

6e Edition
KIDNEY COACHING CLUB
KCC 2017

L'ART DES STATISTIQUES

Olivier Aubert (MD; PhD) et Maud Racapé (PhD)



Paris Translational Research Center
for Organ Transplantation



PROGRAMME

Vendredi 24 novembre

12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

PROGRAMME

Vendredi 24 novembre

12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

Les statistiques : pourquoi faire ?

« La statistique est l'ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation de données d'observation relatives à un groupe d'individus ou d'unités. »

- La statistique utilise des règles et des méthodes sur la collecte des données, pour que celles-ci puissent être correctement interprétées, souvent comme composante d'une aide à la décision.

« L'épidémiologie est la science qui étudie, au sein de populations (humaines, animales, voire végétales), la fréquence et la répartition des problèmes de santé dans le temps et dans l'espace, ainsi que le rôle des facteurs qui les déterminent. »

La statistique: domaine des mathématiques et/ou science des données (en anglais : *Data Science*).

Composante théorique : s'appuie sur la théorie des probabilités et forme avec cette dernière les sciences de l'aléatoire

Composante appliquée : utilisée dans presque tous les domaines de l'activité humaine: ingénierie, management, économie, biologie, informatique, etc.

Les statistiques : pourquoi faire ?

But de ces 2 jours:

Aborder les problèmes que nous rencontrons en tant que statisticiens/épidémiologistes au cours de nos collaborations avec des cliniciens pour faciliter l'interaction future entre les 2

Notamment vous expliquer comment construire une base de donnée qui sera facilement traitable par le stat

Vous donner les notions de base pour faire des test stats simples par vous-mêmes

Vous expliquer ce que sont des régressions logistiques et des analyses de survie, pourquoi ça n'est pas la même chose et pourquoi une courbe de Kaplan Meier ne vous permet pas d'affirmer que votre variable est un facteur de risque indépendamment associé à votre outcome

Les statistiques : Interaction avec le statisticien

Raisons responsables de la difficulté d'interaction avec le statisticien

Liées au statisticien

Votre projet =
Travail supplémentaire

Pas de connaissance
sur votre sujet

N'est pas là pour
réfléchir pour vous

Liées à vous

Absence d'hypothèse
claire

Pas de demande
précise

Base de données
incorrectement faite

Attention: Un statisticien n'est
pas forcément un
épidémiologiste!

Souvent raison #1 de
l'absence de nouvelles du
statisticien

NB: Un statisticien ne « fait pas tourner les analyses » et « n'appuie pas sur un bouton »

PROGRAMME

Vendredi 24 novembre

12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

Créer une base de données : le B-A BA

- ✧ Une base de données, c'est bien; une hypothèse de travail, c'est mieux
- ✧ La base de donnée est le point de départ de tout travail: si elle est mal faite, elle risque de ne pas répondre à la question (et vous risquez d'attendre la réponse longtemps)
- ✧ Sélectionner les variables d'intérêt qui vous aiderons à répondre à la question de départ et connaître votre base est fondamental

✧ **Formats de la base de données:**

- SPSS data files
- excel
- csv
- dta (format stata importable sous R)

✧ **Les variables de la base:**

- Continues: l'âge, le poids, la taille... Peut prendre n'importe quelle valeur
- Catégorielles: Peut prendre un nombre limité et habituellement fixe de valeurs possibles
 - Dichotomiques (ou binaires): 2 niveaux ou catégories (ex: dialyse oui/non)
 - Nominale: ≥ 2 niveaux sans ordre intrinsèque (ex: dialyse non/DP/HD)
 - Ordinales: ≥ 2 niveaux avec ordre intrinsèque (ex: IRA grades 1 à 4)

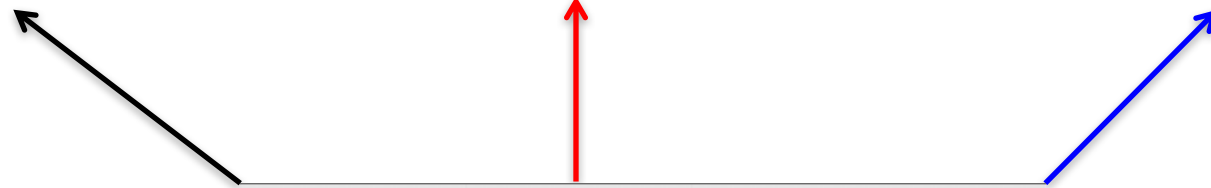
Créer une base de données : le B-A BA

✧ Les sorties STATA

Noir: numérique=nombre

Rouge: alphanumérique=texte

Bleu: libellés attribués à des variables numériques
(ici codage numérique initial en 0=homme et 1=femme)



