Guide d'utilisation de R pour les calculs d'indicateurs quantitatifs en épidémiologie.

Sommaire

Karine CHALVET-MONFRAY, octobre 2016

Introduction

L'utilisation de ce guide nécessite une certaine maîtrise du langage R supposée acquise dans le cadre de l'enseignement de biostatistique de base. Ce guide a pour objectif de présenter les principales fonctions utiles pour réaliser des études de base en épidémiologie tel que le calcul d'un taux de prévalence, d'un taux d'incidence, d'un Risque Relatif ou d'un Odds Ratio avec ou sans ajustement avec les intervalles de confiance associés. Pour cela vous pourrez vous appuyer sur les méthodes vues auparavant, à partir de quelques exemples, exemples qui font parfois appels à des données disponibles sur le réseau de l'ENVL, à tester par vous-mêmes.

Il est auparavant intéressant de présenter ou de revoir quelques points importants.

- **Point 1.** En épidémiologie, on a souvent des études qui portent sur plusieurs sujets pour lesquels on a relevé plusieurs variables qualitatives ou quantitatives. Afin de pouvoir utiliser les données recueillies, une solution facile est de réaliser avec un tableur de type Excel, un tableau où la première ligne correspond aux entêtes des différentes variables qui on été mesurées et où chaque ligne suivante correspond à l'observation d'un sujet donné. S'il existe des données manquantes, il faudra indiquer dans les cases correspondantes « NA » pour « Not Available ». Une fois que le fichier est constitué, il faut le sauvegarder sous le format « Texte(séparateur :tabulation)(*.txt) » (Cf. Poly de guide R pour les statistiques de base).
- **Point 2.** Pour lire les données ainsi sauvées au format « txt », et obtenir un jeu de données reconnu par R, on peut utiliser la fonction `read.table()` (Cf. Poly de guide R pour les statistiques de base).

  `donnees <- read.table("nomfichier.txt",header=TRUE)`

  Ill est possible dans cet exemple, au lieu d'utiliser le nom « donnees », de nommer le résultat de la fonction `read.table()` comme on veut : ex. « d », ou « toto » voire « tarteaupomme ».

  `tarteaupomme <- read.table("nomfichier.txt",header=TRUE)`
- **Point 3.** Afin de ne visualiser qu'une partie du jeu de données, ce qui peut être intéressant en cas d'un nombre important de données, il suffit de préciser la ou les ligne(s) et la ou les colonne(s) à visualiser de la manière suivante : nomdujeudedonnees[numero(s) des lignes, numéro(s) des colonnes] (Cf. Poly de guide R pour les statistiques de base).

  `tarteaupomme[1:2,]`

  Permet de visualiser les deux premières lignes pour l'ensemble des colonnes du jeu de données tarteaupomme.
- **Point 4.** Si l'on souhaite avoir une vue globale de la structure ou un résumé du jeu de donnée les fonctions qui permettent de les obtenir très rapidement sont `str()` et `summary`().

  ```
  str(tarteaupomme)
  summary(tarteaupomme)
  ```
- **Point 5.** Si l'on souhaite, ne sélectionner qu'une partie des données correspondants à certains sujets correspondant à une (des) condition(s), on utilisera la fonction `subset()` (Cf. Poly de guide R pour les statistiques de base). Si l'on veut choisir les sujets mâles du jeu de données « tarteaupomme » pour réaliser un nouveau jeu de données

  `male <- subset(tarteaupomme,sexe=='M')`

  Si l'on souhaite étudier la variable poids pour uniquement le nouveau jeu de données : « male », il faudra utiliser le terme : male$poids ; au contraire si l'on souhaite étudier la variable poids pour l'ensemble du jeu de données tarteaupomme, il faudra utiliser le terme : tarteaupomme$poids.
- **Point 6.** Si l'on souhaite diviser un jeu de données en plusieurs sous-jeux de données correspondant à différentes valeurs d'un facteur, on peut utiliser la fonction `split()`.

  split(tarteaupomme,race)

Où tarteaupomme représente le jeu de données et race est le facteur utilisé pour diviser le jeu de données. On aboutit ainsi à une liste de sous jeux de données correspondants à chaque race.

- **Point 7.** La fonction `nrow()` permet de connaître le nombre de ligne d'un vecteur ou d'un jeu de données. Ainsi pour connaître le nombre de sujets dans un jeu de données, il suffit d'appliquer `nrow()` au jeu de données.

```
nrow(tarteaupomme)
```

- **Point 8.** Il est nécessaire de convertir les dates pour pouvoir les comparer entre elles. Il est alors pratique de faire appel à une fonction `as.Date` converti une date en une variable numérique qui est le nombre de jours depuis le 1er janvier 1970 mais qui reste en apparence une date.

```
as.Date("03/07/2007","%d/%m/%Y")
[1] "2007-07-03".
```

La preuve que la date d'origine est 1er janvier 1970 :

```
(as.Date(or,"%d/%m/%Y")->x)
[1] "1970-01-01"
as.numeric(x)
[1] 0
```

Il est nécessaire d'indiquer dans les arguments de la fonction le format utilisé. Celui utilisé ci-dessus correspond au format : %d/%m/**%Y** dans lequel %d et %m correspondent respectivement au jour et au mois. **Attention** %Y correspond à l'année écrite en 4 chiffres et %y correspond à l'année écrite en 2 chiffres. **En cas d'erreur entre les deux formats, les calculs seront faux et aucun message d'erreur ne sera donné**.

