Biostatistiques

Philippe Carrère, UAG 2014

Biostatistique : objectifs

- Organiser les données provenant d'observations individuelles
- Décrire les phénomènes par des paramètres résumant ces observations
- Estimer les valeurs de ces paramètres
- Comparer ces paramètres
- Prédire la probabilité de survenue d'évènement

Biostatistique : limites

- Hypothèse aléatoire
- Hypothèse de distribution
- Conditions d'application des tests
- Approximation du réel

Variables

- Quantitative
  - Continue
  - Discrète (dénombrement)
- Temporelle
- Qualitative
  - Ordinale
  - Nominale
  - Binaire (dichotomique, booléenne, de Bernouilli)

Discrétisation

- Facilite exploitation et présentation
- Construction d'une échelle de classification
  - Par amplitude (intervalles égaux)
  - Par fréquence (effectifs égaux)
  - Convenance (pertinence clinique)

Mesures

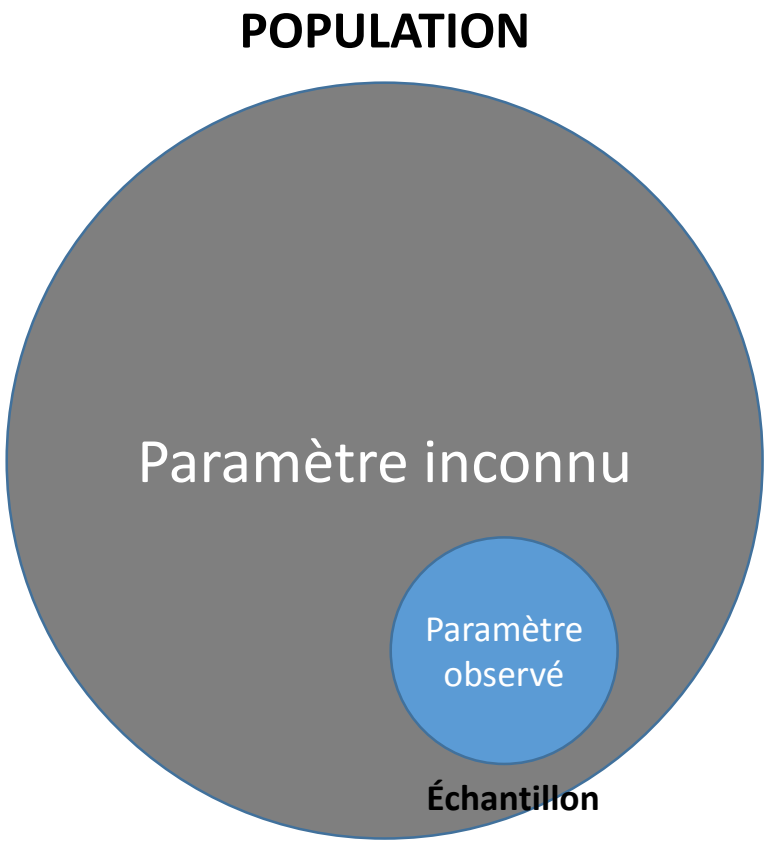
- Paramètres de position
  - Médiane (2 groupes d'effectifs égaux)
  - Quartiles, déciles, percentiles
  - Mode (pic de distribution)
  - Moyenne
  - Fréquence, fréquence cumulée (variable qualitative)
  - Pourcentage (variable binaire)

Mesures

- Paramètres de dispersion
  - Extrêmes
  - Étendue
  - Intervalle interquartile
  - Variance, écart type

Estimation d’un paramètre

- Valeur réelle : inconnue
- Si échantillon représentatif, valeur observée proche de la valeur réelle
- k échantillons => k valeurs observées, comprises dans une fourchette d’estimation



Intervalle de confiance

- Moyenne
  - Écart type de la moyenne (standard error) : $s_m = \frac{s}{\sqrt{n}}$
  - Intervalle de confiance à 95% : $[m - 1{,}96 * s_m \; ; m + 1{,}96 \; * s_m]$
- Pourcentage
  - Écart type d'un pourcentage : $s_p = \sqrt{\frac{p*(1-p)}{n}}$
  - Intervalle de confiance à 95% : $\left[p - 1{,}96 * s_p \; ; p + 1{,}96 \; * s_p\right]$

