", "20-
24","25-29","30-34","35+"))
> summary(ageclasse
         <20   20-24 25-29 30-34  35+
          13    45    98    88    51

> tab1 (ageclasse)
ageclasse :
        Frequency Percent Cum. percent
<20            13     4.4          4.4
20-24          45    15.3         19.7
25-29          98    33.2         52.9
30-34          88    29.8         82.7
35+            51    17.3        100.0
  Total       295   100.0        100.0
```

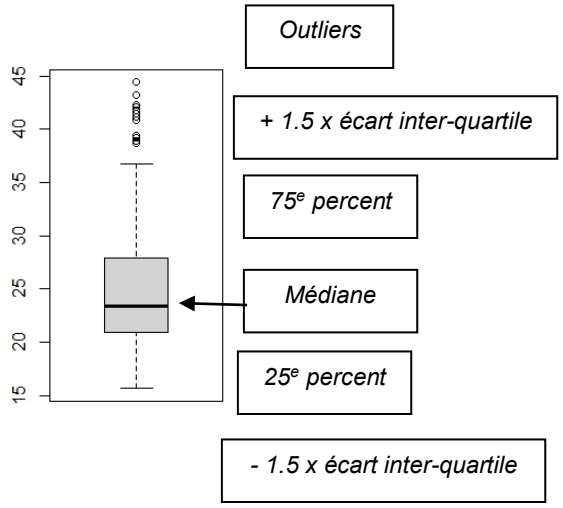
Créer une nouvelle variable avec un seuil et deux classes : âge maternel / 30 ans

```
> age30 <- cut(agemat, breaks=c(0,30,45))
> tab1 (age30)
age30 :
        Frequency Percent Cum. percent
(0,30]        156    52.9         52.9
(30,45]       139    47.1        100.0
  Total       295   100.0        100.0
```

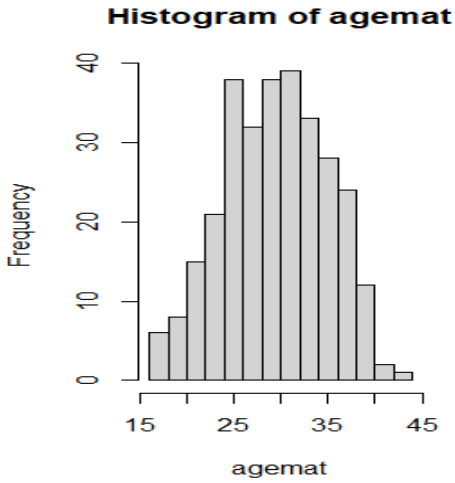
Quelques graphes pour les données quantitatives. On ne fait plus guère d'histogramme dans les articles, mais plutôt des box-plots. Tout est modifiable dans les graphes avec les lignes de commande +++

Voir le chapitre "Graphes" dans le supplément. Un tuto général "Graphes" est disponible sur Episûre

```
> boxplot(imc)
```




```
> hist(agemat)
```



6. Comparaison : la variable de jugement est qualitative

L'analyse dite **univariée**, ou - dans quelques ouvrages et logiciels – dite **bivariée** décrit les relations entre deux variables, l'une correspond au critère de jugement (dans l'exemple CS et VB), l'autre correspond à la variable d'exposition (par exemple le mode de déclenchement, les motifs de non-travail, le poids ou la taille de la mère, la parité etc.….On analyse ici le fichier cas-témoins `decl_temoins`.

Rappel : Résumé des comparaisons et *tests disponibles*

		Critère de jugement, OU Variable Maladie, OU Outcome, OU Variable Cas/Témoins, OU Variable à expliquer, OU Variable dépendante	
		La variable est qualitative ou codée ou catégorielle ou "factor" (en deux classes le plus souvent)	La variable est quantitative
Variable d'exposition, OU Variable explicative	La variable est qualitative	Tableau 2 x 2 *Test du χ^2 ou test de Fisher* Odds ratio ou Risque relatif	Moyenne et dispersion de la var. jugement *Test de t ou Tests non paramétriques (Wilcoxon, Mann-Whitney)*
	La variable est quantitative	Moyenne et dispersion de la var. expo *Test de t ou Tests non paramétriques (Wilcoxon, Mann-Whitney)*	Coefficient de corrélation r ou ρ Droite de régression linéaire *Test de la pente à 0*

a. La variable de jugement et la variable d'exposition sont qualitatives

Principe d'un tableau 2 x 2

Exemple avec `issue` (1 CS, 2 VB) et `primipare` (1 oui, 2 non=multipare). `Issue` est le critère de jugement (outcome) et doit être en colonne (mis en second dans la commande), `primi` est la variable dite d'exposition (mise en 1er dans la commande) et doit être en ligne. Les analyses se font avec les % en

colonnes dans les enquêtes cas-témoins (a/M1 versus b/M2), et avec les % en ligne dans les enquêtes de cohorte (a/E1 versus c/E2).

	Maladie Outcome +	**Maladie Outcome -**	**Total**
Exposition +	a	b	E1
Exposition-	c	d	E0 ou E2
Total	M1	M0 ou M2	Total population

Préparation : mise des variables en factor

```
> primi <- as.factor (primi)
> summary(primi)
   1    2 NA's
 168  126    1

> issue <- factor(issue, levels=c(1,2), labels=c("CS", "VBT"))
> summary(issue)
 CS VBT
146  146
```

Tableau 2 x 2 : table (lignes, colonne)

```
> table (primi, issue)
     issue
primi  CS VBT
    1 113  55
    2  33  93

> addmargins (table (primi, issue))
     issue
primi  CS VBT Sum
  1   113  55 168
  2    33  93 126
  Sum 146 148 294


> options (digits=2)
> prop.table (table(primi, issue)) # % sur le total !
     issue
primi   CS  VBT
    1 0.38 0.19
    2 0.11 0.32

> tableau2 <- prop.table (table (primi, issue),2) # 2 pour % en colonnes dans une enquête cas-témoins
> tableau2
     issue
primi   CS  VBT
    1 0.77 0.37
    2 0.23 0.63

> addmargins (tableau2, 1) # 1 pour % total par colonnes
     issue
primi   CS  VBT
  1   0.77 0.37
  2   0.23 0.63
  Sum 1.00 1.00
```

Avec des recodages pour être plus explicites, création de `labels`