Attention, selon les versions, le format par défaut des dates est modifié : il peut être de la forme jj.mm.aaaa. (ex. le 3 juillet 2007 s'écrit 03.07.2007). Dans ce cas, il faut modifier le format pour la fonction `as.Date()` ainsi :

```
As.Date("03.07.2007","%d.%m.%Y")
```

1 Calcul du taux de prévalence

Dans le cas où nous avons un suivi de cohorte, il est nécessaire dans un premier temps de calculer le nombre de sujets exposés à un moment donné ainsi que le nombre de sujets malades à ce moment donné.

1.1 Données nécessaires

Les données initiales doivent être sous la forme d'un tableau de 4 colonnes où chaque ligne correspond à un sujet de l'étude et les colonnes correspondent à :

• La date d'origine (DO) de chaque sujet (date où le sujet est entré dans l'étude).

• La date de début de maladie (DM) qui peut avoir lieu avant la date d'origine. Si le sujet n'a pas contracté la maladie avant la date des dernières nouvelles, alors le terme NA doit être indiqué pour la date de début de maladie,

• La date de fin de maladie (DFM). Si le sujet n'a pas contracté la maladie avant la date des dernières nouvelles, alors le terme NA doit être indiqué pour la date de fin de maladie. Si le sujet n'est pas guéri avant la date des dernières nouvelles, alors la date des dernières nouvelles (DDN) doit être indiquée pour la date de fin de maladie,

• La date des dernières nouvelles (DDN) de chaque sujet (date où le sujet est sorti de l'étude).

Si on visualise uniquement les trois premières lignes de la base de données on obtient un résultat du type de l'exemple ci-dessous :

```
d[1:3,]
          DO         DDN          DM         DFM
1 27/10/2007 31/01/2008        <NA>        <NA>
2 23/10/2006 15/11/2007 07/10/2007 15/11/2007
3 12/01/2007 15/06/2007 24/02/2007 28/02/2007
```

ici d est le résultat de la fonction read.table qui a permis de lire le jeu de données.
dans cet exemple on a choisi le format de date courte que Windows Vista choisit par défaut en version Française c'est-à-dire **deux chiffres pour les jours/deux chiffres pour les mois/ quatre chiffres pour l'année**.
ATTENTION : R ne comprend pas ces données comme des dates ; d'où la nécessité de faire appel à la fonction `as.Date()`.

1.1.1 Définition des dates du jeu de données de l'étude

On convertit les autres dates du jeu de données date de maladie (DM) et date de dernières nouvelles (DDN), date d'origine (DO) au format numérique

```
d$DDNn<-as.Date(d$DDN,"%d/%m/%Y")
d$DMn<-as.Date(d$DM,"%d/%m/%Y")
d$DFMn<-as.Date(d$DFM,"%d/%m/%Y")
d$DOn<-as.Date(d$DO,"%d/%m/%Y")
```

la fonction `as.Date()` convertit une date en une variable numérique qui correspond au nombre de jour depuis le 1er janvier 1970 (Cf. point 8 de l'introduction) permettant ainsi de pouvoir comparer les dates entre elles.

Les nouvelles dates ont été automatiquement rajoutée au jeu de données (d) car l'attribution était de la forme d$.....

1.1.2 Définition de la date de point (date où l'on calcule de taux de prévalence)

```
DP<-"01/04/2007"
DPn<-as.Date(DP,"%d/%m/%Y")
```

Ici, la date de point correspond au 1er avril 2007.

1.2 Calcul du nombre de sujets exposés (présents à la date de point)

On définit le sous jeu de données qui conserve les sujets qui sont entrés dans l'étude avant la date de point et qui sont sortis de l'étude après la date de point(c-à-d ceux pour lesquelles la date de d'origine est inférieure à la date de point et la date de dernières nouvelles est supérieure à la date de point).

```
dd<-subset(d, DOn<=DPn & DDNn>=DPn)
```

La fonction `subset()` permet d'établir un nouveau jeu de données répondant aux conditions décrites dans les arguments de la fonction (Cf. point 4 de l'introduction). Ici, les conditions sont que le format numérique de la date d'origine soit inférieur ou égale au format numérique de la date de point et que la date de dernière nouvelle soit inférieur ou égale au format numérique de la date de point.
pour rappel les opérateurs logiques : `!` (ne pas), `&` (et), `|` (ou),… peuvent être utilisé pour définir les conditions du `subset()`.
On compte le nombre de lignes, c'est-à-dire de sujets, qu'il y a dans le sous jeu de données dd (les présents à la date de point).

```
N<-nrow(dd)
```

(Cf. point 6 de l'introduction).

```
N
[1] 7
```

1.3 Calcul du nombre de sujets malades (à la date de point)

On définit un nouveau sous jeu de données qui conserve uniquement les sujets du précédent sous jeu de données (les présents à la date de point)qui ont contracté la maladie avant la date de point et qui ont terminé la maladie (guérie ou sortie de l'étude) après la date de point.

```
mdd<-subset(dd,DMn<=DPn & DFMn>DPn)
```

On compte le nombre de lignes, c'est-à-dire de sujets, qu'il y a dans le sous jeu de données mdd .

```
Nm<-nrow(mdd)
```

(Cf. point 6 de l'introduction).

```
Nm
[1] 2
```

1.4 Calcul du taux de Prévalence et de son intervalle de confiance

Calculer l'intervalle de confiance du taux de Prévalence revient à calculer l'intervalle de confiance d'une fréquence et ceci se fait avec la fonction `binom.test()`.

```
binom.test(Nm,N,p=0,alternative="two.sided",conf.level=.95)
```

où Nm est le nombre de malades observés, N est la taille de l'échantillon.

la fonction teste systématiquement si la différence entre la fréquence observée et la fréquence théorique spécifiée dans l'argument p est significative. Dans notre cas, il faut indiquer une valeur arbitraire à l'argument p car aucun test n'est réalisé.