Age_Receveur	type_donneur	Sexe_Receveur2
30.6694	VIV	M
25.90554	SME	M
53.10609	SME	M
43.08829	SME	M
42.40383	SME	M
48.59411	VIV	M
30.95414	SME	M
64.48186	SME	M
47.9206	SME	M
39.22519	SME	M
62.03696	SME	M
22.57632	CA	M
39.48255	SME	M
57.13895	SME	M
31.14853	SME	M



Créer une base de données : le B-A BA

- ✧ Pour commencer... la première ligne de votre base correspond à vos noms de variables
- ✧ C'est elle qui donnera le nom de vos variable dans le logiciel

Excel

A	B	C
XXX STUDY PATIENTS		
Age_Receveur	type_donneur	Sexe_Receveur2
30.6694	VIV	M
25.90554	SME	M
53.10609	SME	M
43.08829	SME	M
42.40383	SME	M
48.59411	VIV	M
30.95414	SME	M
64.48186	SME	M
47.9206	SME	M
39.22519	SME	M
62.03696	SME	M
22.57632	CA	M
39.48255	SME	M
57.13895	SME	M
31.14853	SME	M



STATA

XXX STUDY PATIENTS	B	C
Age_Receveur	type_donneur	Sexe_Receveur2
30.6694	VIV	M
25.90554	SME	M
53.10609	SME	M
43.08829	SME	M
42.40383	SME	M
48.59411	VIV	M
30.95414	SME	M
64.48186	SME	M
47.9206	SME	M
39.22519	SME	M
62.03696	SME	M
22.57632	CA	M
39.48255	SME	M
57.13895	SME	M
31.14853	SME	M

Créer une base de données : le B-A BA

✧ On ne fait pas de sous catégories de la variable

A	B	C	D	E	F
Tx Cardiaque			Tx Rénale		
Age_Receveur	type_donneur	Sexe_Receveur2	Age_Receveur	type_donneur	Sexe_Receveur2
30.6694	VIV	M	31.14853	SME	M
25.90554	SME	M	51.14853	SME	M
53.10609	SME	M	51.718	CA	M
43.08829	SME	M	61.52772	SME	M
42.40383	SME	M	59.80835	SME	M
48.59411	VIV	M	44.41068	VIV	M
30.95414	SME	M	23.22793	VIV	M
64.48186	SME	M	26.25325	SME	M
47.9206	SME	M	53.65092	SME	F
39.22519	SME	M	41.74675	SME	F
62.03696	SME	M	72.6653	SME	M
22.57632	CA	M	35.72895	VIV	M
39.48255	SME	M	24.59138	SME	M
57.13895	SME	M	64.87064	SME	M

Excel

STATA

TxCardiaque	B	C	TxRénale	E	F
Age_Receveur	type_donneur	Sexe_Receveur2	Age_Receveur	type_donneur	Sexe_Receveur2
30.6694	VIV	M	31.14853	SME	M
25.90554	SME	M	51.14853	SME	M
53.10609	SME	M	51.718	CA	M
43.08829	SME	M	61.52772	SME	M
42.40383	SME	M	59.80835	SME	M
48.59411	VIV	M	44.41068	VIV	M
30.95414	SME	M	23.22793	VIV	M
64.48186	SME	M	26.25325	SME	M
47.9206	SME	M	53.65092	SME	F
39.22519	SME	M	41.74675	SME	F
62.03696	SME	M	72.6653	SME	M
22.57632	CA	M	35.72895	VIV	M
39.48255	SME	M	24.59138	SME	M
57.13895	SME	M	64.87064	SME	M
31.14853	SME	M	31.14853	SME	M



Créer une base de données : le B-A BA

✧ Vos sous catégories correspondent à une nouvelle variable.

Excel

A	B	C	D
Age_Receveur	type_donneur	Sexe_Receveur2	Transplantation
30.6694	VIV	M	Cardiaque
25.90554	SME	M	Cardiaque
53.10609	SME	M	Cardiaque
43.08829	SME	M	Cardiaque
42.40383	SME	M	Cardiaque
48.59411	VIV	M	Cardiaque
30.95414	SME	M	Cardiaque
64.48186	SME	M	Cardiaque
47.9206	SME	M	Cardiaque
39.22519	SME	M	Cardiaque
62.03696	SME	M	Cardiaque
22.57632	CA	M	Cardiaque
39.48255	SME	M	Cardiaque
57.13895	SME	M	Cardiaque
31.14853	SME	M	Rénale
51.14853	SME	M	Rénale
51.718	CA	M	Rénale
61.52772	SME	M	Rénale
59.80835	SME	M	Rénale
44.41068	VIV	M	Rénale
23.22793	VIV	M	Rénale
26.25325	SME	M	Rénale
53.65092	SME	F	Rénale
41.74675	SME	F	Rénale
72.6653	SME	M	Rénale
35.72895	VIV	M	Rénale
24.59138	SME	M	Rénale
64.87064	SME	M	Rénale



