Introduction à la régression logistique

Agnès Caille

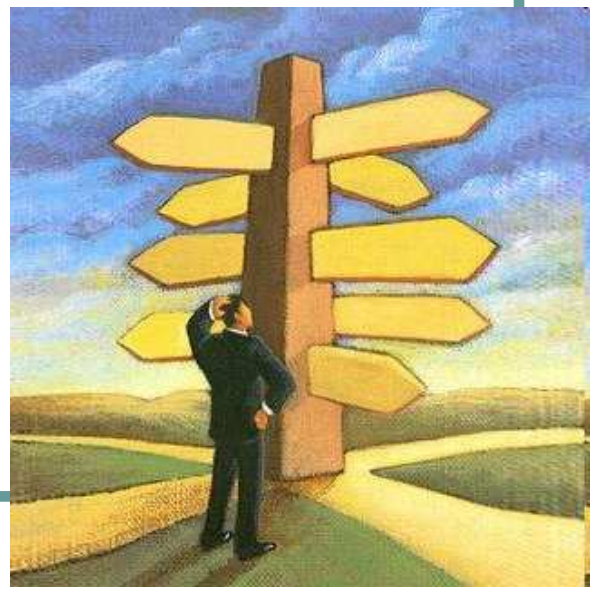
Réunion de Santé Publique

23 février 2012

Choisir le bon test pour la bonne variable

- Dans une étude, les outils statistiques à mettre en œuvre dépendent de la variable d'intérêt (variable à expliquer)
- Il existe différents types de variables : quantitatives, censurées, qualitatives, …
- Variable qualitative binaire (ou dichotomique) :
  - Succès / échec d'une expérience
  - Présence / absence d'une maladie,

  d'une complication
- Tests statistiques
  - Test de comparaison de deux pourcentages
  - Test du khi-deux



Test du khi-deux

- Exemple :
  - Étude de l'association entre la présence de complications infectieuses après une intervention chirurgicale et le traitement préopératoire reçu (Antibiothérapie prophylactique ou Rien)
- Permet de chercher l'association entre la variable binaire d'intérêt et une autre variable qualitative
- Ne permet pas:
  - d'étudier l'effet d'une variable quantitative sur la variable d'intérêt
    - Ex: Age
  - d'étudier l'effet de plusieurs (>2) facteurs à la fois
    - Ex: Traitement préopératoire, Age, Diabète, …
  - d'estimer la force de l'association, si elle existe

La régression logistique

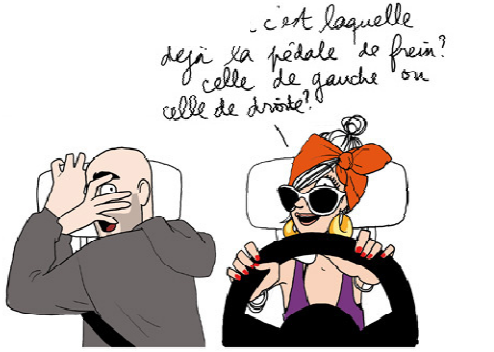
- Etudier la relation entre :
  - Variable à expliquer, qualitative, binaire que l'on veut:
    - Expliquer voire Prédire
    - Ex: Survenue d'une complication infectieuse
  - Variables explicatives, qualitative ou quantitative
    - Facteurs de risque ou facteurs protecteurs
    - Ex: Antibiothérapie, Age, Diabète…

Intérêts de la régression logistique

- Etudier la relation entre une variable binaire et des variables qualitatives et quantitatives
- Quantifier la force de l'association
- En tenant compte de l'effet des autres variables du modèle (ajustement)
- Prédire la probabilité, pour un sujet donné, de connaître l'évènement d'intérêt, en fonction de ses caractéristiques (données cliniques, biologiques, socio-démographiques, antécédents etc…)

Des questions comme …

- Y-a-t-il une relation entre la probabilité de développer un cancer du sein et la consommation d'alcool, la prise de pilule, l'âge, le tabagisme ?
- Y-a-t-il une relation entre la probabilité de trouver un emploi et l'âge, le sexe, le niveau de formation, l'expérience professionnelle ?
- Y-a-t-il une relation entre la probabilité d'obtenir son permis de conduire et le nombre d'heures de conduite, la conduite accompagnée, le sexe ?