Tests

- Objectifs :
  - Comparer des populations
  - Étudier le lien entre des variables
- ! Conditions d'application
- Deux grandes familles :
  - Paramétriques (comparaison de paramètres)
  - Non paramétriques (comparaison de distributions)

Tests

- Hypothèse nulle H_0

  paramètre population 1 ≈ paramètre population 2

  (différence observée : fluctuation échantillonnage)

- Hypothèse alternative H_1
  - Bilatérale

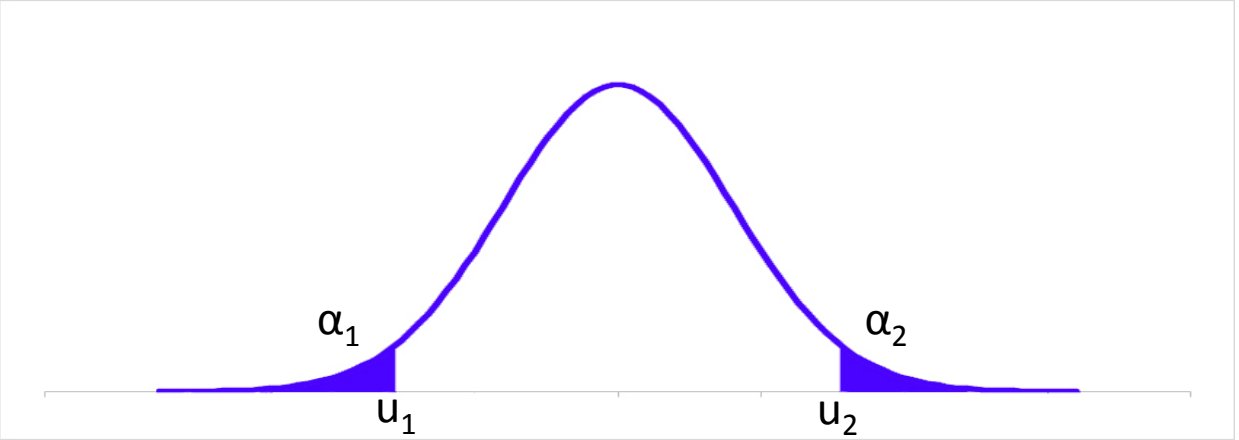
    paramètre population 1 ≠ paramètre population 2
  - Unilatérale

    paramètre population 1 > paramètre population 2

    paramètre population 1 < paramètre population 2

Tests

- Écart observé *u* : réalisation d'une variable aléatoire U
- Sous H_0 , cette variable suit une loi de probabilité



- Si *u* est compris dans $[u_1 ; u_2]$, la différence n'est pas significative : on ne peut pas rejeter l'hypothèse nulle H_0
- Si *u* n'est pas compris dans $[u_1 ; u_2]$, la différence est significative : on rejette H_0 et on accepte l'hypothèse alternative H_1, au risque α près

Risque Alpha et Beta

- Risque α = probabilité de rejeter H_0 alors que H_0 est vraie
  - Dans le graphique précédent, probabilité que $U < u_1$ ou $U > u_2$
  - On consent le plus souvent à un risque α de 5%

- Risque β = probabilité de ne pas rejeter H_0 alors que H_1 est vraie
  = Manque de puissance
  - Intervient dans le calcul du nombre de sujets nécessaires, pas dans le calcul de l'intervalle de confiance d'une estimation