STATA

Age_Receveur	type_donneur	Sexe_Recev~2	Transplant~n
30.6694	VIV	M	Cardiaque
25.90554	SME	M	Cardiaque
53.10609	SME	M	Cardiaque
43.08829	SME	M	Cardiaque
42.40383	SME	M	Cardiaque
48.59411	VIV	M	Cardiaque
30.95414	SME	M	Cardiaque
64.48186	SME	M	Cardiaque
47.9206	SME	M	Cardiaque
39.22519	SME	M	Cardiaque
62.03696	SME	M	Cardiaque
22.57632	CA	M	Cardiaque
39.48255	SME	M	Cardiaque
57.13895	SME	M	Cardiaque
31.14853	SME	M	Rénale
51.14853	SME	M	Rénale
51.718	CA	M	Rénale
61.52772	SME	M	Rénale
59.80835	SME	M	Rénale
44.41068	VIV	M	Rénale
23.22793	VIV	M	Rénale
26.25325	SME	M	Rénale
53.65092	SME	F	Rénale
41.74675	SME	F	Rénale
72.6653	SME	M	Rénale
35.72895	VIV	M	Rénale
24.59138	SME	M	Rénale
64.87064	SME	M	Rénale

Créer une base de données : le B-A BA

✧ Les logiciels de traitement de données ne reconnaissent pas les couleurs

Excel

A	B	C
Homme	Femme	
Age_Receveur	type_donneur	Age_Donneur
25.90554	SME	20
26.78166	VIV	28
30.6694	VIV	28
30.95414	SME	37
41.74675	SME	47
41.859	VIV	33
42.40383	SME	37
43.08829	SME	32
48.59411	VIV	57
51.03901	SME	50
53.10609	SME	55
53.65092	SME	54
55.09651	SME	48
60.28748	SME	61
64.48186	SME	79



STATA

Homme	Femme	C
Age_Receveur	type_donneur	Age_Donneur
25.90554	SME	20
26.78166	VIV	28
30.6694	VIV	28
30.95414	SME	37
41.74675	SME	47
41.859	VIV	33
42.40383	SME	37
43.08829	SME	32
48.59411	VIV	57
51.03901	SME	50
53.10609	SME	55
53.65092	SME	54
55.09651	SME	48
60.28748	SME	61
64.48186	SME	79

Créer une base de données : le B-A BA

✧ A EVITER!!

Variables numériques:

- Attention à ne pas mélanger les carottes et les patates: un poids rentré à la fois en kg et en gr est inutilisable! (si si, la créatinine en mg/L et en umol/mol dans la même variable, on a vu)

- Tout caractère entré dans une variable supposément numérique telle que le poids la rend alphanumérique (texte), ce qui oblige le stat à trouver tous les caractères qui empêchent la variable d'être numérique et à la transformer manuellement au lieu que le processus soit automatique d'excel vers stata
Il préférera que la case soit laissée vide.

Age_Receveur
25.90554
26.78166
30.6694
30.95414
41.74675
41.859
42.40383
43.08829
48.59411
51.03901
53.10609
53.65092
55.09651
60.28748
64.48186



Age_Receveur
25.90554
26.78166
30.6694
30.95414
41.74675
41.859
42.40383
43.08829
48.59411
51.03901
?
53.65092
55.09651
60.28748
64.48186



Créer une base de données : le B-A BA

✧ Donc: on choisit : soit numérique, soit alphanumérique!

Exemple de base reçue:

M1	M2	M3	M6	M9	M12	M24
4	1.9	1.7	1.6	1.6	1.3	1.2
.7	.6	.7	.7	.7	.7	.8
1.2	1.2	1.1	1	1	1	.9
1.3	1.5	1.3	1.2	1.3	1.2	1.1
1.6	1.4	1.5	1.35	1.32	.9	.
4.8	2.9	2.2	2.4	2.3	2.3	.
1.4	1.1	1.3	1.3	1.5	1.3	.
1.8	1.4	1.5	1.4			.
2.3	.9	.9				.
1.9	1.5	1.6				.
1.5	.9	.7	.9	.8	1	.9
1.6	1.5	1.6	1.4	1.1	1.1	.
1.2	1.3	1.6	2	4.1	6.1	.
1.6	1.4	1.7	NA	NA	NA	.
1.7	1.6	1.5	1.48			.
1.9	1.9	2.3				.
1.1	.	.				.

Créer une base de données : le B-A BA

✧ **Donc: on choisit : soit numérique, soit alphanumérique!**

Exemple de base reçue:

Excel

CIT
22 h 55 min
21 hours 5 minutes
25 hours 38 minutes
23 hours 31 minutes
24 hours 10 minutes
15 hours 52 minutes
12 hours 47 minutes
24 hours 21 minutes
25 hours 48 minutes
19 hours 5 minutes
19.72 hours
17,32
14.37 hours
5.9 hours
13.067 hours

STATA

CIT
22 h 55 min
21 hours 5 minutes
25 hours 38 minutes
23 hours 31 minutes
24 hours 10 minutes
15 hours 52 minutes
12 hours 47 minutes
24 hours 21 minutes
25 hours 48 minutes
19 hours 5 minutes
19.72 hours
17.32
14.37 hours
5.9 hours
13.067 hours

= variable inexploitable



NB: Le temps d'ischémie froide est habituellement donné en minutes (et les délais en jours)
Il est aisé de transformer la variable de base en heures, jours ou autre tant qu'elle est bien codée

Créer une base de données : le B-A BA

✧ A EVITER!!

