Mesures d'associations
et
Tests statistiques

Rappels

Mesures d'association

- Objectif des études
  - Descriptif
  - Analytique → recherche de facteurs associés à un événement clinique / caractéristique particulière

- 2 principales mesures d'association
  - Risque Relatif → étude de cohorte
  - Odds ratio / rapport de cotes
    - étude cas-témoin
    - voir étude transversale

Rappel : Risque Relatif

$$RR = \frac{\text{Risque de développer la maladie si exposé au facteur (R1)}}{\text{Risque de développer la maladie si non exposé (R0)}}$$

	Malades	Non malades	
Exposés	**a**	**b**	**R1=a/(a+b)**
Non exposés	**c**	**d**	**R0=c/(c+d)**

RR=1 → le facteur n'est pas lié à la maladie

RR< 1 → le facteur est associé à une diminution du risque de maladie (facteur protecteur)

RR> 1 → le facteur est associé à une augmentation du risque de maladie (facteur de risque)

Estimation possible dans le cadre de cohorte → pas de contrôle de la proportion de malades et de non malades + suivi dans le temps

Exemple

	choriorétinite	pas de choriorétinite	
Autres lésions Toxoplasmiques	19	16	35
Pas d'autres Lésion	60	232	292
	79	248	327

RR=(19/35)/(60/292)=2.64

Les enfants présentant d'autres signes de toxoplasmose congénitale à la naissance ont un risque multiplié par plus de 2.5 de développer une choriorétinite au cours de leurs enfance que ceux n'ayant pas de lésions

OR

	Malades	Non malades
Exposés	**a**	**b**
Non exposés	**c**	**d**

$$OR = \frac{R1}{(1-R1)} \times \frac{(1-R0)}{R0} = \frac{a \times d}{b \times c}$$

- Même interprétation que le RR
- Bonne estimation du RR si la maladie est rare dans la population (prévalence faible)
- Utilisé dans les études cas-témoins (contrôle de la proportion de malades/non malades → estimation RR impossible)

Exemple

	toxoplasmose (cas)	pas de toxoplasmose	
Viande insuffisamment cuite	44	15	59
Pas de consommation de viande insuf. cuite	36	65	101
	80	80	160

OR=(44x65)/(36x15)=5.3

Les femmes ayant consommé de la viande insuffisamment cuite sont plus fréquemment contaminées par le toxoplasma gondii que les femmes n'en ayant pas consommé

Autres mesures d'association

- Plus accessoire
  - Excès de risque
    - ΔR=R1-R0
    - Si ΔR=0→ pas d'association entre exposition et maladie
  - Risque attribuable
    - Proportion de cas de maladie dans la population qui seraient évités si l'exposition au facteur était supprimée
    - $$RA = \frac{p \times (RR - 1)}{p \times (RR - 1) + 1}$$

      avec p, fréquence de l'exposition dans la population
  - Fraction étiologique du risque (RA chez les exposés)
    - FER=(RR-1)/RR

Mesures d'association

- 2 principales
  - Risque Relatif
  - Odds ratio / rapport de cotes
- Question : association significative ?

Rôle des statistiques

Population générale

Généralisation si les caractéristiques de la population cible ne sont pas trop éloignées de celles de la population générale

Population cible

Inférence statistique
À moduler suivant la représentativité de l'échantillon

Echantillon

Estimations, Intervalles de confiances, tests

Signification statistique

- 2 méthodes
  - Intervalle de confiance
  - Test

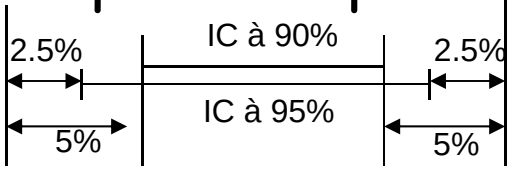
Intervalle de confiance

- Intervalle dans lequel la vraie valeur a une probabilité p de se situer

  Exemple IC à 95%:

  - La vraie valeur a 95% de chances d'être à l'intérieur de l'intervalle,
  - et 5% de chances d'être à l'extérieur (2,5% au dessus de la borne sup, 2,5% en dessous de la borne inf)

- Deux règles
  - Plus la taille de l'échantillon est grande, plus l'IC est étroit
  - Plus le risque d'erreur alpha accepté est grand, plus l'IC est étroit



Intervalle de confiance

- RR ou OR
  - Si l'intervalle de confiance comprends 1
    → pas d'association entre l'exposition et l'événement étudié
  - Si l'intervalle de confiance ne comprends pas 1
    → association significative entre l'exposition et l'événement étudié
  - Exemple
    - RR de développer une choriorétinite en cas de calcification intracrânienne chez des enfants atteints de toxoplasmose congénitale : 2.61 (1.49-4.59)