Significativité : p-value

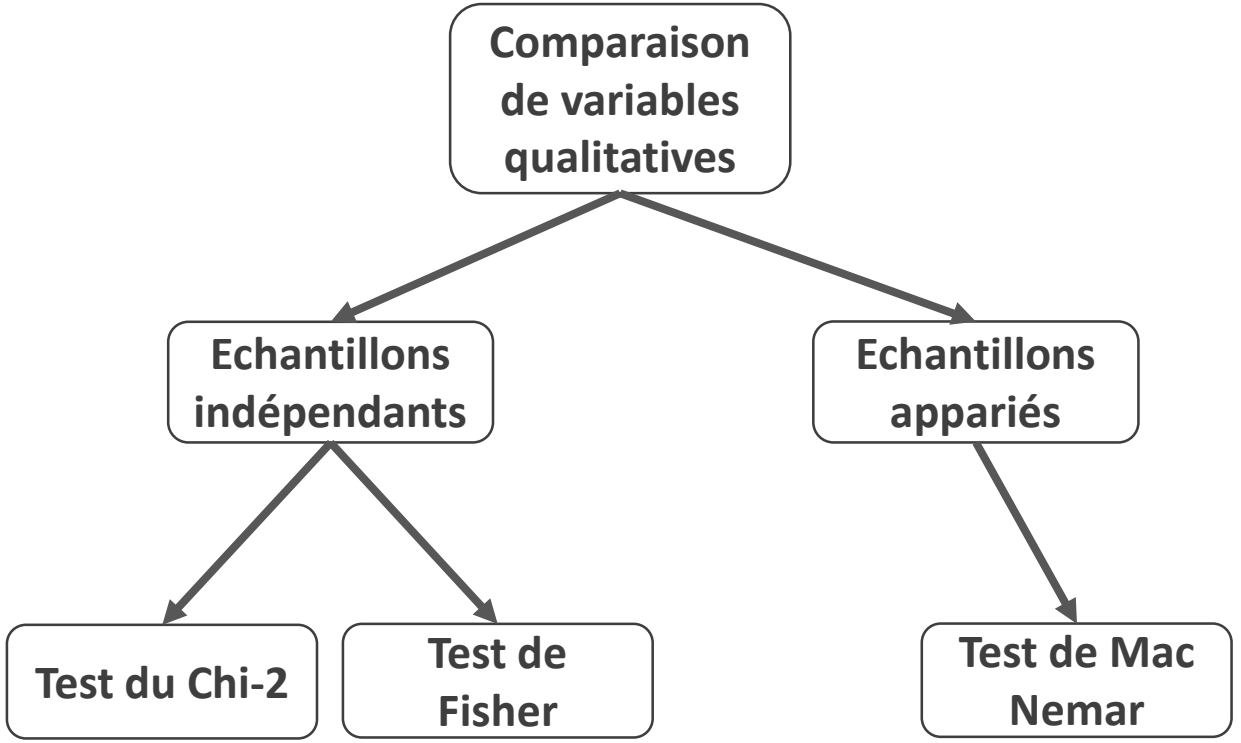
Le p n'est pas…	**Le p est…**
p n'est pas la probabilité de l'hypothèse nulle $Pr(H_0)$	p est la probabilité d'obtenir le résultat observé si l'hypothèse nulle est vraie, p est la probabilité conditionnelle du résultat sous l'hypothèse nulle, p = $Pr(\text{résultat} \mid H_0)$
p n'est pas la probabilité d'absence de différence 1-$Pr(H_1)$	p est la probabilité d'observer une différence au moins aussi importante si en réalité il n'y a pas de différence
p n'est pas la probabilité que le traitement n'ait pas d'effet	p est la probabilité d'obtenir le résultat qui a été observé si le traitement est en réalité inefficace
p : 0,05 ne signifie pas qu'il y a 5% de risque que le traitement soit sans effet	p : 0,05 signifie qu'il y a 5% de risque d'observer le résultat obtenu si le traitement est sans effet

Analyse bivariée

- Consiste à comparer deux groupes ou à évaluer la relation entre deux variables
- Indices d'effet :
  - Comparer deux groupes : difference de risque, de moyenne…
  - Relation entre deux variables : OR (tableau à 4 cases), RR…
- Significativité : p-value et intervalle de confiance de l'indice d'effet