Variables textes:

- Habituellement inutilisable telle quelle par le statisticien, doit être résumée/simplifiée
- 2 solutions pour le stat:
 - Recoder la variable pour que chaque texte corresponde à une catégorie (ici 9 catégories différentes pour 14 lignes)
→ Inexploitable si chaque cellule comprend du texte différent
 - Recoder la variable en interprétant et en faisant des regroupements si nécessaire (ici, nécessaire...)
→ Ça n'est pas le boulot du statisticien!

Causededécès
Défaillance multi-organes
Septicémie
Accident vasculaire cérébral (ischémique ou hémorragique)
Défaillance non spécifique du greffon
choc cardiogénique
Défaillance multi-organes
Septicémie
Infection fongique ou à protozoaire d'autres organes
choc cardiogénique/dysfonction greffon
Septicémie
Septicémie
Défaillance primaire
Septicémie
sepsis: récupérer le CRH du CHU de Montpellier

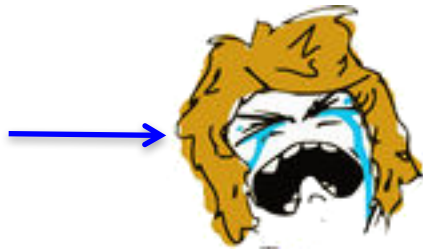


Créer une base de données : le B-A BA

✧ A EVITER!!

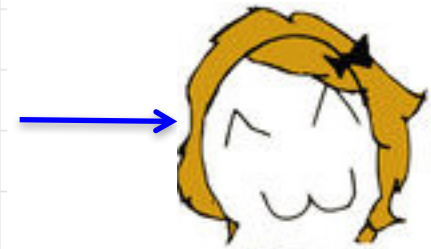
Les variables qui devraient être catégorielles...Mais ne le sont pas:

Sexe
Homme
Femme
femme
masculin
homme
homme
Masculin
féminin
Masculin
fem
0
1
1
0



Stat

Sexe	Sexe
Homme	0
Femme	1
Femme	1
Homme	0
Homme	0
Homme	0
Homme	0
Femme	1
Homme	0
Femme	1
Homme	0
Femme	1
Femme	1
Homme	0



Stat

Avec les labels
évidemment!

NB: Une variable binaire sera généralement codée en 0/1

Créer une base de données : le B-A BA

✧ Exemple DIVAT : Néphropathie initiale

Maladie initiale
Diabète de type I (insulino-dépendant)
Diabète de type I (insulino-dépendant)
Maladie polykystique rénale adulte (type dominant)
Maladie kystique de la médullaire (néphronophtise incluse)
Inconnue ou indéterminée
Néphro. e dépôts d'IgA (diagnostic prouvé par immunofluorescence)
Glomérulo. avec hyalinose segmentaire et focale de l'adulte (av.syndrôme néphrotique).
Maladie polykystique rénale adulte (type dominant)
Néphro.origine vasc.due e hypertension (absence néphropathie primitive)
Inconnue ou indéterminée
Inconnue ou indéterminée
Purpura rhumatoïde
Inconnue ou indéterminée
Hypoplasie rénale congénitale – type non spécifiée
Inconnue ou indéterminée
Diabète de type I (insulino-dépendant)
Néphropathie héréditaire / familiale non précisée
Maladie polykystique rénale adulte (type dominant)
Inconnue ou indéterminée
Pyélo. ou néphro. interstitielle cause non précisée
Inconnue ou indéterminée
Cystinose
Maladie de système
Lupus érythémateux
Diabète de type I (insulino-dépendant)
Pyélo. ou néphro. interstitielle due e lithiase des voies excrétrices urinaires
Tuberculose
Maladie polykystique rénale adulte (type dominant)
Néphro.origine vasc.due e hypertension (absence néphropathie primitive)
Maladie polykystique rénale adulte (type dominant)
Diabète de type I (insulino-dépendant)
Inconnue ou indéterminée
Glomérulo. avec hyalinose segmentaire et focale de l'adulte (av.syndrôme néphrotique).
Inconnue ou indéterminée
Néphro. e dépôts d'IgA (diagnostic prouvé par immunofluorescence)
Inconnue ou indéterminée
Inconnue ou indéterminée
Maladie polykystique rénale adulte (type dominant)
Maladie polykystique rénale adulte (type dominant)
Glomérulo. avec hyalinose segmentaire et focale de l'adulte (av.syndrôme néphrotique).
Glomérulo. membrano-proliférative type II avec dépôts denses (prouvée par immunofluorescence et/ou microscopie électronique)
Autre
Maladie polykystique rénale adulte (type dominant)
Inconnue ou indéterminée
Maladie polykystique rénale adulte (type dominant)
Inconnue ou indéterminée
Néphro. e dépôts d'IgA (diagnostic prouvé par immunofluorescence)
Reins polykystiques – type non spécifiée
Inconnue ou indéterminée
Maladie polykystique rénale adulte (type dominant)
Glomérulo. avec hyalinose segmentaire et focale de l'adulte (av.syndrôme néphrotique).
Glomérulonephrite, sans examen histologique
Sclérodermie / sclérose systémique
Inconnue ou indéterminée
Inconnue ou indéterminée
Maladie polykystique rénale adulte (type dominant)
Inconnue ou indéterminée