Indiquer "two.sided" pour un intervalle de confiance bilatéral (valeur par défaut), "less" pour un intervalle de confiance unilatéral de la forme [0;t], "greater" pour un intervalle de confiance unilatéral de la forme [t;1].

Pour un risque α=5%, indiquer conf.level=0.95 et pour un risque α=1%, indiquer conf.level=0.99

- **Exemple 1** où le nombre de malade est de 3 pour un échantillon de 100 sujets et que l'on souhaite un intervalle de confiance bilatéral à 95%

```
binom.test(3,100,p=0,alternative="two.sided",conf.level=.95)

        Exact binomial test

data:  3 and 100
number of successes = 3, number of trials = 100, p-value < 2.2e-
16
alternative hypothesis: true probability of success is not equal
to 0
95 percent confidence interval:
 0.006229972 0.085176053
sample estimates:
probability of success
                  0.03
```

Le taux de Prévalence est de 3% avec un intervalle de confiance bilatéral à 95% de [0.623% ; 8.52%].

- **Exemple 2** où le nombre de malade est de 0 pour un échantillon de 100 sujets et que l'on souhaite calculer le seuil au dessous duquel on est sûr à 95% que se trouve le taux de Prévalence théorique. Ceci revient à calculer un intervalle de confiance unilatéral de type [0;t] à 95%.

```
binom.test(0,100,p=0,alternative="less",conf.level=.95)

        Exact binomial test

data:  0 and 100
number of successes = 0, number of trials = 100, p-value = 1
alternative hypothesis: true probability of success is less than
0
95 percent confidence interval:
```

```
 0.00000000 0.02951305
sample estimates:
probability of success
                     0
```

Le taux de Prévalence seuil vaut 2.95%.

C'est-à-dire que, que l'on est sûr à 95% que la prévalence est inférieure à 2.95%. Plus précisément, l'intervalle de confiance est calculé de telle manière qu'en moyenne pour différents échantillons de 100 sujets, l'intervalle de confiance unilatéral [0 ;t] contient la valeur théorique dans 95% des cas.

2 Calcul du taux d'Incidence

2.1 Calcul du Temps de Participation et du Nombre de sujet-temps

2.1.1 Données nécessaires

Les données initiales doivent être sous la forme d'un tableau de 3 colonnes où chaque ligne correspond à un sujet de l'étude et les colonnes correspondent à :

- La date d'origine (DO) de chaque sujet (date où le sujet est entré dans l'étude).
- La date de début de maladie (DM) qui peut avoir lieu avant la date d'origine. Si le sujet n'a pas contracté la maladie avant la date des dernières nouvelles, alors le terme NA doit être indiqué pour la date de début de maladie,
- La date des dernières nouvelles (DDN) de chaque sujet (date où le sujet est sorti de l'étude).

Donc même si il y a guérison, le sujet est considéré comme « malade » à partir du moment où il a contracté la maladie.

Si on visualise uniquement les deux premières lignes de la base de données on obtient un résultat du type de l'exemple ci-dessous :

```
d[1:2,]
          DO          DM         DDN
1 02/12/2006        <NA> 02/04/2008
2 23/10/2006 24/01/2007 15/11/2007
```

ici d est le résultat de la fonction read.table qui a permis de lire le jeu de données.

dans cet exemple on a choisi le format de date courte que Windows Vista choisit par défaut en version Française c'est-à-dire **deux chiffres pour les jours/deux chiffres pour les mois/ quatre chiffres pour l'année**.

ATTENTION : R ne comprend pas ces données comme des dates ; d'où la nécessité de faire appel à la fonction `as.Date()`.

2.1.2 Définition des dates et du sous jeu de données de l'étude

On convertit les autres dates du jeu de données date de maladie (DM) et date de Dernières nouvelles (DDN), date d'origine (DO) au format numérique

```
d$DDNn<-as.Date(d$DDN,"%d/%m/%Y")
d$DMn<-as.Date(d$DM,"%d/%m/%Y")
d$DOn<-as.Date(d$DO,"%d/%m/%Y")
```

la fonction `as.Date()` convertit une date en une variable numérique qui correspond au nombre de jour depuis le 1er janvier 1970 (Cf. point 8 de l'introduction) permettant ainsi de pouvoir comparer les dates entre elles.

Les nouvelles dates ont été automatiquement rajoutée au jeu de données (d) car l'attribution était de la forme d$.....

Si besoin on définit le sous jeu de données qui ne conserve pas les sujets qui sont malades avant la date d'origine (c-à-d ceux pour lesquelles la date de maladie est inférieure à la date d'origine).

```
dd<-subset(d,(DMn>DOn|is.na(DM)))
```
La fonction `subset()` permet d'établir un nouveau jeu de données répondant aux conditions décrites dans les arguments de la fonction (Cf. point 4 de l'introduction). Ici, les conditions sont que le format numérique de la date de maladie soit supérieure ou égale au format numérique de la date de d'origine ou que la date de la maladie ne soit pas disponible (NA)
pour rappel les opérateurs logiques : `!` (ne pas), `&` (et), `|` (ou),… peuvent être utilisé pour définir les conditions du `subset()`.