```
> primi <- factor (primi, levels=c(1,2), labels=c("Primi", "Multi"))
> tableau3 <- prop.table (table (primi, issue),2)
> addmargins (tableau3, 1)
       issue
primi     CS  VBT
  Primi 0.77 0.37
  Multi 0.23 0.63
  Sum   1.00 1.00
```

 # Lecture (en colonnes) : les "CS" sont primipares à 77 %, tandis que les "VB" le sont à 38 %. La primiparité semble associée fortement avec l'issue des déclenchements avec césarienne. Il faut le prouver avec un test statistique.

Tester la différence entre deux % de variables qualitatives

Pour comparer deux pourcentages dans un tableau 2 x 2, c'est le test du χ^2 de Pearson : commande `tab2by2.test` du package "epiR"

```
> install.packages("epiR")  ou cocher dans la fenêtre "Packages" du coin inférieur droit
> library (epiR)
> tab2by2.test (primi,issue)
         Outcome
Predictor  CS VBT
    Primi 113  55
    Multi  33  93

$p.value
         two-sided
Predictor midp.exact fisher.exact chi.square
    Primi         NA           NA         NA
    Multi      2e-12      2.1e-12    3.2e-12

$correction
[1] FALSE
```

Le test du `chi-square` montre un χ^2 = 3.5 x 10^{-12}, très inférieur à p<0.05 (en fait c'est p = 0,avec 12 zéros..3.2, ou 3.2 * 10^{-12}), soit très significative. On ne "prendra" le "`ficher.exact`" qu'en cas de petits effectifs.

Autre commande possible :

```
> summary (table (issue, primi))
Number of cases in table: 294
Number of factors: 2
Test for independence of all factors:
        Chisq = 49, df = 1, p-value = 3e-12
```

Autre commande possible

```
> chisq.test(issue, primi)
        Pearson's Chi-squared test with Yates' continuity correction
data:  issue and primi
X-squared = 47, df = 1, p-value = 7e-12
```

La correction de Yates utilisée ici est réservée en cas de petits effectifs (pas utile dans cet exemple)

Autre commande possible avec le test de Fisher :

```
> fisher.test(primi,issue)
        Fisher's Exact Test for Count Data
data:  primi and issue
p-value = 2e-12
```

Pour mémoire : Grands effectifs : χ^2 de Pearson

Petits effectifs : Correction de Yates (peu utilisée), Test exact de Fisher, Test exact Mid-P

Calcul de l'odds ratio (pour les enquêtes cas-témoins)

On a donc le tableau 2 x 2 suivant

- Les odds : odds des CS (113 : 33) et odds des VB (55 : 93)
- Rapport des odds (OR) "à la main" vaut : OR = (113 : 33) / '55 : 93) = (on croise) 113 x 93 / 33 x 55 = 5.79

	Césarienne CS	Voie basse VB
Primi	113	55
Multi	33	93
Total colonnes	148	148

Il y a (en approximation) 5.79 fois plus de CS chez les primipares que chez les multipares. C'est le rapport de risque de la ligne des primi / la ligne des multi. On peut dire que le risque pour les multi est de "1" et de "5.79" pour les primi.

Ne pas utiliser les commandes suivantes de `epitools` qui affiche des résultats trompeurs. Exemple montré pour comprendre ce qu'est un OR avec `epitab` ou `oddsratio` et les erreurs possibles d'interprétation.

```
> epitab (table(primi, issue), method="oddsratio")
$tab
       issue
primi     CS    p0 VBT     p1 oddsratio lower upper   p.value
  Primi  113 0.774  55 0.3716      1.00    NA    NA        NA
  Multi   33 0.226  93 0.6284      5.79 3.472 9.656 2.106e-12
```

`epitab` a choisi de mettre OR = 5.79 sur la ligne des multi ce qui est intuitivement trompeur : attention donc ! +++ . Il faudrait le mettre sur la ligne des primi puisque c'est un OR des primi / multi +++. L'intervalle de confiance est bon et vaut 3.47 – 9.66 (très différent de 1).

Autre mauvaise manière pour l'odds ratio

```
> oddsratio (primi,issue, method=c ("fisher"), verbose=TRUE)
$measure
         odds ratio with 95% C.I.
Predictor estimate lower upper
    Primi    1.000    NA    NA
    Multi    5.751 3.371 9.995
```

Même erreur de ligne +++++ ! Valeur de l'OR légèrement différente (calcul avec `midp` plutôt que direct).

Pour le sens des lignes et des colonnes, en cas de codage 0 et 1 par exemple (le 1 étant toujours pathologique), on peut changer les valeurs de références (1ère pour les lignes ou les colonnes), avec `rev` pour `rows` à inverser par exemple.

Le mieux est d'utiliser la commande `epi.2by2` du package `"epiR"` (visiblement fait par des épidémiologistes) et la partie `method="case.control"` pour les études cas-témoins . Voir en fin de document la liste des commandes des packages utilisés. Là c'est plus simple et c'est bon

```
> View(decl_temoins)
> attach(decl_temoins)
> install.packages("epiR")
> library(epiR)
> issue <- as.factor(issue)
> primi <- as.factor(primi)

> options (digits=4)
> tableOR <- table (primi, issue)
> epi.2by2 (tableOR, method="case.control")
             Outcome+    Outcome-      Total
Exposed +         113          55        168
Exposed -          33          93        126
Total             146         148        294
                         Odds
Exposed +  2.05 (1.51 to 2.91)
Exposed -  0.35 (0.24 to 0.52)
Total      0.99 (0.78 to 1.24)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Exposure odds ratio                              5.79 (3.47, 9.66)
Attrib fraction (est) in the exposed (%)        82.73 (71.24, 89.63)
Attrib fraction (est) in the population (%)     64.03 (58.36, 69.90)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 48.584 Pr>chi2 = <0.001
Fisher exact test that OR = 1: Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
```

Même commande avec la variable `obese` (IMC ≥ 30 kg/m² codée en "1", et "non obèse codée en 2)