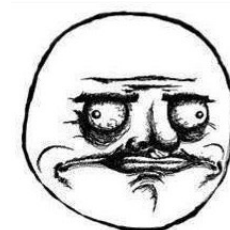


Créer une base de données : le B-A BA

✧ Exemple DIVAT : Néphropathie initiale --> Ensemble des données

. table Maladieinitiale

Maladieinitiale	Freq.		
Amyloidose	22	Néphropathie de reflux	2
Autre	111	Néphropathie due a la ciclosporine A	29
Autre affection renale identifiée	232	Néphropathie due au cis-platinum	3
Autre maladie renale identifiée et précisée	5	Néphropathie due aux analgésiques	5
Cystinose	5	Néphropathie due à un autre médicament	24
Diabète de type I (insulino-dépendant)	285	Néphropathie endémique des Balkans	3
Diabète type II (non insulino-dépendant)	307	Néphropathie héréditaire - autre	8
Dysplasie rénale congénitale avec ou sans	29	Néphropathie héréditaire / familiale non	15
Glomérulo. avec diagnostic histologique	186	Néphropathie héréditaire avec surdité (s)	62
Glomérulo. avec hyalinose segmentaire et	126	Néphropathies tubulo- interstitielles ac	2
Glomérulo. extra-membraneuse	71	Oxalose primitive	7
Glomérulo. membrano-prolif., type I (par	32	Perte de rein origine trauma. ou chirurg	12
Glomérulo. membrano-proliférative type I	55	Polyangéite microscopique et glomérulone	1
Glomérulo. progressive en l'absence de m	6	Purpura rhumatoïde	33
Glomérulonephrite post-infectieuse	2	Pyélo. ou néphro. interstitielle cause	71
Glomérulonephrite, sans examen histologi	152	Pyélo. ou néphro. interstitielle associée	6
Goutte	2	Pyélo. ou néphro. interstitielle due à a	36
Granulomatose de Wegener	14	Pyélo. ou néphro. interstitielle due à l	26
Hypoplasie rénale avec oligoméganéphroni	6	Pyélo. ou néphro. interstitielle due à r	121
Hypoplasie rénale congénitale - type non	63	Pyélo. ou néphro. interstitielle due à u	27
Inconnue ou indéterminée	800	Pyélo. ou néphro. interstitielle due à ur	75
Insuf. rénale chron. terminale d'étiologie	155	Reins polykystiques - autre variété reco	5
Lupus érythémateux	99	Reins polykystiques - type non spécifiée	85
Maladie de Berger	38	Sclérodémie / sclérose systémique	6
Maladie de Fabry	6	Synd. néphrotique sévère avec sclérose f	41
Maladie de système	16	Syndrôme d'agénésie des muscles abdomina	5
Maladie kystique de la médullaire (néphr	27	Syndrôme de Drash ou de Frasier	1
Maladie polykystique rénale adulte (type	619	Syndrôme de Goodpasture	16
Maladie polykystique rénale enfant (type	6	Syndrôme hémolytique et urémique typique	2
Maladie vasculaire rénale - origine conn	61		
Mégauretère	3		
Myélome/maladie des chaînes légères	6		



= Travail de recodage exténuant!!

Néphro. origine vas. due à artérite diffus	9	Néphro. origine vas. due à hypertension (	196	Néphroblastome	2	Uropathies et hypodysplasies sans précises	1
Néphrocalcinose ou néphro. due à hyperca	5	Néphropathie de reflux	2			Valves de l'urètre	2



Créer une base de données : le B-A BA

✧ ATTENTION!

- Les dates: (excel vs STATA)



Grosse prise
de tête les
dates...