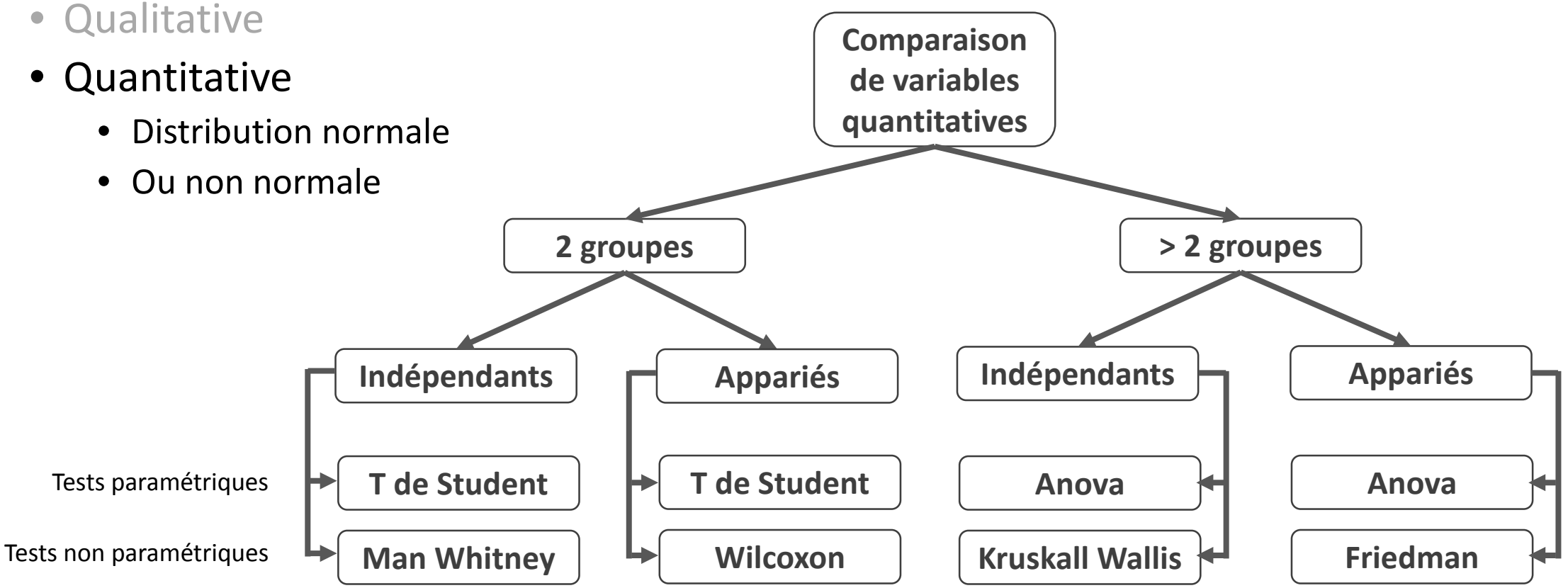
Analyse bivariée

- Calcul de la p-value : le test à utiliser dépend du type de variable…
  - Qualitative



Analyse bivariée

- Calcul de la p-value : le test à utiliser dépend du type de variable…
  - Qualitative
  - Quantitative
    - Distribution normale
    - Ou non normale



Analyse bivariée : régression simple

- Si l'on cherche à estimer une relation entre une variable à expliquer Y et une variable explicative X, le modèle de régression s'écrit :

$$E(Y) = \beta_0 + \beta_x X$$

Si $\beta_x = 0$, pas de relation

Si $\beta_x > 0$, relation positive

Si $\beta_x < 0$, relation négative

- Un intervalle de confiance de β_x peut être calculé, la significativité de la relation peut être testée

Analyse multivariée

- Consiste à tester la relation entre une variable dépendante et plusieurs variables indépendantes
  - Pour contrôler des facteurs de confusion existant dans la relation entre une variable à expliquer et une variable explicative
  - Pour explorer les effets de plusieurs variables explicatives sur une même variable à expliquer