```
> tableOR2 <- table (obese, issue)
> epi.2by2 (tableOR2, method="case.control")
             Outcome+    Outcome-      Total
Exposed +          26          22         48
Exposed -         116         120        236
Total             142         142        284
                         Odds
Exposed +  1.18 (0.66 to 2.20)
Exposed -  0.97 (0.75 to 1.25)
Total      1.00 (0.79 to 1.27)
Point estimates and 95% CIs:
-------------------------------------------------------------------
Exposure odds ratio                              1.22 (0.66, 2.28)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 0.401 Pr>chi2 = 0.527
Fisher exact test that OR = 1: Pr>chi2 = 0.635
 Wald confidence limits
 CI: confidence interval
```

L'OR n'est pas significatif (on dit aussi significativement différent de 1), avec un intervalle de confiance qui comprend 1, et un test du χ^2 non significatif avec $p > 0.05$.

Calcul du risque relatif pour les enquêtes de cohorte

Un risque relatif (RR) s'applique aux enquêtes de cohorte, où le rapport de risque est un rapport d'incidence de la maladie chez les exposés sur les non-exposés. La disposition du tableau 2 x 2 est la même que pour le calcul de l'OR, mais on fait les totaux sur les lignes.

Supprimer les fichiers téléchargés antérieurement avec la commande `rm ()`, puis télécharger le fichier `decl_cohorte` à partir des fichiers dézippés de Epi-Sûre

```
> attach (decl_cohorte)
> options (digits=2)
> install.packages("epitools")
> library(epitools)
> issue <- factor(issue, levels=c(1,2),labels=c("CS", "VBC"))
> primi <- factor(primi, levels=c(1,2),labels=c("Primi", "Multi"))

> table (primi, issue)
      issue
primi   1   2
    1 113 322
    2  33 405
> addmargins (table (primi, issue))
      issue
primi   1   2 Sum
  1   113 322 435
  2    33 405 438
  Sum 146 727 873
```

	Césarienne	Voie basse	Total
Primi	113	322	435
Multi	33	405	438

Les totaux et les % se font en ligne et le RR (à la main) vaut :

(113 : 435) / (33 : 438 = 0.2598 / 0.0753 = 3.45

```
> tableauRR <- prop.table(table(primi, issue),1)
> tableauRR
      issue
primi     1     2
    1 0.260 0.740
    2 0.075 0.925
> addmargins (tableauRR,2)
      issue
primi     1     2   Sum
    1 0.260 0.740 1.000
    2 0.075 0.925 1.000
```

Attention, la commande `riskratio` de "Epitools" risque d'être trompeuse : elle écrit un RR sur la mauvaise ligne comme pour l'odds ratio :

Le mieux est d'utiliser la commande `epi.2by2` du package `"epiR"` comme pour les OR, mais avec la partie `method="cohort.count"`

```
> table_coh <- table(primi, issue)
> epi.2by2(table_coh, method="cohort.count")
             Outcome+    Outcome-      Total                 Inc risk *
Exposure+         113         322        435     25.98 (21.92 to 30.37)
Exposure-          33         405        438       7.53 (5.24 to 10.42)
Total             146         727        873     16.72 (14.31 to 19.37)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 3.45 (2.39, 4.96)
Inc odds ratio                                 4.31 (2.85, 6.52)
Attrib risk in the exposed *                   18.44 (13.64, 23.25)
Attrib fraction in the exposed (%)            71.00 (58.45, 79.86)
Attrib risk in the population *                9.19 (5.69, 12.69)
Attrib fraction in the population (%)         54.95 (46.21, 63.35)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 53.302 Pr>chi2 = <0.001
Fisher exact test that OR = 1: Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Le RR vaut 3.45 (2.39 – 4.96) et est significativement différent de 1 (p < 0.001) : il y a 3.45 fois plus de césariennes chez les primipares par rapport aux multipares.

b. La variable de jugement est qualitative et la variable d'exposition est quantitative

La variable de jugement a 2 modalités qualitatives : comparer deux moyennes

On revient au fichier cas-témoins `decl_temoins.xlsx`

```
> library(readxl)
> decl_temoins <- read_excel ()
```
avec le chemin du fichier `decl_temoins.xlsx`
```
> attach(decl_temoins)
> issue <- as.factor (issue)
```

Décrire la variable par catégories : moyennes…. (`pn` = poids de naissance en g)

```
> tapply (pn, issue, mean)
       1        2
3165.714 3331.953

> tapply (pn, issue, sd)
       1        2
676.0058 534.9619

> issue <- factor(issue, levels=c(1,2), labels=c("CS", "VBT"))
> options (digits=4)
> tapply(pn,issue, mean)
      CS     VBT
    3166    3331
```

Lorsqu'il existe des valeurs manquantes ("NA" ou "NA's"), dans une des variables (ici "taille" a 10 valeurs manquantes sur 295 observations), les ennuis commencent…..

Certaines commandes comme tapply() ou cor() (voir plus loin) ne fonctionnent pas et un résultat s'affiche avec "NA" ou "pas de la même longueur" (cas où une variable n'a pas le même nombre de lignes remplies qu'une autre variable comparative).

Diverses solutions "casse-tête" :

= Essayer d'enlever "NA" dans tout le fichier avec création d'un nouveau fichier : `> new_fichier <- na.omit (fichier_en_cours).` Risque de supprimer toutes les valeurs !

= Essayer d'enlever les "NA" dans chaque variable à étudier avec : `> taille <- na.omit (taille).` Ne marche pas toujours

= Créer des nouveaux fichiers successifs sans les "NA" avec la commande subset et la commande !is.na qui enlève des NA :

```
decl1 <- subset (decl_temoins,!is.na (taille))
decl2 <- subset (decl1,!is.na(pn))
decl3 <- subset (decl2, !is.na(parite))
attach (decl3)
nrow (decl3)
[1] 285
length (taille)
length (pn)
length (parite)
length (issue)
summary (taille)
summary (parite)
summary (pn)
tapply (taille,issue, mean)
       1        2
1.611690 1.647063
```

OU commande agregate

```
> moy <-aggregate(taille~issue, FUN=mean)
> moy
  issue   taille
1    CS  1.611690
2    VBT 1.647063
> options (digits=3)
> moy
  issue taille
1    CS   1.61
2   VBT   1.65
```

Comparer les moyennes et faire les tests

On veut comparer les poids de naissance des nouveau-nés ainsi que la taille des mères selon l'issue (césarienne ou voie basse). Noter bien le caractère tilde « ~ » qui veut dire « en fonction de » (touche sous le "2" ou le "é" sur le clavier avec alt gr à droite de la barre d'espace, sinon touches Alt et 126, puis écrire le caractère à suivre).