Tests statistiques

- Résultat sur un échantillon → m1≠m2
- Quel serait le résultat sur la population cible?
  → problème des fluctuations d'échantillonnage
  - Population m1 ≠m2

    Si tire 100 échantillons différents → m1=m2 dans un certain nombre d'entre eux
  - Population m1=m2

    Si tire 100 échantillons différents → m1 ≠ m2 dans un certain nombre d'entre eux

Tests statistiques et risque d'erreurs

- Test statistique
  - consiste à rejeter ou non une certaine hypothèse (hypothèse nulle) qui concerne la population étudiée aux vues des résultats observés sur l'échantillon → inférence à la population cible

  - En pratique, calcul de la probabilité p (p-value) d'observer sur un échantillon quelconque de même effectif
    - une différence entre proportions (moyennes)
    - qui soit ≥ à celle observée sur l'échantillon d'étude

    du fait du hasard (pas de différence dans la population cible)

Tests statistiques et risque d'erreurs

- **Deux types d'erreurs en statistique** :
  - α: probabilité de conclure à une différence alors qu'elle n'existe pas
    - $p < \alpha$ (en général, 5%): différence significative c-a-d la différence observée ne peut pas être due au hasard
    - $p \geq \alpha$: différence non significative , mais on ne peut pas exclure que cette différence existe réellement
  - β: probabilité de conclure qu'il n'y a pas de différence alors qu'elle existe réellement

  → α et β paramètres sont fixés en début d'étude (éléments du calcul du nombre de sujets nécessaire)

Tests statistiques et risque d'erreurs

Conclusion du test	Réalité dans la population : Pas de différence (H0)	Réalité dans la population : Différence (H1)
Pas de différence significative	Conclusion vraie 1 - α	Conclusion fausse B
Différence significative	Conclusion fausse α	Conclusion vraie 1 - β 'Puissance'

En pratique

- Test de l'association entre la présence d'un nouveau marqueur et la survenue d'un IDM
  - Données
    - P1=proportion d'IDM chez les patients porteurs du marqueur
    - P2= proportion d'IDM chez les patients non porteurs
  - Hypothèses
    - Hypothèse nulle : P1=P2
    - Hypothèse alternative : P1≠P2

En pratique

- Au risque α de 5% fixé a priori
  - Résultat du test
    - P=0.002 :
      - » on a moins de 2 chances pour 1000 de se tromper en considérant que les deux pourcentages diffèrent et qu'il existe une association.
      - » La différence observée ne peut être due au hasard
    - P=0.15 → test non significatif : on a 15% de chance de se tromper si on considère que les deux pourcentages diffèrent
      - » Soit effectivement pas de différence dans la population
      - » Soit manque de puissance : on ne peut pas mettre en évidence une différence alors qu'elle existe

En pratique

- p-value (p) = probabilité de rejeter à tort l'hypothèse nulle
  - Risque alpha de 5% fixé a priori
  - p-value calculée pour le test > 0.05
    → pas de rejet de H0 (test non significatif)
  - p-value calculée pour le test ≤ 0.05
    → rejet de H0 (test significatif)

Principaux tests

variables	Qualitative	Quantitative
Qualitative	Chi-2 de Pearson - CA : Effectif théorique ≥5 - **Alternative** : * correction de Yates * test exact de Fisher	>2 groupes : ANOVA - CA : *normalité de la variable quant. dans chaque groupe * homogénéité des variances - Alternative : Kruskal-Wallis
Quantitative	2 groupes : test de Student - CA : *normalité de la variable quant. dans chaque groupe * homogénéité des variances - Alternative : Mann-Whitney	Corrélation de Pearson / régression linéaire - **CA** : normalité - **Alternative** : corrélation des rangs de Spearman

Principaux tests

- Mesures répétées chez les mêmes sujets
  - Comparaison de pourcentage → test=chi2 de McNemar
  - Comparaison de moyennes → test?
    - 2 groupes : test de Student apparié
    - >2 groupes : ANOVA pour mesures répétées
    - Non paramétrique : Wilcoxon
  - Concordance?
    - Variables qualitatives : coefficient Kappa
    - Variables quantitatives : coefficient de corrélation intraclasse

Analyse multivariée/multivariable

- Permet de tenir compte de facteurs de confusion
  - Facteur lié à l'exposition et à la maladie
  - Possible explication association entre exposition et maladie
  - Exemple :
    - 1ère observation : association du stade du cancer et du risque de décès
    - 2ème observation : indication ou non d'une chimiothérapie en fonction du stade du cancer
    - Si étude
      - de l'association entre chimiothérapie et risque de décès
        → effet péjoratif de la chimio
      - Idem + prise en compte du stade dans l'analyse
        → effet bénéfique
    - Le stade est un facteur de confusion dans l'étude de la relation entre chimio et décès