Analyse multivariée : régression multiple

- Si l'on cherche à estimer une relation entre une variable à expliquer Y et une variable explicative X avec plusieurs facteurs de confusion C_1 à C_n, le modèle de régression s'écrit :

$$E(Y) = \beta_0 + \beta_x X + \beta_{C1} C_1 + \ldots + \beta_{Cn} C_n$$

  Si $\beta = 0$, pas de relation

  Si $\beta > 0$, relation positive

  Si $\beta < 0$, relation négative

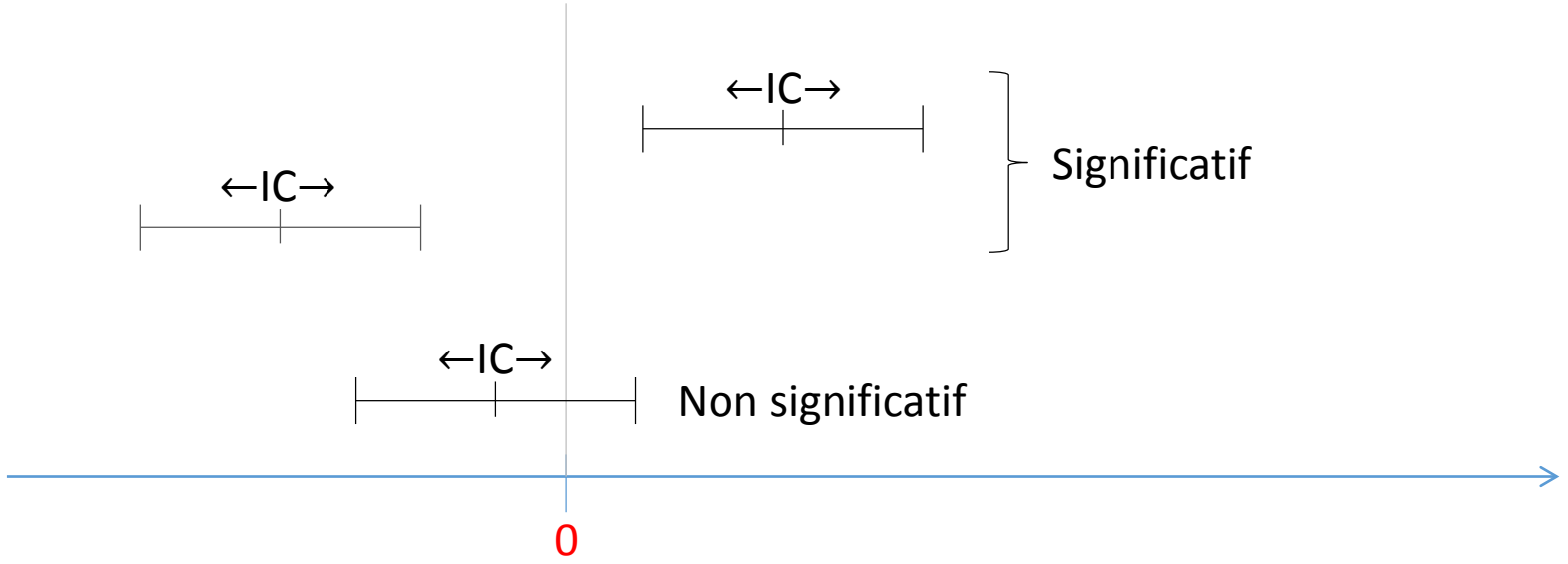
- Un intervalle de confiance de chaque β peut être calculé, la significativité de chaque relation peut être testée

Analyse multivariée

- Variable dépendante quantitative
  - Régression linéaire (! Conditions d'application)
  - Éventuellement à effet aléatoire ou modèle mixte
  - …
- Variable dépendante qualitative
  - Régression logistique
  - Régression multiniveau
  - …