Test t de Student : on suppose les variances égales

```
> t.test (taille~issue, var.equal=TRUE)
Two Sample t-test
data:  taille by issue
t = -4.6, df = 283, p-value = 6e-06
alternative hypothesis: true difference in means between group CS and group VBT
is not equal to 0
95 percent confidence interval:
 -0.05051 -0.02023
sample estimates:
mean in group CS mean in group VBT
           1.612             1.647
```

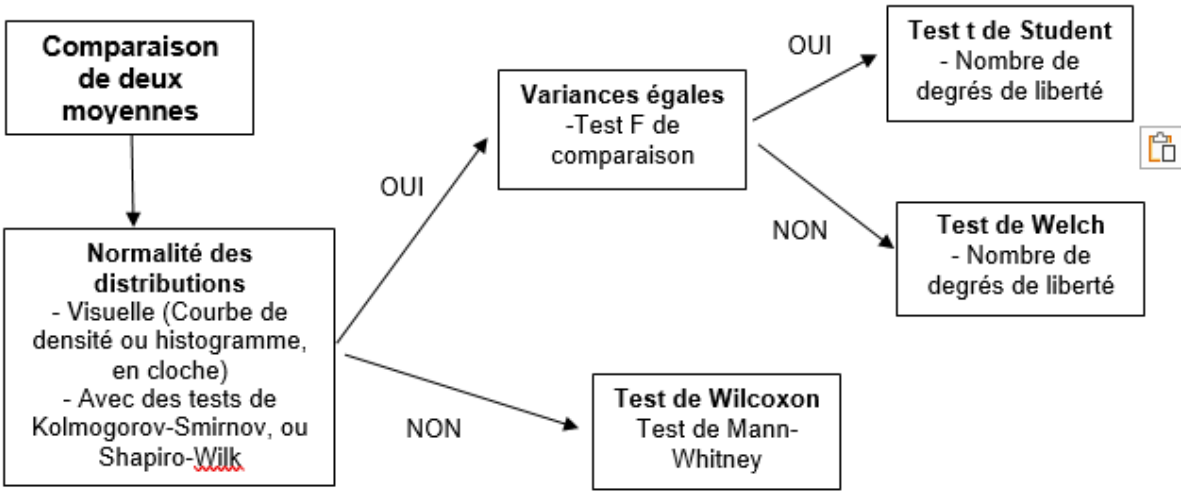
Les femmes avec CS mesurent 1 m 61 tandis que les femmes avec VB mesurent 1 m 64 et c'est très significatif ($p<10^{-5}$) : les femmes petites ont un plus grand risque de CS (quand on les déclenche) (causalité à explorer). Noter : `df` = degrees of freedom (degrés de liberté) pour adapter la loi de Student et la valeur de t au nombre d'observations (ici 284 observations – 1 = 283). Pour des grands nombres comme ici, la valeur de t pour le $p<0.05$ est voisine de $\varepsilon \geq |1.96|$

Test de Welch : on suppose les variances inégales

```
> t.test (pn~issue, var.equal=FALSE)
Welch Two Sample t-test
data:  pn by issue
t = -2.3, df = 277, p-value = 0.02
alternative hypothesis: true difference in means between group CS and group VBT
is not equal to 0
95 percent confidence interval:
 -306.03  -26.45
sample estimates:
 mean in group CS mean in group VBT
             3166              3332
```

La différence de poids de naissance est de 3 166 g - 3 332 g = - 164 g (intervalle de confiance) à 95 % notée ici de – 306 g à – 26 g, ne comprend pas "0" donc significativement différent de "0". A noter que le `pn` des enfants nés par césarienne est moins élevé que celui des enfants nés par voie basse ➔ revenir à la clinique obstétricale….+++ (remarquer que les femmes avec CS sont plus petites et donc peuvent avoir des enfants plus petits en poids).

Note qui dépasse ce tuto : le test de Student suppose la normalité des distributions de la variable quantitative et l'égalité des variances. Quand la normalité est établie mais que les variances ne sont pas égales, le test de Welch est à utiliser. Quand la normalité n'est pas établie, voir le test de Wilcoxon. "R" propose des commandes pour vérifier ces présupposés : voir `var.test ()`, *et* `shapiro.test()` *par exemple, plus les commandes de graphes (voir infra). Pour tester la normalité, il existe aussi des commandes ; pour le fichier ici, en raison des grands effectifs, on peut se dédouaner de prouver la normalité des distributions.*



Test non-paramétrique (basé sur les rangs)

Avec un test non paramétrique (ici test de Wilcoxon non apparié) sinon test de Kruskall-Wallis."R" ne donne pas les valeurs moyennes ou médianes…

```
> wilcox.test (parite~issue)
Wilcoxon rank sum test with continuity correction
data:  parite by issue
W = 6396, p-value = 1e-11
alternative hypothesis: true location shift is not equal to 0

> kruskal.test (pn~issue)
Kruskal-Wallis rank sum test
data:  pn by issue
Kruskal-Wallis chi-squared = 3.5, df = 1, p-value = 0.06
```

La variable de jugement a plus de 2 catégories qualitatives : analyse de 3 moyennes ou plus = analyse de variance (ANOVA)

Supprimer les fichiers téléchargés antérieurement avec la commande `rm ()`, puis télécharger le fichier `decl_cohorte` à partir des fichiers dézippés de Epi-Sûre.