A	B
Date1	Date2
20/10/04	8/22/2004
22/10/04	02/01/04
24/10/04	04/01/04
24/10/04	17/01/04
26/10/04	1994
28/10/04	27/01/04
02/11/04	27/01/04
05/11/04	09/02/04
06/11/04	1/1/2003
06/11/04	14/10/04
18/11/04	2/10/2004
01/12/04	11/5/2004

Date1	Date2
10/20/2004	8/22/2004
10/22/2004	1/2/2004
10/24/2004	1/4/2004
10/24/2004	1/17/2004
10/26/2004	1994
10/28/2004	1/27/2004
11/2/2004	1/27/2004
11/5/2004	2/9/2004
11/6/2004	1/1/2003
11/6/2004	10/14/2004
11/18/2004	2/10/2004
12/1/2004	11/5/2004

Sélectionner la colonne
Clic droit
Format
Nombre

A	B
Date1	Date2
20/10/04	8/22/2004
22/10/04	37988,00
24/10/04	37990,00
24/10/04	38003,00
28/10/04	38013,00
02/11/04	38013,00
05/11/04	38026,00
06/11/04	1/1/2003
06/11/04	38274,00
18/11/04	2/10/2004
01/12/04	11/5/2004

Format anglo-saxon
MDY

Homogénéiser les
formats de date est
très important!!!

Créer une base de données : le B-A BA

- ✧ Lors de l'import d'une base, n'envoyer que la base. L'ensemble de vos calculs ou figures ne peuvent être reconnus.

	JA					
70728107		18/07/2015	IA		CAD	
200613753	RL	15/06/2016	IIB		CAD	
200344171	ZF	08/07/2016	IIB		CAD	
	MP					
200255420		08/08/2016	IA		CAD	
200401140	GA	10/09/2016	IA		CAD	
	DA				CAD	
200500504		29/09/2016	IA			
71433838	DA	20/10/2016	IA		CAD	

As of 11/2/17 the following are number of patients for GIA = 31; GIB = 19; GIIA = 10; GIIB = 30

Sera reconnu par le logiciel comme le résultat de la variable de la première colonne!

Créer une base de données : le B-A BA

✧ **Toujours avoir un **identifiant unique** pour vos patients dans votre base et ne pas le changer**

- Cela permet d'envoyer une base anonyme
- Cela permet de fusionner les bases,
 - en cas d'ajout d'une variable notamment
 - en cas de correction ponctuelle à donner au statisticien pour un patient

✧ **Attention, la modification d'une donnée pour un patient peut changer tous les résultats et obliger le statisticien à tout recommencer!**

- Il faut vous assurer que votre base est finale
- Cela impose une vraie discipline pour la vérification de votre base

✧ EXEMPLES DE BASE DE DONNEES: do's and don'ts



Paris Transplant Group
Kidney, Heart & Lung

Créer une base de données : le B-A BA

LFH Co	Flexion dorsale	Flexion plantaire
5	15	15
5	15	30
5	10	30
4	30	50
3	0	30
4	10	30
5	10	45
3	30	15
5	15	30
5	35	35
5	0	45
5	20	10

Créer une base de données : le B-A BA

1. Pas de reconnaissance des couleurs

LFH Co	Flexion dorsale	Flexion plantaire
5	15	15
5	15	30
5	10	30
4	30	50
3	0	30
4	10	30
5	10	45
3	30	15
5	15	30
5	35	35
5	0	45
5	20	10

Créer une base de données : le B-A BA

Colonne4 ▼	Colonne5 ▼	Colonne6 ▼	Colonne7 ▼	Colonne8 ▼	Colonne9 ▼	Colonne10 ▼	Colonne11 ▼	Colonne12 ▼	Colonne13 ▼
10,9	1	oui	1++	10	appui	11	-4		NON
15,6	1	oui	1 2	4,9	appui	15	-2,6		NON
14,9	2	oui	1++	0,8	2mm	15	-2		NON
17,9	1	oui	1++	-3		16	-3,1		NON
18,9	1	oui	1++	4,2	5,8mm	16	-1		NON
16,2	2	oui	1 2	3	2mm	17	-1		NON
13,6	1	non	1	4,8	appui	18	1		NON
10,9	2	oui	1	0,6	appui	18	1		NON
7,9	2	oui	1	2,7	3,5mm	18	1,8		NON
7	1	oui	1 2	2	1mm	18	0,5		NON
14,7	1	oui	1 2	2	appui	19	1,5		NON
12,1	1	oui	1 2	4,9	2mm	20	2,4		mobilisation
12,6	2	oui	1	0,9	appui	22	0,9		NON
5,7	2	oui	1	1,1	appui	23	0,5		NON
18,1	1	oui	1 2	14,6	appui	25	3,6		NON
12,3	1	oui	1 2	0	appui	25	3,2		NON
7 2	1		1 2	1 2	appui	25	2 2		NON