On définit date de fin (DFn) pour chaque sujet qui est la valeur minimum entre DDN et DM.
```
DFn<-pmin(dd$DDNn,dd$DMn,na.rm=T)
```
Pour cela on utilise la fonction `pmin()` qui renvoie la valeur minimum pour chaque ligne c'est-à-dire pour chaque sujet de l'étude. `Na.rm=T` permet de ne pas obtenir la valeur « NA » quand dd$DMn = « NA » pour les sujets pas (encore) malade.

On définit date d'entrée (DEn) qui est la DO préalablement convertie en format numérique.
```
DEn<-dd$DOn
```

2.1.3 Calcul des Temps de Participation et du Nombre de sujet-temps

```
TP<-DFn-DEn
```
On définit le temps de participation de chaque sujet (TP) qui est la différence entre DFn et DEn.
```
Nst<-sum(TP)
```
On définit le nombre de sujet-temps (Nst) qui est la somme des temps de participation.
```
Nst
[1] 6.551677
```

2.2 Calcul du Nombre de nouveaux malades

On définit un nouveau sous jeu de données qui conserve uniquement les sujets du précédent sous jeu de données (les sujets qui ne sont pas malades avant à la date d'origine) qui sont malades après la date de d'origine.
```
mdd<-subset(dd,DMn>DOn)
```
On compte le nombre de lignes, c'est-à-dire de sujets, qu'il y a dans le sous jeu de données mdd.
```
Nnm<-nrow(mdd)
```
(Cf. point 6 de l'introduction).
```
Nnm
[1] 4
```

2.3 Calcul du taux d'Incidence (I) et de son intervalle de confiance

Calculer l'intervalle de confiance du taux d'incidence revient à calculer l'intervalle de confiance d'un nombre d'événements se réalisant dans un temps donné. Or le nombre d'événements se réalisant dans un temps donné, si ces événements sont indépendants entre eux, suit une loi de Poisson. Son intervalle de confiance peut être calculé avec la fonction `pois.exact()`. Cependant, cette fonction ne se trouve pas dans la bibliothèque de base. Il est nécessaire de faire appel à une autre bibliothèque d'outils complémentaires : « epitools » qui contient d'autres outils appliqués à l'épidémiologie.
```
library(epitools)
pois.exact(Nnm, Nst, conf.level = 0.95)
```
#ou Nnm et Nst sont respectivement le nombre de nouveaux malades et le nombre de Sujet-Temps défini aux paragraphes 2.1.3 et 2.2.
Pour un risque α=0.05 indiquer conf.level=0.95 et pour un risque α=0.01, indiquer conf.level=0.99

- Pour l'exemple précédent

pois.exact(4, 6.551677, conf.level = 0.95)

	x	pt	rate	lower	upper	conf.level
1	4	6.551677	**0.6105307**	**0.1663527**	**1.563201**	0.95

On obtient le taux d'incidence 0.610 cas par un sujet-année avec un intervalle de confiance à 95% [0.166 ; 1.563]. On remarque ici que la borne 1 peut être dépassée car le taux d'incidence n'est pas une fréquence.

3 Calcul du taux d'Incidence pour une période donnée

Il convient tout d'abord d'enlever du jeu de données les sujets qui sont tombés malades ou qui sont sortis de l'étude avant la Date de Début de période ainsi que ceux qui sont rentrés dans l'étude après la date de la fin de la période.

3.1 Calcul du Temps de Participation et du nombre de Sujets-Temps pour une période donnée

Il suffit de procéder comme pour le 2.1. avec pour différence :

- la définition d'une date de Début de période (DD) et sa conversion au format numérique DDn.
- la définition d'une date de fin de Période (DP) et sa conversion au format numérique DPn.
- la date d'entrée (DEn) est alors la valeur maximum entre DOn et DDn.
- la date de fin (DFn) est alors la valeur minimum entre DDNn, DM net DPn)

Le Temps de Participation (TP) est toujours la différence entre la date Finale (DFn) et la date d'entrée (DEn).

3.2 Calcul du Nombre de nouveaux malades pour une période donnée

On procède comme pour 2.2. pour créer un nouveau jeu de données qui ne conserve cette fois-ci que les sujets dont la date de maladie se situe entre la date de Début de période et la date de fin de période.

3.3 Calcul du taux d'Incidence (I) et de son intervalle de confiance pour une période donnée

Il faut utiliser la même fonction `pois.exact()` présentée au paragraphe 2.3.

4 Calcul du Risque Relatif et de son intervalle de confiance et test de comparaison du Risque Relatif à 1

Pour calculer un Risque relatif et son intervalle de confiance, il est nécessaire de faire appel à une bibliothèque d'outils complémentaires : « epitools ».

```
library(epitools)
riskratio(exposition,maladie)
```

Rentrer la variable correspondant à l'exposition en premier terme ;

Rentrer la variable correspondant à la maladie en second terme ;

S'il existe plus de deux niveaux d'exposition, la fonction `riskratio()` calculera un RR par niveau d'exposition et prendra pour niveau de référence celui dont l'intitulé est le premier dans l'ordre alphabétique ou dans l'ordre croissant.