On analyse le fichier total dans lequel la variable `issue` en codée en 3 catégories.

```
> View(declenche_2025)
> attach(declenche_2025)
> options (digits=4)
> issue <- as.factor (issue)

> summary (issue)
  1   2   3
148 149 606

> moy <-aggregate(pn~issue, FUN=mean)
> moy
  issue   pn
1     1 3152
2     2 3331
3     3 3291
```

```
> anova (aov(pn~issue))
Analysis of Variance Table
Response: pn
           Df   Sum Sq Mean Sq F value Pr(>F)
issue      2 2.87e+06 1434512    4.48  0.012 *
Residuals 893 2.86e+08  320254
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Les 3 moyennes de `pn` (3 152 g, 3 331 g et 3 291 g) diffèrent entre elles avec p = 0.012. On peut explorer laquelle diffère des deux autres ; ce sont les tests dits "post-hoc" de Tukey ou Bonferroni par exemple. Pas dans ce tuto. A "vue de nez", c'est issue = 1 qui paraît différent des deux autres, qui semblent proches.

7. Comparaison : la variable de jugement est quantitative

Il arrive que le critère de jugement soit quantitatif, comme la durée de grossesse, la durée de séjour, la durée de vie, un score de qualité de vie ou de pathologie. Deux manières de l'analyser :

- Garder le caractère quantitatif en explorant bien la répartition numérique de cette variable (dite dépendante). Les variables d'exposition (appelées aussi explicatives ou indépendantes) peuvent être qualitatives, ou quantitatives utilisées comme telles ou transformées en qualitatives avec un ou des seuils
- Transformer la variable quantitative en qualitative avec un ou des seuils : ces seuils peuvent être définis à partir des données observées ou en référence avec des seuils scientifiquement établis (comme l'IMC < et ≥ 30 kg/m² pour l'obésité, ou la prématurité avec un âge gestationnel < 37 SA).

a. La variable de jugement et la variable d'exposition sont quantitatives

Le coefficient de corrélation (de - 1 à + 1 en passant par 0)

r de Pearson entre l'âge maternel et le poids de naissance

```
> cor(agemat,pn)
[1] 0.06227
```

En cas de réponse "NA" ou "pas de la même longueur", voir page 19 (suite commande `tapply ()`)

ρ **(rho) de Spearman** pour les petits effectifs

```
> cor (agemat,pn, method="spearman")
[1] 0.0338
```

Test du r à 0

```
> cor.test(agemat,pn)
Pearson's product-moment correlation
data:  agemat and pn
t = 1.1, df = 293, p-value = 0.3
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0523  0.1752
sample estimates:
    cor
0.06227
```

La droite de régression linéaire (y = ax + b)

```
> lm (pn~agemat)
Call:lm(formula = pn ~ agemat)
Coefficients:
(Intercept)       agemat
    3038.81         7.03
```

La formule de la droite s'écrit : pn (en grammes) = 7.03 * agemat (années) + 3038.81 (g)

On dit que "pn" est "expliqué" par l'âge maternel, qui est une variable "explicative", ou encore que "pn" est la variable dépendante et âge maternel la variable indépendante.

Le coefficient d'âge maternel est > 0 : plus la femme est âgée, plus le nouveau-né est gros.

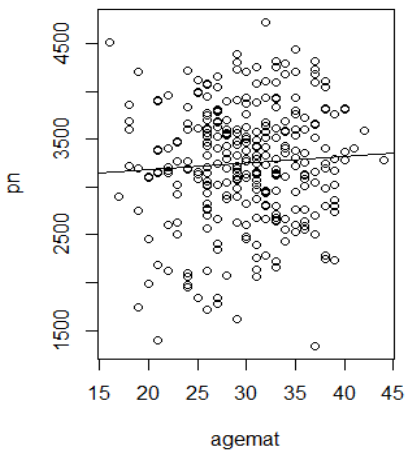
La pente de la droite est de 7.03 : est-ce différent de "0" ?

```
> reg_lin <- lm(pn~agemat)
> summary (reg_lin)
Residuals:
    Min      1Q  Median      3Q     Max
-1958.8  -337.3    21.4   414.3  1456.3
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3038.81     200.15   15.18   <2e-16 ***
agemat          7.03       6.58    1.07     0.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 614 on 293 degrees of freedom
Multiple R-squared:  0.00388,  Adjusted R-squared:
0.000477
F-statistic: 1.14 on 1 and 293 DF,  p-value: 0.286
```

La pente de la droite n'est pas significativement différente de 0 (p = 0.286).

Ce test de la pente correspond au même "p" que le coefficient de corrélation (0.3 noté).



Le graphe de la droite de régression (plot (x,y))

```
> plot (agemat,pn) # de type x,y
> abline (lm(pn~agemat)) # de type y~x
```

Pour présenter au mieux les graphes, voir document supplémentaire.

b. La variable de jugement est quantitative et la variables d'exposition est qualitative

La variable d'exposition peut être qualitative par nature ou après transformation avec des seuils. Par exemple, pour comparer le pn avec l'âge maternel, on peut analyser le pn selon un seuil (cutpoint) d'âge maternel à 30 ans (< 30 ans versus ≥ 30ans), ou deux seuils à 25 et 35 ans (< 25 ans, 25 – 35 ans et ≥ 35 ans). Dans le 1er cas, on compare la variable de jugement quantitatif avec les variables codées avec un test de t de Student ou de Welch, ou un test de Wilcoxon ou de Kruskall-Wallis. Dans le 2ème cas, avec une ANOVA (voir page 20).

Souvent, quand on compare, par exemple `pn` et `age maternel` par la régression linéaire, ou par une ANOVA avec trois classes, par exemple, on trouve des résultats significativement équivalents, mais pas toujours car établir des seuils c'est changer (ou réduire) l'information.

Exemple avec le poids de naissance selon l'âge maternel en 3 classes :

```
> ageclasse=cut(agemat,breaks=c(0,25,35,45))
> summary (ageclasse)
 (0,25] (25,35] (35,45]
     58     186      51
> anova(aov(pn~ageclasse))
Analysis of Variance Table
Response: pn
            Df   Sum Sq Mean Sq F value Pr(>F)
ageclasse    2 6.77e+05  338560     0.9   0.41
Residuals  292 1.10e+08  377225
```

A comparer avec p = 0.29 avec la régression linéaire

Autre exemple d'analyse : le lien entre taille de la mère et l'issue avec trois classes :

```
> taille_cl2=cut(taille,breaks=c(0,1.50,1.60,1.80))

> summary (taille_cl2)
  (0,1.5] (1.5,1.6] (1.6,1.8]      NA's
       11       109       165         10

> table (issue, taille_cl2)
     taille_cl2
issue (0,1.5] (1.5,1.6] (1.6,1.8]
    1       9        72        61
    2       2        37       104

> options (digits=2)
> tab11 <- prop.table (table(issue, taille_cl2),2)

> tab11
     taille_cl2
issue (0,1.5] (1.5,1.6] (1.6,1.8]
    1    0.82      0.66      0.37
    2    0.18      0.34      0.63