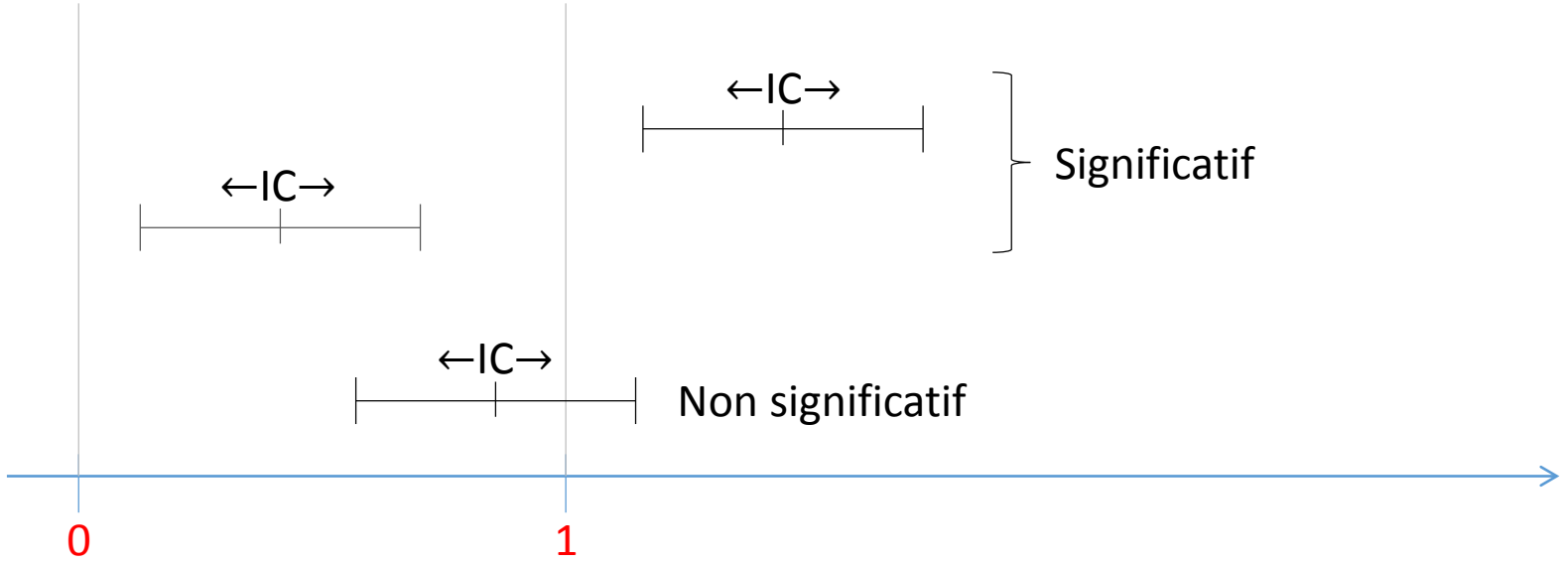
Significativité : intervalle de confiance

- La significativité d'une association peut s'apprécier grâce à l'intervalle de confiance de l'indice choisi pour mesurer la force de l'association
- Dans le cas d'un <u>coefficient, d'une différence de risque ou de moyenne</u>



Significativité : intervalle de confiance

- La significativité d'une association peut s'apprécier grâce à l'intervalle de confiance de l'indice choisi pour mesurer la force de l'association
- Dans le cas d'un odds ratio ou d'un risque relatif $(= e^{\beta})$



Données censurées : analyse de survie

- Variable à expliquer : durée avant apparition d'un évènement d'intérêt
- Suivi des participants : essai clinique ou étude de cohorte
- Date d'origine, de point, de dernières nouvelles, durée de suivi
- Censure
  - à droite (exclus vivants ou perdus de vue)
  - à gauche
  - par intervalle

  Indépendante du temps de survie (non informative ou aléatoire) ?
- Analyse de survie : éviter une perte d'information / censure