La méthode proposée par défaut pour calculer l'intervalle de confiance est la méthode de Wald qui est basée sur l'approximation par la loi Normale du ln(RR). Ceci n'est valable que si les effectifs sont grands. Il n'y a pas de valeurs « seuil » à partir desquelles on est sûr que l'on puisse employer cette approximation. On peut toutefois indiquer que dans les conditions suivantes :

- RR>10 ou ;
- Quand l'effectif d'un groupe d'exposition < 100 ou ;
- Quand les conditions du χ^2 ne sont pas vérifiées

il est hasardeux d'utiliser cette approximation et qu'il faut mieux demander conseils auprès de statisticiens.

- **Exemple 1**. Lors d'une intoxication alimentaire durant un banquet, une enquête est réalisée sur l'ensemble de la population présente au banquet. On souhaite vérifier si le fait de manger du Thon mayonnaise (codé par la variable binomiale « thonmayo » dans la base de données) est lié au fait d'avoir été intoxiqué (codé par la variable binomiale « Intox » dans la base de données). Comme l'ensemble de la population a été enquêtée, il est possible de calculer le Risque Relatif.

```
riskratio(d$thonmayo,d$Intox)
$data
         Outcome
Predictor non oui Total
    non   223  21   244
    oui    30 178   208
    Total 253 199   452

$measure
         risk ratio with 95% C.I.
Predictor estimate    lower    upper
      non 1.000000       NA       NA
      oui 9.943223 6.581235 15.02267

$p.value
         two-sided
Predictor midp.exact fisher.exact   chi.square
      non         NA           NA           NA
      oui          0 1.806235e-67 1.170244e-60

$correction
[1] FALSE

attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
>
```

Dans cet exemple, les deux variables ne sont pas indépendantes (P=1.80e-67). Le risque relatif du thon mayonnaise pour l'intoxication est de 9.94 avec un intervalle de confiance de [6.58 ; 15.02]. Les personnes qui ont mangé du thon mayonnaise ont 9.94 fois plus de chance d'avoir une intoxication que les personnes qui n'ont pas mangé de thon mayonnaise. On vérifie que les conditions d'utilisation de l'approximation de la loi normale pour l'estimation du RR sont vérifier.(i) RR<10, (ii) tous les groupes ont des effectifs supérieurs 100 (iii) les conditions du test du chi2 sont vérifiées : tous les effectifs calculés > 5

```
chisq.test(d$thonmayo,d$Intox)$expected
        Intox
thonmayo      non       oui
     non 136.5752 107.42478
     oui 116.4248  91.57522
```

- **Exemple 2.** En élevage laitier, on souhaite savoir comment le taux de gestation à 56 jours dépend du délai vêlage-début du traitement d'induction de l'ovulation. Pour cela une étude de cohorte a été mise en route ou plusieurs groupes on été constitués : Groupe a : délai>77 jours, groupe b : délai compris entre 64 et 76 jours, groupe c : délai compris entre 50 et 63 jours, groupe d : délai < 50 jours. On souhaite indiquer les différents risques relatifs en prenant le groupe a comme groupe de référence.
  Le groupe « a.>77d » est pris automatiquement comme groupe de référence, car c'est l'intitulé qui arrive le premier par ordre alphabétique.

```
riskratio(d$Exposition,d$pregnancy_rate_56d)
$data
          Outcome
Predictor  non oui Total
  a.>77d   104 254   358
  b.64-76d 100 248   348
  c.50-63d 115 240   355
  d.<50d   135 219   354
  Total    454 961  1415

$measure
          risk ratio with 95% C.I.
Predictor   estimate     lower     upper
  a.>77d   1.0000000        NA        NA
  b.64-76d 1.0044348 0.9142779 1.1034821
  c.50-63d 0.9528668 0.8640277 1.0508403
  d.<50d   0.8719472 0.7848195 0.9687474

$p.value
          two-sided
Predictor  midp.exact fisher.exact chi.square
  a.>77d           NA           NA         NA
  b.64-76d 0.92692639   0.93401756 0.92652912
  c.50-63d 0.33473269   0.37188214 0.33313663
  d.<50d   0.01043793   0.01113162 0.01026391

$correction
[1] FALSE

attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

Dans cet exemple, on a trois risques relatifs correspondants respectivement aux trois groupes (b, c et d) qui sont comparés au groupe de référence (a). Ici seul le groupe d à un RR significativement différent de 1. Pour le groupe d (en comparaison avec le groupe a), le risque relatif est de 0.872 avec un intervalle de confiance de [0.785 ; 0.969]. En d'autres termes, les vaches du groupe a ont en moyenne 1.147 (1/0.872= 1.146858) fois plus de chance d'être gestante que les vaches du groupe d'ou encore le risque d'être gestante pour les vaches du groupe d est 0.872 fois le risque d'être gestante pour les vaches du groupe a. On vérifie que les conditions d'utilisation de l'approximation de la loi normale pour l'estimation du RR sont vérifier.(i) RR<10, (ii) tous les groupes ont des effectifs supérieurs 100 (iii) les conditions du test du chi2 sont vérifiées sont vérifiées : tous les effectifs calculés > 5

```
chisq.test(Exposition,pregnancy_rate_56d)$expected
          pregnancy_rate_56d
Exposition      non      oui
  a.>77d   114.8636 243.1364
  b.64-76d 111.6551 236.3449
```

```
c.50-63d 113.9011 241.0989
d.<50d   113.5802 240.4198
```

5 Calcul de l'Odds Ratio et de son intervalle de confiance ainsi que le test de comparaison de l'Odds Ratio à 1

Pour calculer un Odds Ratio et son intervalle de confiance, il est nécessaire de faire appel à une bibliothèque d'outils complémentaires : « epitools ».