> addmargins(tab11,1)
     taille_cl2
issue (0,1.5] (1.5,1.6] (1.6,1.8]
  1      0.82      0.66      0.37
  2      0.18      0.34      0.63
  Sum    1.00      1.00      1.00
> chisq.test(issue, taille_cl2)
Pearson's Chi-squared test
X-squared = 27, df = 2, p-value = 1e-06
> anova (aov(taille~issue))
Analysis of Variance Table
Response: taille
            Df Sum Sq Mean Sq F value  Pr(>F)
issue        1  0.089  0.0891    21.1 6.4e-06 ***
Residuals  283  1.193  0.0042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Le taux de césarienne est de 82 % pour les femmes de moins de 1m50, de 66 % pour les femmes de 1m50 à moins de 1m60, et de 37 % pour les femmes de 1m60 et plus. Le χ^2 des 3 proportions est significatif, de même que l'ANOVA sur la taille en quantitatif. Pour les savants, on peut aussi tester le χ^2 de tendance ("χ^2 for trend") des trois proportions qui marque une progression ou une diminution progressive dans les trois classes `> prop.trend.test (x,y)`

8. Construire un tableau de comparaison avec une seule commande

On peut utiliser le package "`tableone`" pour comparer toutes les variables selon le critère de jugement et faire un tableau directement. Commande `CreateTableOne` (avec les 3 majuscules…)

Tableau par défaut avec les catégories des variables "factor" choisies par la commande

La commande choisie le codage le plus élevé (ici "2") ; passer par un changement de codage en 0 et 1 pour n'avoir que le "1"

```
> install.packages("tableone")
> library(tableone)

> expo <- c("primi", "modedecl","rpde12","taille")
> table1 <- CreateTableOne(vars=expo, factorVars=c("primi", "rpde12",
"modedecl"), data=decl_temoins, strata="issue")
> print (table1)
                     Stratified by issue
                      1            2            p      test
  n                    147          148
  primi = 2 (%)         33 (22.6)    93 (62.8)  <0.001
  modedecl (%)                                  <0.001
     1                  78 (53.1)    42 (28.4)
     2                  50 (34.0)    99 (66.9)
     3                  19 (12.9)     7 ( 4.7)
  rpde12 = 2 (%)       100 (68.0)   119 (80.4)   0.022
  taille (mean (SD)) 1.61 (0.07)  1.65 (0.06)  <0.00
```

Tableau par défaut avec toutes les catégories des variables "factor", plus une variable quanti

Avec `showAllLevels` dans `print ()`

```
> print (table1, showAllLevels=TRUE)
                     Stratified by issue
                      level 1            2            p
  n                          147          148
  primi (%)           1      113 (77.4)    55 (37.2)  <0.001
                      2       33 (22.6)    93 (62.8)

  modedecl (%)        1       78 (53.1)    42 (28.4)  <0.001
                      2       50 (34.0)    99 (66.9)
                      3       19 (12.9)     7 ( 4.7)

  rpde12 (%)          1       47 (32.0)    29 (19.6)   0.022
                      2      100 (68.0)   119 (80.4)

  taille (mean (SD))        1.61 (0.07)  1.65 (0.06)  <0.001
```

Tableau avec le choix des catégories des variables "factor"

```
> library(readxl)
```

Lire le fichier `decl_temoins.xslx`

```
install.packages ("tableone")
library(tableone)
attach (decl_temoins)
table (primi)
primi= replace (primi, primi==2,0)
table (primi)
primi <- as.factor (primi)
table (primi)
modedecl<-as.factor(modedecl)
rpde12 <- as.factor (rpde12)
primi <- as.factor (primi)
issue <- factor(issue, levels=c(1,2), labels=c("CS", "VBT"))
expo <- c("primi", "modedecl","rpde12","taille")
table1 <- CreateTableOne(vars=expo, factorVars=c("primi", "rpde12",
"modedecl"), data=decl_temoins, strata="issue")
print (table1)
```

Exporter le tableau créé en format Excel dans un fichier, pour pouvoir le copier dans Word ensuite

```
install.packages ("writexl")
library (writexl)
table2 <- print (table1, showAllLevels=TRUE)
```

Créer une data_frame

```
table2_df <- as.data.frame (print(table2, quote=FALSE, noSpaces=TRUE,
printToggle=FALSE))"
```

Préciser le chemin du fichier avec la barre oblique "/ (pas" \")

```
write_xlsx (table2_df, "/Chemin/tabl3.xlsx")
```

Voir pour importer ce résultat directement dans un traitement de texte ??? Y ajouter les OR ?????

	Issue = 1 (CS)	**Issue = 2 (VB)**	**p**
Primipares (%)	113 (77.4)	55 (37.2)	< 0.001
Mode de déclenchement (%)			
Prostaglandines	78 (53.1)	42 (28.4)	
Ocytocine	50 (34.0)	99 (66.9)	<0.001
Les deux	19 (12.9)	7 (4.7)	
Rupture poche des eaux > 12 h (%)	47 (32.0)	29 (19.6)	0.02
Taille (m)	1.61 ± 0.07	1.65 ± 0.06	< 0.001

9. Analyse multivariée

Voir un tutoriel sur le principe de l'analyse multivariée : site https://www.episure.fr/tutoriels

a. Régression multiple/ ANOVA

La variable de jugement est quantitative et les variables explicatives sont quantitatives.

C'est comme une régression linéaire, mais avec plusieurs variables.

La formule générale est $y = ax_1 + b_{x2} \ldots +$ constante, ou (idem) $y = a + b_1x_1 + b_2x_2 \ldots. b_nx_n$

```
> reg_multi <- lm ((pn~agemat+poids+taille+ag))
> summary(reg_multi)

Call:
lm(formula = (pn ~ agemat + poids + taille + ag))

Residuals:
     Min       1Q   Median       3Q      Max
-1102.47  -319.13   -27.23   309.54  1307.59

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7679.0568   990.2206  -7.755 4.07e-13
agemat         -0.1377     5.5648  -0.025  0.98028
poids           4.9536     2.0945   2.365  0.01896
taille       1471.1024   460.4973   3.195  0.00162
ag            207.4212    16.7295  12.399  < 2e-16

(Intercept) ***
agemat
poids       *
taille      **
ag          ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 440.9 on 205 degrees of freedom
  (86 observations effacées parce que manquantes)
Multiple R-squared:  0.4606,   Adjusted R-squared:  0.4501
F-statistic: 43.77 on 4 and 205 DF,  p-value: < 2.2e-16
```

Le poids de naissance est significativement lié au poids et taille de la mère, ainsi qu'à l'âge gestationnel, mais non avec l'âge maternel. Les coefficients de la formule "Estimate" sont > 0 et donc sont liés positivement avec le `pn` : plus la mère est grande, a un poids élevé ou un âge gestationnel élevé, plus le `pn` est élevé. Le R^2 - coefficient de détermination - correspond à une sorte de pourcentage de la variance expliqué par le modèle : ici 0.46, soit 46 % de la variabilité du poids de naissance est expliqué par les 4 variables. Le p-value global est significatif.