Créer une base de données : le B-A BA

2. Le statisticien n'est pas dans la tête du créateur de la base. Les noms des variables doivent être clairs

Colonne4	Colonne5	Colonne6	Colonne7	Colonne8	Colonne9	Colonne10	Colonne11	Colonne12	Colonne13
10,9	1	oui	1++	10	appui	11	-4		NON
15,6	1	oui	1 2	4,9	appui	15	-2,6		NON
14,9	2	oui	1++	0,8	2mm	15	-2		NON
17,9	1	oui	1++	-3		16	-3,1		NON
18,9	1	oui	1++	4,2	5,8mm	16	-1		NON
16,2	2	oui	1 2	3	2mm	17	-1		NON
13,6	1	non	1	4,8	appui	18	1		NON
10,9	2	oui	1	0,6	appui	18	1		NON
7,9	2	oui	1	2,7	3,5mm	18	1,8		NON
7	1	oui	1 2	2	1mm	18	0,5		NON
14,7	1	oui	1 2	2	appui	19	1,5		NON
12,1	1	oui	1 2	4,9	2mm	20	2,4		mobilisation
12,6	2	oui	1	0,9	appui	22	0,9		NON
5,7	2	oui	1	1,1	appui	23	0,5		NON
18,1	1	oui	1 2	14,6	appui	25	3,6		NON
12,3	1	oui	1 2	0	appui	25	3,2		NON
7 3	1		1 2	1 2	appui	25	3 2		NON

Créer une base de données : le B-A BA

er	hypo	hyper	hyper	PPV keller deficit exten
	hypo	normal	hyper	Rigido-valgus
al	hypo	hyper	hypo	Rigido-valgus (Weil fait)
	hypo	hyper	hypo	Récidive HV + HR (Weil
			Résultats	
er = >30				
al = 16<N<30		Coté	Pourcentage	
i = <16		D	53,33	
		G	48,28	
		Arrière Pied	Pourcentage	
		Valgus	40,00	
		Varus	36,67	
		Normal	23,33	
		Position Frontale	Pourcentage	
		Parallèle	43,33	
		Pronation	40,00	
		Supination	16,67	
		Position Horizontale	Pourcentage	
		Normoaxé	63,33	
		Varus / Ecartement 1-2	20,00	
		Valgus / Conflit 1-2	16,67	



Créer une base de données : le B-A BA

3. Impossibilité d'utiliser des « morceaux » de base

er	hypo	hyper	hyper	PPV keller deficit exten
	hypo	normal	hyper	Rigido-valgus
al	hypo	hyper	hypo	Rigido-valgus (Weil fait)
	hypo	hyper	hypo	Récidive HV + HR (Weil
			Résultats	
er = >30				
al = 16<N<30		Coté	Pourcentage	
i = <16		D	53,33	
		G	48,28	
		Arrière Pied	Pourcentage	
		Valgus	40,00	
		Varus	36,67	
		Normal	23,33	
		Position Frontale	Pourcentage	
		Parallèle	43,33	
		Pronation	40,00	
		Supination	16,67	
		Position Horizontale	Pourcentage	
		Normoaxé	63,33	
		Varus / Ecartement 1-2	20,00	
		Valgus / Conflit 1-2	16,67	

Créer une base de données : le B-A BA

AF
Colonne9
appui
appui
2mm
5,8mm
2mm
appui
appui
3,5mm
1mm
appui
2mm
appui
appui
appui
appui
appui
appui
appui
1mm
14,6mm
2,7mm
5mm
appui
3,5mm
5,8mm
2mm

Créer une base de données : le B-A BA

4. On choisit : numérique ou alphanumérique, et on ne mélange pas les unités!

AF
Colonne9 ▼
appui
appui
2mm
5,8mm
2mm
appui
appui
3,5mm
1mm
appui
2mm
appui
appui
appui
appui
appui
appui
appui
appui
1mm
14,6mm
2,7mm
5mm
appui
3,5mm
5,8mm
2mm

Créer une base de données : le B-A BA

	A	B	C	D	E	F	G	H	I
1	Age Tx thorax			Délais IRCT			Modalités EER		
2	Cœur	Poumons CP	Tous	Cœur	Poumons CP	Tous	Cœur	Poumons CP	Tous
3	48	30	48	32	63	32	Both	HD	Both
4	56	25	56	88	112	88	HD	HD	HD
5	45	20	45	153	12	153	HD	HD	HD
6	40	33	40	136	0	136	HD	HD	HD
7	46	19	46	116	180	116	HD	HD	HD
8	34	17	34	56	84	56	Both	HD	Both

Créer une base de données : le B-A BA

✧ On ne fait pas de sous catégories de la variable

	A	B	C	D	E	F	G	H	I
1	Age Tx thorax			Délais IRCT			Modalités EER		
2	Cœur	Poumons CP	Tous	Cœur	Poumons CP	Tous	Cœur	Poumons CP	Tous
3	48	30	48	32	63	32	Both	HD	Both
4	56	25	56	88	112	88	HD	HD	HD
5	45	20	45	153	12	153	HD	HD	HD
6	40	33	40	136	0	136	HD	HD	HD
7	46	19	46	116	180	116	HD	HD	HD
8	34	17	34	56	84	56	Both	HD	Both

Créer une base de données : le B-A BA

✧ Vos sous catégories correspondent à une nouvelle variable.