```
library(epitools)
oddsratio(exposition,maladie)
```

Rentrer la variable correspondant à l'exposition en premier terme ;
Rentrer la variable correspondant à la maladie en second terme ;
S'il existe plusieurs niveaux d'exposition, la fonction `oddsratio()` calculera un OR par niveau d'exposition et prendra pour niveau de référence celui dont l'intitulé est le premier dans l'ordre alphabétique.
Par défaut l'estimation de l'intervalle de confiance n'est pas réalisé par la méthode de Wald basée sur l'approximation de la loi normale mais par la méthode exacte de mid-p ; elle est plus adaptée pour les effectifs petits et pour des OR élevés.

- **Exemple 1** Si l'on reprend les variables et les valeurs de l'exemple 1 du paragraphe 4. On obtient des résultats différents. Car dans cet exemple, le risque d'être malade est élevé et donc l'OR et le RR sont très différents. Par ailleurs, il n'est absolument pas intéressant de calculer un OR quand on peut calculer directement un RR.

```
oddsratio(d$thonmayo,d$Intox)
$data
         Outcome
Predictor non oui Total
    non   223  21   244
    oui    30 178   208
    Total 253 199   452

$measure
         odds ratio with 95% C.I.
Predictor estimate    lower    upper
      non  1.00000       NA       NA
      oui 61.56871 34.81183 114.4274

$p.value
         two-sided
Predictor midp.exact fisher.exact   chi.square
      non         NA           NA           NA
      oui          0 1.806235e-67 1.170244e-60

$correction
[1] FALSE
attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

  Ainsi, ici l'OR vaut 61.57 [34.81 ;114.4] ce qui est différent des résultats obtenus pour le RR =9.94 [6.58 ; 15.02]. Il est à noter qu'ici l'estimation de l'intervalle de confiance n'est pas basée sur l'approximation de la loi normale et il est adapté pour les effectifs petits et pour des OR élevés.

- **Exemple 2** Dans une enquête cas-témoin à l'abattoir, où il y avait 92 bovins parasités par des paramphistomes et 70 bovins indemnes de paramphistomes, on a recherché la race des bovins. Les races ont été regroupées en 4 catégories : Charolais, Limousine, Montbéliarde et

autres. La catégorie « autres » est automatiquement choisie comme groupe de référence pour calculer les différents OR, car c'est l'intitulé qui arrive le premier par ordre alphabétique.

```
oddsratio(d$Race,d$Paramphistome)
```

```
$data
             Outcome
Predictor      Non Oui Total
  autres        14  14    28
  Charolais     27  68    95
  Limousine     15   3    18
  Montbéliarde  14   7    21
  Total         70  92   162

$measure
             odds ratio with 95% C.I.
Predictor       estimate      lower     upper
  autres       1.0000000         NA        NA
  Charolais    2.4956681 1.04046998 6.0292052
  Limousine    0.2132008 0.04008573 0.8392874
  Montbéliarde 0.5104826 0.14917251 1.6445340

$p.value
             two-sided
Predictor      midp.exact fisher.exact chi.square
  autres               NA           NA         NA
  Charolais    0.04056197   0.04136639  0.0332749
  Limousine    0.02590770   0.03030955  0.0222625
  Montbéliarde 0.26277578   0.38193832  0.2433450

$correction
[1] FALSE

attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

Les charolais sont significativement plus atteints de paramphistome que le groupe de référence.
OR =2.496 [1.040,6.029] p<0.05
Les limousins sont significativement moins atteints de paramphistome que le groupe de référence.
OR =0.21 [0.04009,0.8393] p<0.05
Les montbéliardes ne sont pas significativement moins atteints de paramphistome que le groupe de référence.
OR =0.5105 [0.1492,1.645] p>0.05 NS

6 Calcul du Rapport de Taux d'incidence de son intervalle de confiance ainsi que le test de comparaison de ce rapport à 1

Pour calculer un Rapport de Taux d'Incidence et son intervalle de confiance, il est nécessaire de faire appel à une bibliothèque d'outils complémentaires : « epitools ».

```
library(epitools)
rateratio(c(x1,x2,st1,st2)
# x1 représente le nombre de cas dans le groupe  non exposé ;
# x2 représente le nombre de cas dans le groupe  exposé;
# st1 représente le nombre de sujet-temps dans le groupe non exposé  ;
```

st2 représente le nombre de sujet-temps dans le groupe exposé;
Par défaut l'estimation de l'intervalle de confiance n'est pas réalisé par la méthode de Wald basée sur l'approximation de la loi normale mais par la méthode exacte de mid-p ; elle est plus adaptée pour les effectifs petits et pour des RT élevés.
Pour le calcul du nombre de cas et du nombre de sujet-temps, il suffit d'adapter les méthodes de calcules présentés aux chapitres 2 et 3 de ce document.

- **Exemple 1** Dans un suivi de cohorte, il a été observé dans le groupe non exposé 41 cas de mammites pour un nombre de vache-mois de 1201. Tandis que dans le groupe exposé, il y a eu 35 cas de mammites pour une observation de 632 vache-mois.