Pour vérifier la validité du modèle, de nombreuses analyses et commandes sont disponibles dans "R". Pas dans ce cours d'initiation (voir en particulier ce qu'on appelle "les résidus").

L'âge gestationnel étant forcément corrélé avec le poids de naissance (la valeur de "t" est la plus forte et de loin), on peut établir un nouveau modèle (une nouvelle formule) sans cette variable. Les coefficients des autres variables sont alors changés :

```
> reg_multi2 <- lm ((pn~agemat+poids+taille))
> summary(reg_multi2)

Call:
lm(formula = (pn ~ agemat + poids + taille))

Residuals:
     Min       1Q   Median       3Q      Max
-1871.27  -343.91    60.11   380.02  1242.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  593.498    896.740   0.662   0.5086
agemat         4.830      6.583   0.734   0.4638
poids          5.355      2.361   2.268   0.0241 *
taille      1318.190    555.405   2.373   0.0183 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 600.7 on 280 degrees of freedom
  (12 observations effacées parce que manquantes)
Multiple R-squared:  0.05413,  Adjusted R-squared:  0.04399
F-statistic: 5.341 on 3 and 280 DF,  p-value: 0.001363
```

On voit que les coefficients ont été modifiés, mais que surtout le R^2 est ridiculement petit (5.4 %) montrant, si besoin était, que c'était bien l'âge gestationnel qui "emportait" la majorité du lien +++

Note : l'introduction, dans un modèle de régression multiple, de variables qualitatives est a priori malaisée et n'est pas envisagée ici.

b. Régression logistique (LR)

La variable de jugement (dite aussi dépendante) est qualitative binaire (LR binaire) ou qualitative à plus de 2 codages (LR ordinale) non envisagée ici.

Les variables d'exposition sont qualitatives

Importer le fichier `decl_temoins.xlsx`

```
> View(decl_temoins)
> attach(decl_temoins)
> install.packages(("epiDisplay"))
> library(epiDisplay)
> primi <- as.factor (primi)
> summary(primi)
   1    2 NA's
 169  126    1

> rpde12 <- as.factor(rpde12)
> summary(rpde12)
  1   2
 76 220
```

```
> modedecl <- as.factor (modedecl)
> summary(modedecl)
  1   2   3
120 150  26
```

Il faut que la variable dépendante soit codée en « 0 » et «1 » (elle doit être numérique à ce stade), avec la fonction `replace`

```
> issue2 <- replace (issue, issue==2,0)   # avec 2 signes = à la suite
> issue2 <- as.factor (issue2)
> summary(issue2)
  0   1
149 147

> reg_log <- glm (issue2~primi+rpde12+modedecl, family=binomial)
> logistic.display (reg_log)

Logistic regression predicting issue2 : 1 vs 0
                crude OR(95%CI)    adj. OR(95%CI)     P(Wald's test)
primi: 2 vs 1   0.18 (0.11,0.29)  0.24 (0.14,0.42)  < 0.001
rpde12: 2 vs 1  0.51 (0.3,0.87)   0.74 (0.41,1.33)  0.31
modedecl: ref.=1
   2            0.26 (0.16,0.44)  0.45 (0.26,0.79)  0.005
   3            1.46 (0.57,3.76)  1.65 (0.61,4.44)  0.326

                P(LR-test)
primi: 2 vs 1   < 0.001

rpde12: 2 vs 1  0.309

modedecl: ref.=1 0.003
```

L'odds ratio brut de `primi` par rapport à `issue` est de 0.18 : la référence est sans doute mauvaise car on sait que la primiparité est un facteur de césarienne (1 "oui" par rapport à 0 "non"). Il faut donc changer la référence de l'odds ratio : mettre `primi` 1 vs 2. Idem pour `rpde12` :

```
> primi<- relevel(primi, ref="2")
> rpde12 <- relevel (rpde12, ref="2")
> primi <- as.factor (primi)
> rpde12 <- as.factor (rpde12)
> modedecl <- as.factor (modedecl)
> reg_log1 <- glm (issue2~primi+rpde12+modedecl, family=binomial)
> logistic.display(reg_log1)


                crude OR(95%CI)    adj. OR(95%CI)     P(Wald's test) P(LR-test)
primi: 1 vs 2   5.79 (3.47,9.66)  4.18 (2.41,7.24)  < 0.001        < 0.001

rpde12: 1 vs 2  1.95 (1.14,3.32)  1.35 (0.74,2.44)  0.328          0.327

modedecl: ref.=1                                                  0.004
   2            0.27 (0.16,0.44)  0.46 (0.26,0.81)  0.007
   3            1.46 (0.57,3.76)  1.65 (0.61,4.47)  0.323

Log-likelihood = -172.2227
No. of observations = 294
AIC value = 354.4454
```

Le sens est le bon : la **primiparité** et la **rupture de la poche des eaux** sont des facteurs associés à l'échec du déclenchement (par césarienne) avec des OR ajustés (ORa dans les articles) de 4.18 et 1.35 respectivement (ils sont inférieurs aux OR bruts). L'OR de `primi` est significativement différent de 1 (p<0.01), mais pas `rpde` (p = 0.31) qui était significatif en brut, mais pas en ajusté ++++
Pour le **mode de déclenchement** codée en 3 catégories : "1" pour maturation du col (prostaglandines), "2" pour l'oxytocine, et "3" pour les deux. L'OR de l'oxytocine par rapport aux prostaglandines est significatif (p=0.007) et inférieur à 1 (l'oxytocine donne moins de césariennes que les prostaglandines), mais pas quand on utilise les deux (p = 0.32) avec un OR > 1… Une explication obstétricale est facile : quand le col est fermé, on utilise d'abord les prostaglandines (et le risque de césarienne augmente) par rapport au col déjà ouvert pour lequel on utilise l'oxytocine.