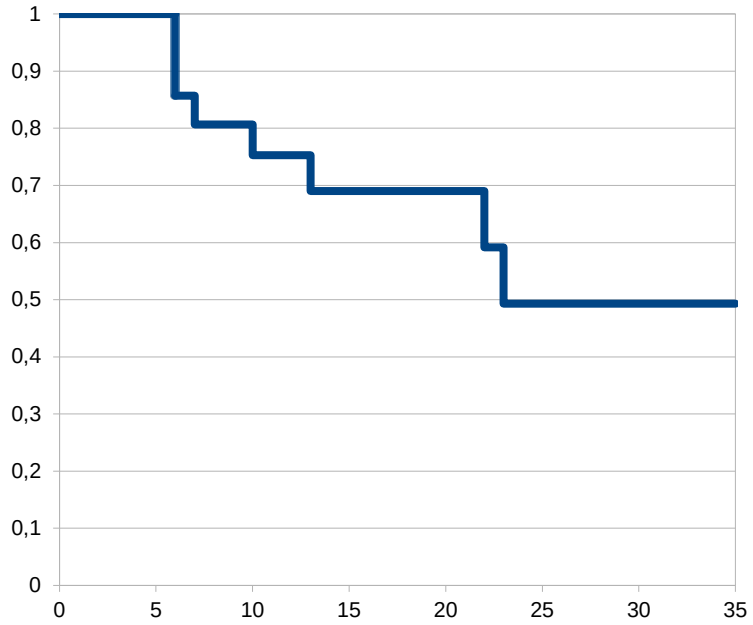
Données censurées : analyse de survie

- Deux questions :
  - Durée de « survie » différente selon les groupes étudiés ?
    - Courbes de survie : méthode de Kaplan Meier, méthode actuarielle…
    - Test de la différence : Log Rank…
  - Facteurs pronostiques (ou contrôle des facteurs de confusion) ?
    - Modèle de Cox…

Méthode de Kaplan Meier

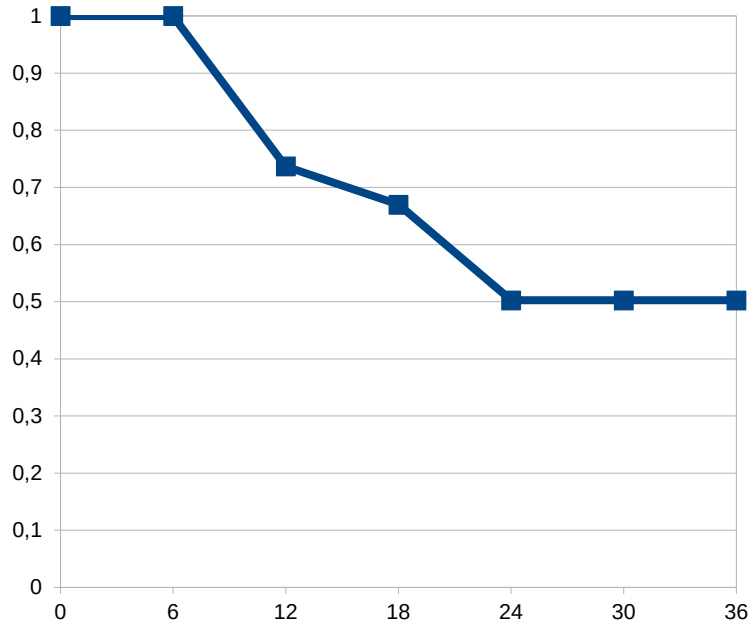
Suivi	Nb de survivants	Nb censure	Nb décès	Nb de patients à risque	Survie conditionnelle	Survie
T	N	C	D	X	Sc	S
0	21	0	0	21	1	1
]0-6]	21	0	3	21	0,86	0,86
]6-7]	18	1	1	17	0,94	0,81
]7-10]	16	1	1	15	0,93	0,75
]10-13]	14	2	1	12	0,92	0,69
]13-16]	11	1	0	10	1	0,69
]16-22]	10	3	1	7	0,86	0,59
]22-23]	6	0	1	6	0,83	0,49
]23-35]	5	0	0	5	1	0,49

X : nombre de sujets à risque juste avant T : N-C
Sc : 1 - (D/X)
S(T) : probabilité de survie après T : Sc x S(T-1)



Méthode actuarielle

Suivi	Nb de survivants	Nb censure	Nb décès	Nb de patients à risque	Survie conditionnelle	Survie
T	N	C	D	X	Sc	S
[0-6[	21	0	0	21	1	1
[6-12[	21	4	5	19	0,74	0,74
[12-18[	12	2	1	11	0,91	0,67
[18-24[	9	2	2	8	0,75	0,50
[24-30[	5	1	0	4,5	1	0,50
[30-36[	4	4	0	2	1	0,50



Log Rank

- Deux courbes de survie
  - 1 : groupe intervention ou exposition
  - 2 : groupe témoin
- H_0 : survie 1 ≈ survie 2
- Test du Log Rank : rejet de H_0 ($p < 10^{-3}$)