```
rateratio(c(41,35,1201,632))
$data
          Outcome
Predictor  Cases Person-time
  Exposed1    41        1201
  Exposed2    35         632
  Total       76        1833

$measure
          rate ratio with 95% C.I.
Predictor  estimate    lower    upper
  Exposed1 1.000000       NA       NA
  Exposed2 1.623489 1.027673 2.550396

$p.value
          two-sided
Predictor  midp.exact       wald
  Exposed1         NA         NA
  Exposed2 0.03797212 0.03377064

attr(,"method")
[1] "Median unbiased estimate & mid-p exact CI"
```

Le rapport de taux d'incidence est significativement supérieur à 1.
RT =1.623 [1.028,2.055] p<0.05.
Les vaches du groupe exposé sont significativement plus atteintes de mammites.

7 Calcul d'un indicateur de risque (Odds Ratio ou Risque Relatif) ajusté en prenant en compte un facteur de confusion.

Le calcul d'un indicateur de risque ajusté consiste à calculer une valeur moyenne pondérée entre les valeurs observées pour les différentes strates du facteur de confusion. Cette valeur moyenne est calculée sur les logarithmes des indicateurs de risque et ces logarithmes sont pondérés. Plus l'intervalle de confiance d'un logarithme de l'indicateur de risque est étroit, plus il est précis, plus il y aura de poids dans le calcul de la valeur ajustée. Cependant, pour que le calcul d'une valeur ajustée ait un sens, il faut que les valeurs observées pour les différentes strates du facteur de confusion ne diffèrent pas trop. En d'autres termes, il ne doit pas y avoir d'interaction entre le facteur de confusion et le facteur d'exposition sur le risque.

7.1 Calcul dans chaque strate de l'indicateur de risque (OR_i) et de son intervalle de confiance.

Dans un premier temps, il est important de diviser le jeu de données en plusieurs sous-jeux de données correspondant à chaque niveau du facteur de confusion avec la fonction `split()`. La fonction `split()` renvoit une liste constituée des différents sous jeux de données.

```
a<-split(x,x$y)
```

où x correspond au jeu de donnée
où y correspond au vecteur du facteur de confusion
Dans un second temps, afin d'alléger le script il est avantageux d'utiliser la fonction `lapply()` qui permet d'appliquer une fonction à chaque élément d'une liste.

```
lapply(a,function(sd)oddsratio(sd$exposition,sd$maladie))
```

cette ligne de code calcule l'Odds Ratio pour chaque élément de la liste a (i.e., chaque jeu de données correspondant à une strate.
où a correspond au résultat de la fonction split précédente
la fonction `oddsratio()` peut être remplacée par la fonction `riskratio()` si l'on souhaite calculer des risques relatifs.

- **Exemple 1** Si l'on souhaite diviser un jeu de données selon la catégorie d'âge qui est considéré comme facteur de confusion, puis calculer les OR_i de la castration (exposition) pour le mélanome (maladie) pour chaque niveau de la strate –age (confusion). Nous supposerons que notre jeu de données est nommé d.

```
splitd<-split(d,d$Age)
lapply(splitd,function(sd)oddsratio(sd$Castration,sd$Melanome))
$A1
$A1$data
         Outcome
Predictor non oui Total
    non    26   1    27
    oui    19   1    20
    Total  45   2    47
$A1$measure
         odds ratio with 95% C.I.
Predictor estimate      lower    upper
      non 1.000000         NA       NA
      oui 1.359166 0.03340447 55.30361
$A1$p.value
         two-sided
Predictor midp.exact fisher.exact chi.square
      non         NA           NA         NA
      oui  0.8510638            1  0.8276744
$A1$correction
[1] FALSE
attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"

$A2
$A2$data
         Outcome
Predictor non oui Total
    non     7   1     8
    oui    31   8    39
    Total  38   9    47
$A2$measure
         odds ratio with 95% C.I.
Predictor estimate     lower    upper
      non 1.000000        NA       NA
      oui 1.624853 0.2259971 45.87289
$A2$p.value
         two-sided
Predictor midp.exact fisher.exact chi.square
      non         NA           NA         NA
      oui  0.6722349            1  0.5998022
$A2$correction
[1] FALSE
attr(,"method")
```

```
[1] "median-unbiased estimate & mid-p exact CI"

$A3
$A3$data
         Outcome
Predictor non oui Total
    non     2   1     3
    oui    23  15    38
    Total  25  16    41
$A3$measure
         odds ratio with 95% C.I.
Predictor estimate      lower    upper
      non 1.000000         NA       NA
      oui 1.225969 0.09165118 40.90165
$A3$p.value
         two-sided
Predictor midp.exact fisher.exact chi.square
      non         NA           NA         NA
      oui  0.8818011            1  0.8337476
$A3$correction
[1] FALSE
attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"

$A4
$A4$data
         Outcome
Predictor non oui Total
    non     1   7     8
    oui    14  30    44
    Total  15  37    52
$A4$measure
         odds ratio with 95% C.I.
Predictor  estimate      lower    upper
      non 1.0000000         NA       NA
      oui 0.3450484 0.01263460 2.294834
$A4$p.value
         two-sided
Predictor midp.exact fisher.exact chi.square
      non         NA           NA         NA
      oui  0.3078224    0.4120757  0.2672526
$A4$correction
[1] FALSE
attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