Analyse multivariée

- Permet de tenir compte de facteurs de confusion → **ajustement** +++
  - =estimation de
    - l'effet **propre** de la variable d'intérêt sur le risque de survenue de l'événement étudié
    - indépendamment de l'impact des facteurs de confusion
  - Méthode : modèle **multivarié/multivariable**

    → Relation de type $Y=\alpha+\beta X$

    Y= décès/autre critère de jugement=variable à expliquer

    X= variables explicatives (chimio, stade, age, sexe,…)

    Exemple: $Y=f(\alpha+\beta_1 chimio+\beta_2 stade+\beta_3 age+\beta_4 sexe)$

Les éléments qui suivent seront revus en MM1

Modèles possibles en fonction de la variable à expliquer (Y)

- Variable quantitative
  → régression linéaire : $Y = \alpha + \beta X$
  - Surtout étude transversale / biologique

Régression linéaire

- **Exemple** : effet d'un traitement et de l'âge sur le taux de lymphocytes CD4 chez des patients VIH+
  - Variable à expliquer = CD4
  - Variables explicatives: X1=traitement (X= 1 quand le patient reçoit le traitement et 0 dans le cas contraire) et X2=âge (en années)

    → Equation : $CD4= \alpha+\beta_1X_1+\beta_2X_2$
  - CA : CD4 → distribution # normale
  - Exemple :
    - $CD4= 100+200X_1 + (-2) X_2$
    - Interprétation pour un patient de 30 ans
      - Âge : le taux de CD4 diminue de 2 cellules/mm3 pour chaque année d'âge supplémentaire
      - Traitement : le taux de CD4 est égal à 100 +200 -2x30=240 si le patient reçoit le traitement et à 100-2x30=40 sinon

Modèles possibles en fonction de la variable à expliquer

- Variable à expliquer = compte d'événement peu fréquent dans une population (λ)

  → régression de Poisson : $\log(\lambda)=\alpha+\beta X$

  - Surtout
    - études d'incidence, de mortalité, en population
    - Études de cohortes

Régression de Poisson

- Interprétation :
  - log(λ) augmente de β chaque fois que X augmente d'une unité
  - Exemple : log(λ)=0.0025 + 0.25(département X vs département de référence)+ 0.02 âge - 0.2(sexe)
  - **Exp(β)=risque relatif**
    - RR département X vs département de référence
      = exp(0.25)=1.28

      → l'incidence est 1.28 fois plus importante dans le dept X que dans le dept de référence
    - RR pour l'âge : pour chaque augmentation d'une année d'âge l'incidence est augmenté de : exp(0.02)=1.02, soit 2%

Modèles possibles en fonction de la variable à expliquer

Y = Variable qualitative à deux classes (dichotomique)

→ régression logistique Y= proportion =P

$$\text{logit}(Y) = \log(P/(1-P)) = \alpha + \beta X$$

avec exp(β)=OR

- Exemple : réponse au traitement
- Surtout utilisé dans
  - Études cas-témoin
  - Études transversales
  - Études pronostiques avec établissement d'un score