On peut ajouter comme variables explicatives la taille de la mère (significative en univariée) en mètres.

```
> reg_log2 <- glm (issue2~primi+rpde12+modedecl+taille, family=binomial)
> logistic.display(reg_log2)

Logistic regression predicting issue2 : 1 vs 0
                    crude OR(95%CI)   adj. OR(95%CI)    P(Wald's test) P(LR-
test)
primi: 1 vs 2       6.08 (3.6,10.26)  4.07 (2.28,7.26)  < 0.001        < 0.001
rpde12: 1 vs 2      1.88 (1.1,3.23)   1.31 (0.7,2.44)   0.394          0.394
modedecl: ref.=1                                                       0.001
   2                0.25 (0.15,0.41)  0.37 (0.2,0.67)   0.001
   3                1.3 (0.5,3.38)    1.35 (0.48,3.82)  0.568

taille (cont. var.) 0 (0,0.01)        0 (0,0.01)        < 0.001        < 0.001

Log-likelihood = -156.3137
No. of observations = 285
AIC value = 324.6273
```

La taille est significative, mais l'unité "en mètres" entraîne beaucoup trop de décimales après la virgule.
On passe à la taille en cm

```
> taille = taille*100
> reg_log2 <- glm (issue2~primi+rpde12+modedecl+taille, family=binomial)
> logistic.display(reg_log2)
Logistic regression predicting issue2 : 1 vs 0

                    crude OR(95%CI)   adj. OR(95%CI)    P(Wald's test) P(LR-
test)
primi: 1 vs 2       6.08 (3.6,10.26)  4.07 (2.28,7.26)  < 0.001        < 0.001

rpde12: 1 vs 2      1.88 (1.1,3.23)   1.31 (0.7,2.44)   0.394          0.394

modedecl: ref.=1                                                       0.001
   2                0.25 (0.15,0.41)  0.37 (0.2,0.67)   0.001
   3                1.3 (0.5,3.38)    1.35 (0.48,3.82)  0.568

taille (cont. var.) 0.92 (0.89,0.96)  0.92 (0.88,0.96)  < 0.001        < 0.001
```

L'OR ajusté de la taille en quantitative (en cm) est < 1 et très significatif : plus la taille augmente, ~~moins~~ l'issue se fait en césarienne +++. On peut même dire que à chaque centimètre en plus, le risque de césarienne diminue de 8 % (100 – 0.92)... sous condition de linéarité ++ (pas dans ce tuto : pour info, faire des classes de taille et voir si les ORa successifs sont "à peu prés" linéaires…).

On peut vérifier si c'est linéaire en faisant 4 classes de taille :

```
> classe_t3 <- cut (taille, breaks=c(145,150,155,160,185))
> classe_t3 <- as.factor (classe_t3)
> summary (classe_t3)
(145,150] (150,155] (155,160] (160,185]      NA's
       10        28        81       165        11
> classe_t3 <- relevel (classe_t3, ref="(160,185]")
> reg_log5 <- glm (issue2~primi+rpde12+modedecl+classe_t3, family=binomial)
> logistic.display(reg_log5)

Logistic regression predicting issue2 : 1 vs 0

                 crude OR(95%CI)    adj. OR(95%CI)    P(Wald's test) P(LR-test)
primi: 1 vs 2    6.03 (3.57,10.17)  4.43 (2.42,8.09)   < 0.001        < 0.001

rpde12: 1 vs 2   1.84 (1.07,3.16)   1.47 (0.78,2.78)   0.232          0.231

modedecl: ref.=1
   2             0.25 (0.15,0.42)   0.37 (0.2,0.68)    0.002          0.002
   3             1.32 (0.51,3.43)   1.32 (0.45,3.85)   0.611

classe_t3: ref.=(160,185]                                            < 0.001
   (145,150]     6.82 (1.4,33.16)   4.63 (0.89,24.14)  0.069
   (150,155]     6.25 (2.4,16.27)   7.29 (2.53,20.98)  < 0.001
   (155,160]     2.75 (1.59,4.76)   3.68 (1.93,7.03)   < 0.001

Log-likelihood = -150.6121
No. of observations = 284
AIC value = 317.2242
```

On a pris comme référence les tailles "grandes" des mères (160 cm et +). Les OR de césarienne sont tous > 1 pour les tailles inférieures à 160 cm et significatives pour les femmes avec une taille > 150 cm. Ce n'est pas significatives pour les femmes < 150 cm (mais il n'y en que 10).

Le modèle est « complet » avec toutes les variables ajustées les unes par rapport aux autres. Une méthode « **pas-à-pas ascendant** » permet de déterminer la variable d'exposition la plus importante, suivie progressivement des suivantes…

```
> reg_step <- step(reg_log1, dir="forward")
> summary (reg_step)
> reg_step <- step(reg_log3, dir="forward")
Start:  AIC=321.96
issue2 ~ primi + rpde12 + modedecl + classe_taille
> summary (reg_step)
```

Résultats non montrés. Pour établir une courbe ROC après une régression logistique, voir le document supplémentaire.

a. Les autres analyses possibles dans le 1er contact avec R : voir supplément ++

Analyse de survie

Description, comparaison univariée, comparaisons multivariées (Modèle de Cox)
Package = "survival"
Commandes = `surv (), surfit (), survdiff (), coxph ()`
Courbes de survie : Package = "surwiner", et ggsurvplot ( ) et `ggplot2`

Tests diagnostiques et courbes ROC

Packages "epiR" et commande `epi.test`

Propensity score (score de propension)
Possible avec de multiples méthodes

b. Rappel pour terminer

Il y a encore loin de l'analyse vers l'écriture et la publication. Il faut d'abord mettre en forme les résultats dans des tableaux et dans le texte pour pouvoir être lus en quelques minutes. Les tableaux répondent à des règles précises qui peuvent varier selon les revues (qui facilitent plus ou moins bien la lecture).
Voir **https://www.episure.fr/tutoriels** la page "Comment faire des tableaux" et le résultat de l'étude avec seulement les tableaux.

==========

Table des matières

==========