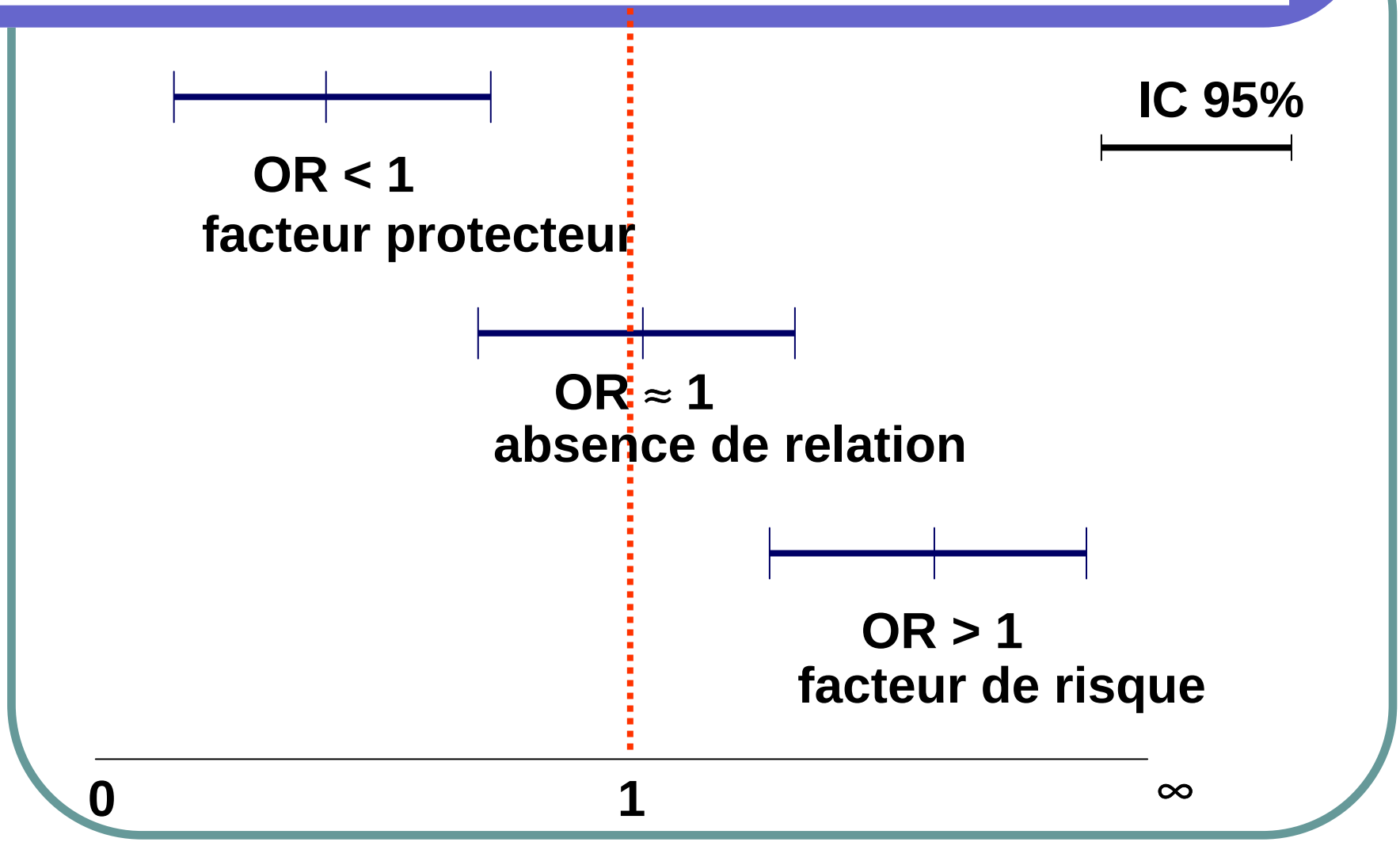


Odds ratio (OR)

- Relation entre une exposition (E) et une maladie (M) ?
- P_1=P(M+/E+)=a/n_1
- P_0=P(M+/E-)=c/n_0
- OR= $\frac{P_1/(1-P_1)}{P_0/(1-P_0)} = \frac{ad}{bc}$
- OR compris entre 0 et + ∞
- OR = 1 : Absence de relation entre E et M
- OR et son IC 95 %

	M+	M-	
E+	a	b	n1
E-	c	d	n0
	m1	m0	

Interprétation d'un OR et de son IC



Causalité

Un facteur de risque n’est pas forcément causal.

FDR ≠ Cause

Tests des paramètres

- Tester la significativité des paramètres du modèle de régression logistique

- Hypothèse nulle = Absence de relation entre E et M (OR=1)
  - Intervalle de confiance contient 1, l'OR n'est pas significativement différent de 1. On conclut qu'il n'existe pas, au risque 5%, d'association entre les deux variables.