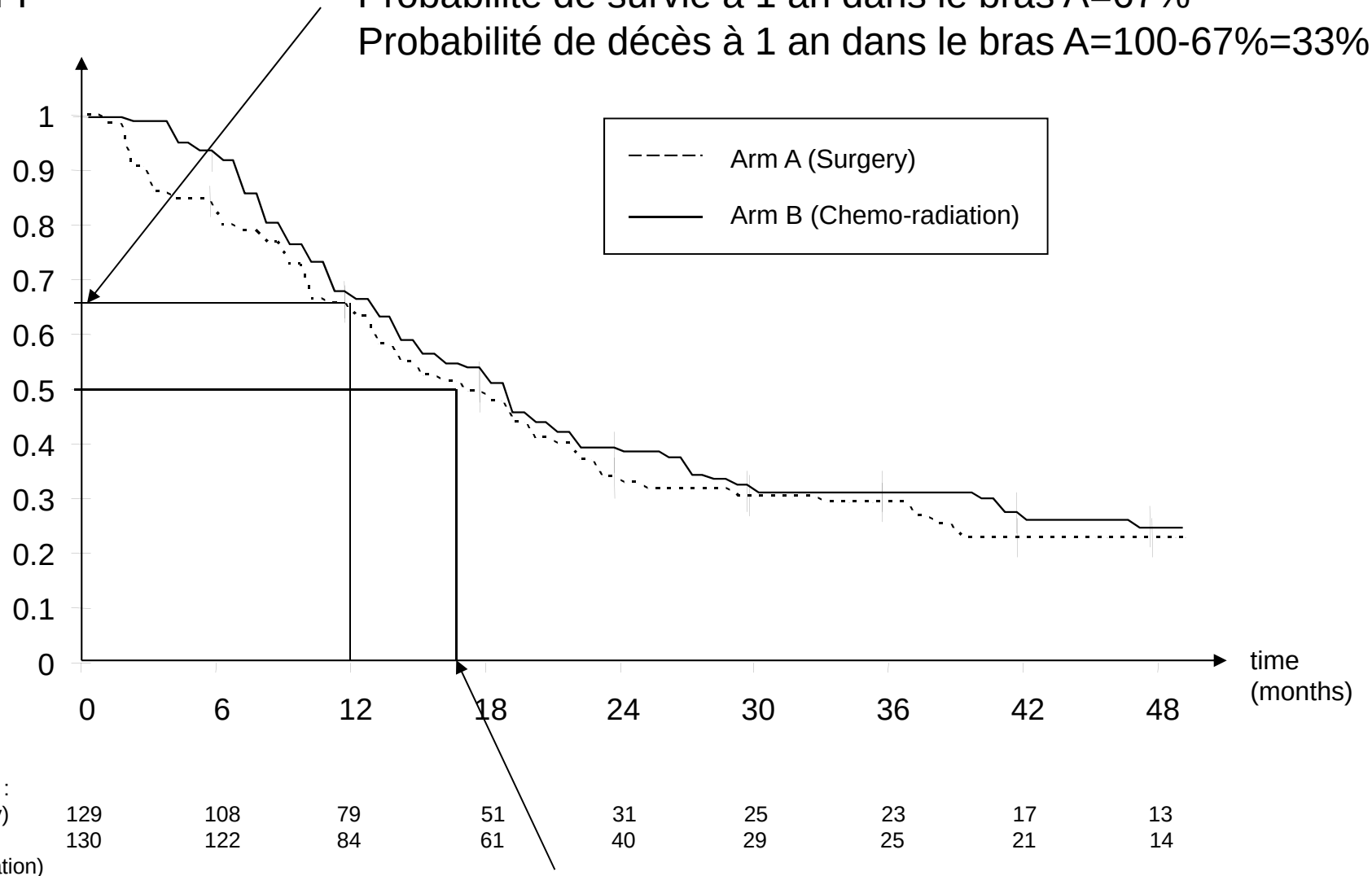
Modèles possibles en fonction de la variable à expliquer

- Délai de survenue d'un événement (décès, récidive, infection opportuniste)
  - → Analyse de survie (time-to-event analysis)
  - Application :
    - études de cohortes, études pronostiques, essais thérapeutiques
    - Très utile en cas de suivi inégal entre les sujets :
      - » estimation des probabilités de survenue d'un événement à chaque temps
      - » en tenant compte des individus restant encore à risque dans l'étude à ce temps là et de ce qui s'est passé au temps précédent
  - Univarié
    - courbes de **Kaplan-Meier** → estimation de la probabilité de survenue de l'événement à un temps donné
    - Test du **logrank** → comparaison des courbes

ITT

Probabilité de survie à 1 an dans le bras A=67%

Probabilité de décès à 1 an dans le bras A=100-67%=33%



Patients at risk :									
Arm A (surgery)	129	108	79	51	31	25	23	17	13
Arm B (chemo-radiation)	130	122	84	61	40	29	25	21	14

Médiane du délai de survie

Modèles possibles en fonction de la variable à expliquer

- Délai de survenue d'un événement (décès, récidive, infection opportuniste) → analyse de survie (time-to-event analysis)
  - Univarié
  - Multivarié
    - Modèles paramétriques : exponentiel, Weibull
    - Modèle semi-paramétrique des risques proportionnels = **modèle de Cox**
    - → estimation du RR=exp(β)

En résumé

- Analyse multivariée

  → estimation de

  - l'effet **propre** de la variable d'intérêt X sur le risque de survenue de l'événement étudié Y
  - indépendamment de l'impact des facteurs de confusion

  → Choix du modèle dépend de la nature de Y

  - Y=Variable quantitative → régression linéaire
  - Y=Taux (incidence)→ régression de Poisson (RR)
  - Y= Variable dichotomique (0/1) → régression logistique (OR)
  - Y=Délai de survenue d'un événement
    → analyse de survie / modèle de Cox (RR)

  → On obtient des mesures d'association (RR ou OR) ajustées, mais l'interprétation finale dépend de l'intervalle de confiance de ces mesures