7.2 Recherche d'une éventuelle interaction entre facteur de confusion et le facteur d'exposition sur le risque

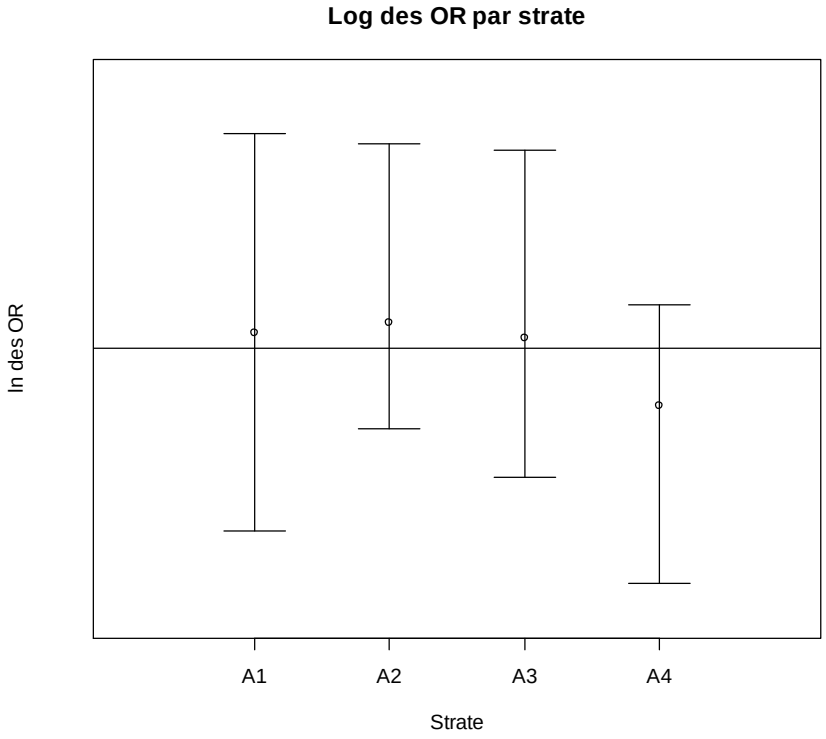
La visualisation des différents intervalles de confiance pour les différentes strates permet de se faire une opinion sur la présence ou non d'une interaction.

Le code ci-dessous permet de visualiser les intervalles de confiances des différents OR_i calculés précédemment. D'une manière générale, si l'on souhaite réutiliser le code vu au paragraphe 7.1 et ici pour d'autres cas, il est juste nécessaire de remplacer :

a. « Age » par le facteur de confusion,
b. « Castration » par le facteur d'exposition
c. « Melanome » par le facteur indiquant la maladie
d. d'adapter les limites de y pour le graphe afin d'avoir les intervalles complets et bien visibles ex. `ylim=c(-5,5)`.

Il est important de copier l'ensemble de lignes, même celles qui ne contiennent qu'un { et de sélectionner l'ensemble des lignes quand on execute le code.

```
traceICOR<-function(i)
{
      ori<-oddsratio(splitd[[i]]$Castration,splitd[[i]]$Melanome)
      est<-ori$measure[2,1]
      inf<-ori$measure[2,2]
      sup<-ori$measure[2,3]
      arrows(i,log(inf),i,log(sup),angle = 90,code =3)
      points(i,log(est))
}
plot(c(0,0),type="n",xlim=c(0,nlevels(d$Age)+1),ylim=c(-5,5), axes=F,
xlab="Strate",ylab="ln des OR",main="Log des OR par strate")
axis(side=1, at=c(1:nlevels(d$Age)),labels=levels(d$Age))
box()
abline(0,0)
lapply(1:nlevels(d$Age),traceICOR)
```



On trouve que les différents OR_i ne sont pas significativement différents de 1 et ne semblent pas graphiquement différer les uns des autres. En d'autres termes, il ne semble pas qu'il y ait une interaction. Un test approprié permettrait de juger plus rigoureusement de la significativité de l'interaction. Un tel test est disponible par exemple dans la librairie « epiR ».

7.3 Calcul de la valeur ajustée d'un OR (ORajusté) et de son intervalle de confiance ainsi que le test de comparaison à 1

S'il parait raisonnable de considérer qu'ils n'y a pas d'interaction, on peut calculer la valeur ajustée. Le résultat est obtenu avec la fonction `mantelhean.test()` si la variable d'exposition n'a que deux niveaux (exposé et non exposé)

```
mantelhaen.test(exposition,maladie,confusion)
```

Rentrer la variable correspondant au facteur d'exposition en premier terme ;
Rentrer la variable correspondant à la maladie en second terme ;
Rentrer la variable correspondant au facteur de confusion en troisième terme ;

- **Exemple 1** portant sur l'exemple précédent, un facteur d'exposition (Castration), une maladie (Melanome), un facteur de confusion (Age)

```
mantelhaen.test(d$Castration,d$Melanome,d$Age)
        Mantel-Haenszel chi-squared test without continuity
correction

data:  Castration and Melanome and Age
Mantel-Haenszel X-squared = 0.0574, df = 1, p-value = 0.8106
alternative hypothesis: true common odds ratio is not equal to
1
95 percent confidence interval:
 0.2946209 2.5692578
sample estimates:
common odds ratio
        0.8700328
```

Dans cet exemple, l'OR de la castration sur la présence de mélanome ajusté sur l'âge vaut 0.87 [0.29 ; 2.57]. Il n'est pas significativement différent de un.