    ⇨ non rejet de l'hypothèse nulle, p >0.05

  - Intervalle de confiance ne contient pas 1, l'OR est significativement différent de 1. On conclut au risque 5% qu'il existe une relation entre les deux variables

    ⇨ rejet de l'hypothèse nulle, p ≤ 0.05

Un exemple

- Etude des facteurs associés au risque de survenue d'une infection par le VIH en Tanzanie
- Nov 1991 à Déc 1992
- 763 femmes de plus de 15 ans
  - 189 séropositives
  - 574 séronégatives
- Facteurs étudiés: transfusion, nombre de partenaires et niveau d'éducation

Etude de la relation avec la transfusion sanguine

Procédure FREQ

Table de Transfusion par VIH

Transfusion	VIH		
Fréquence Pourcentage Pctage en ligne	0	1	Total
0	553 72.57 75.96	175 22.97 24.04	728 95.54
1	20 2.62 58.82	14 1.84 41.18	34 4.46
Total	573 75.20	189 24.80	762 100.00

Valeur(s) manquante(s) = 1

Codage:

Pas de transfusion = 0

Transfusion = 1

$$OR = \frac{14 \times 553}{175 \times 20} = 2.21$$

Résultats:

OR Transfusion = 2.21 IC 95% [1.09 – 4.47] p=0.03

1. 1 n'est pas dans l'intervalle de confiance
2. OR est supérieur à 1 : le risque d'infection par le VIH est plus élevé pour les femmes ayant eu une transfusion sanguine

Etude de la relation avec le nombre de partenaires

Procédure FREQ

Table de NbPart par VIH

NbPart	VIH		
Fréquence Pourcentage Pctage en ligne	0	1	Total
0	263 34.47 84.29	49 6.42 15.71	312 40.89
1	311 40.76 68.96	140 18.35 31.04	451 59.11
Total	574 75.23	189 24.77	763 100.00

Codage:

≤ 1 partenaire = 0

> 1 partenaire = 1

Résultats:

OR NbPart = 2.42 IC 95% [1.68 – 3.48] p<.0001

1. 1 n'est pas dans l'intervalle de confiance
2. OR est supérieur à 1 : le risque d'infection par le VIH est plus élevé pour les femmes ayant eu plus d'un partenaire

Etude de la relation avec le niveau d'éducation

Procédure FREQ

Table de Education par VIH

Education	VIH		
Fréquence Pourcentage Pctage en ligne	0	1	Total
0	378 49.61 71.86	148 19.42 28.14	526 69.03
1	195 25.59 82.63	41 5.38 17.37	236 30.97
Total	573 75.20	189 24.80	762 100.00

Valeur(s) manquante(s) = 1

Codage:

Pas de scolarité = 0

Scolarité = 1

Résultats:

OR Education = 0.54 IC 95% [0.37 – 0.79] p=0.002

1. 1 n'est pas dans l'intervalle de confiance
2. OR est inférieur à 1 : le risque d'infection par le VIH est moins élevé pour les femmes ayant eu une scolarité

Etude de la relation entre le nombre de partenaires et le niveau d'éducation

Procédure FREQ

Table de Education par NbPart

Education	NbPart		
Fréquence Pourcentage Pctage en ligne	0	1	Total
0	152 19.95 28.90	374 49.08 71.10	526 69.03
1	160 21.00 67.80	76 9.97 32.20	236 30.97
Total	312 40.94	450 59.06	762 100.00

Valeur(s) manquante(s) = 1

- Pourcentage de femmes ayant eu plus d'un partenaire
  - 71.1% pour les femmes sans scolarité
  - 32.2% pour les femmes avec scolarité
  - Khi-deux p <.0001

Interprétation: « toutes variables égales par ailleurs »

Procédure LOGISTIC

Estimations par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	-1.5182	0.1831	68.7248	<.0001
NbPart 1	1	0.7767	0.1971	15.5242	<.0001
Education 1	1	-0.3349	0.2110	2.5185	0.1125

Estimations des rapports de cotes

Effet	Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
NbPart 1 vs 0	2.174	1.477	3.200
Education 1 vs 0	0.715	0.473	1.082

- OR Nb Part = 2.17 IC95% [1.48-3.20]
- OR du Nb Part ajusté sur le niveau d'éducation
- OR du NbPart pour un niveau constant d'éducation
- Nb Part = facteur de risque indépendant d'infection par le VIH

Interprétation: « toutes variables égales par ailleurs »

Procédure LOGISTIC

Estimations par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	-1.5182	0.1831	68.7248	<.0001
NbPart 1	1	0.7767	0.1971	15.5242	<.0001
Education 1	1	-0.3349	0.2110	2.5185	0.1125

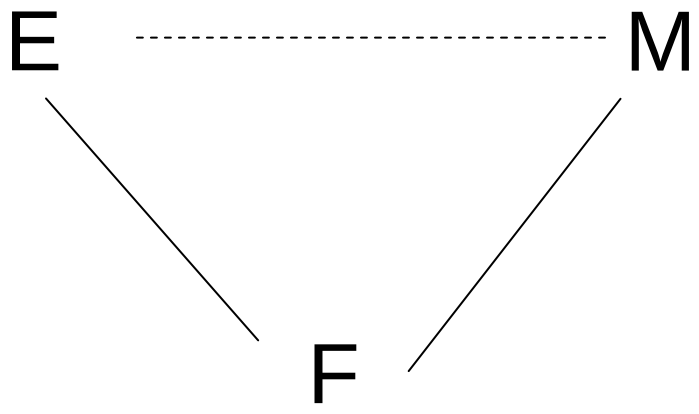
Estimations des rapports de cotes

Effet		Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
NbPart	1 vs 0	2.174	1.477	3.200
Education	1 vs 0	0.715	0.473	1.082

- OR Education = 0.72 IC95% [0.47-1.08]
  1. 1 est dans l'intervalle de confiance

- Après prise en compte (ajustement) du nombre de partenaires, le niveau d'éducation n'est plus associé au risque d'infection par le VIH
- On parle de confusion, ici le nombre de partenaires est facteur de confusion dans la relation entre niveau d'éducation et infection par le VIH

Facteur de confusion



- L’exposition étudiée est associée à un facteur
- Ce facteur est indépendamment associé à la maladie
- Mesure de l’association entre E et M est perturbée par F, facteur de confusion

Autre exemple

- Etude des facteurs associés au risque de survenue d'une pneumonie chez les enfants éthiopiens
- Jan 1989 à Déc 1993
- Etude cas-témoins
  - 500 cas et 500 témoins
  - Enfants de moins de 5 ans
- Facteurs étudiés: rachitisme et allaitement maternel

Etude de la relation brute avec le rachitisme et l'allaitement maternel

Résultats analyses brutes (univariées) :

- **OR Rachitisme** = 22.11 IC 95% [11.34 – 43.12] p<.0001
  1. 1 n'est pas dans l'intervalle de confiance
  2. OR est supérieur à 1 : le risque de pneumonie est plus élevé pour les enfants souffrant de rachitisme

- **OR Allaitement** = 0.940 IC 95% [0.913 – 0.968] p=0.01
  - Allaitement = variable quantitative en mois
  - Il s'agit de l'OR pour une augmentation d'un mois de la durée de l'allaitement maternel
  1. 1 n'est pas dans l'intervalle de confiance
  2. OR est inférieur à 1 : le risque de pneumonie est moins élevé pour les enfants ayant été allaité plus longtemps

Etude de la relation avec le rachitisme et l'allaitement maternel après ajustement

- Résultats analyses brutes (univariées) :
  - **OR Rachitisme** = 22.11 IC 95% [11.34 – 43.12] p<.0001
  - **OR Allaitement** = 0.940 IC 95% [0.913 – 0.968] p=0.01
- L'allaitement est lié à la survenue d'une pneumonie mais il est aussi probablement lié à la survenue du rachitisme.
- Résultats analyses ajustées (multivariées) :
  - **OR Rachitisme** = 13.37 IC 95% [8.08 – 24.22] p=.0001
  - **OR Allaitement** = 0.962 IC 95% [0.931 – 0.995] p=0.03
  - Après ajustement le rachitisme reste un facteur de risque de la pneumonie mais la force de la relation est atténuée par l'ajustement

La sélection des variables

1. Description des variables
2. Analyse univariée:
   - Association de la variable à expliquer avec chacune des variables « explicatives »
   - OR bruts
3. Sélection des variables du modèle complet
   - *A priori* +++
   - Facteurs de confusion connus
   - Variables d'intérêt clinique
   - Variables avec p< 0.20 en analyse univariée
   - Une limite = le nombre d'événements (10 EPV)

Pour aller plus loin

- Falissard, Bruno. *Comprendre Et Utiliser Les Statistiques Dans Les Sciences De La Vie*. 2e éd. Paris: Masson, 1998.

- Cours de master 2 de Jean Bouyer:
- http://www.hal.inserm.fr/docs/00/12/43/35/PDF/Reg_log_M2.pdf

- Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd Revised ed. John Wiley & Sons Inc; 2000.

Procédure LOGISTIC

Analyse des effets Type 3

Effet	DDL	Khi-2 de Wald	Pr > Khi-2
Transfusion	1	6.6418	0.0100
NbPart	1	15.7007	<.0001
Education	1	3.3570	0.0669

Estimations par l'analyse du maximum de vraisemblance

Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	-1.5560	0.1845	71.0882	<.0001
Transfusion 1	1	0.9592	0.3722	6.6418	0.0100
NbPart 1	1	0.7841	0.1979	15.7007	<.0001
Education 1	1	-0.3917	0.2138	3.3570	0.0669

Estimations des rapports de cotes

Effet		Valeur estimée du point	Intervalle de confiance de Wald à 95 %	
Transfusion	1 vs 0	2.610	1.258	5.413
NbPart	1 vs 0	2.190	1.486	3.228
Education	1 vs 0	0.676	0.445	1.028