	A	B	C	D	E
1	ID	ORGANE	AGE_TX_THO	DELAIS_IRCT	EER
2	7	CŒUR	1	273	DP
3	12	CŒUR	1	249	HD
4	45	CP	8	71	HD
5	49	CP	10	148	HD
6	10	CŒUR	11	185	HD
7	15	CŒUR	12	228	HD
8	13	CŒUR	13	0	HD
9	51	CP	16	7	HD
10	43	CP	17	84	HD
11	46	CP	17	117	HD
12	50	CP	17	122	HD

Créer une base de données : le B-A BA

129	TTT X	TTT Y				
130	Variable A			Variable B		
131	A	B	C	D	E	F
132	0	0	0	0	0	0
133	0	0	0	0	0	0
134	ans) puis pAM	0	ans) puis pAM	1	0	1
135	0	0	0	1	1	1
136	0	0	0	1	0	1
137	0	0	0	0	1	0
138	0	0	0	0	0	0

Créer une base de données : le B-A BA

✧ Les logiciels de traitement de données ne reconnaissent pas les couleurs

129	TTT X	TTT Y				
130	Variable A			Variable B		
131	A	B	C	D	E	F
132	0	0	0	0	0	0
133	0	0	0	0	0	0
134	ans) puis pAM	0	ans) puis pAM	1	0	1
135	0	0	0	1	1	1
136	0	0	0	1	0	1
137	0	0	0	0	1	0
138	0	0	0	0	0	0

Créer une base de données : le B-A BA

✧ EXEMPLES DE BASE DE DONNEES: do's and don'ts

rein_greff~G	Greffemult~e	BIGREFFE	SexeDonneur	AgeDonneur	Taillee~neur	PoidsenkG~r	HTA_D_FINAL	DIABETE_D_~L	CREAT_D_FI~L	IF_MIN_FINAL	ECD	AB0Donneur	HLAA1Donneur	HLAA2Donneur	HLAB1Donneur	HLAB2Donneur	HLADR1Donn~r	HLADR2Donn~r	HLADQ1Donn~r	HLA
RD	N	0	M	47	170	76	0	0	47	1280	N	0	2	32	44	62	11	13	6	
RD	N	0	F	54	170	80	0	N	47	690	0	0	1	68	8	14	8	17	2	
RG	N	0	F	75	155	64	0	N	69	1380	0	B	1	11	13	37	10	12	1	
RG	N	0	M	50	175	75	0	N	88	630	N	A	3	25	18	35	1	15	5	
RD	N	0	F	61	170	68	0	N	45	1079	0	0	2	68	35	62	11	13	7	
RG	N	0	F	43	169	53	N	N	70	120	N	0	3	0	15	35	4	13	3	
RD	N	0	F	82	162	56.4	0	N	61	885	0	0	1	3	7	8	15	95	1	
RG	N	0	M	53	170	80	N	N	114	1080	N	A	1	11	8	60	4	17	2	
RG	N	0	M	43	165	82	N	N	80	1970	N	0	1	2	51	8	3	0	.	
RG	N	0	M	69	180	85	N	N	68	1070	0	0	2	68	18	51	11	17	1	
RG	N	0	M	30	172	60	N	N	131	780	N	A	3	26	38	39	4	14	1	
RG	N	0	M	44	180	80	N	N	79	120	N	0	24	26	8	38	3	13	2	
RG	N	0	F	67	155	70	0	N	56	120	N	B	1	2	44	57	4	7	3	
RG	N	0	F	63	158	62	0		71	1170	0	0	2	0	38	72	13	16	1	
RM	N	1	F	66	170	52		0	86	1350	0	0	2	25	27	57	7	0	2	
RD	N	0	M	37	178	75	N	N	126	1755	N	0	24	0	7	18	7	17	.	
RG	N	0	F	55	160	64	N	N	41	990	N	0	2	24	44	27	4	7	2	
RD	N	0	M	54	186	83	N	N	168	800	0	A	2	26	18	0	4	8	3	
RD	N	0	F	44	155	44	N	N	44	780	N	A	29	31	44	39	7	11	2	
RD	N	0	M	58	180	70	N	N	200	1020	N	A	1	24	51	57	13	0	6	
RD	N	0	F	74	155	78.6	N	N	56	840	0	A	3	31	7	35	4	15	3	
RG	N	0	F	70	180	80	N	N	42	630	0	0	29	30	45	18	4	7	2	
RD	N	0	F	66	159	60	0	N	33	685	0	A	1	2	7	35	7	9	2	
RD	N	0	M	58	170	83	N	N	66	930	N	A	23	29	44	49	7	15	2	
RM	N	1	F	83	164	65	N	N	77	960	0	A	2	0	51	49	4	11	7	
RD	N	0	M	50	185	90	N	N	82	1200	N	0	2	3	35	58	13	0	1	
RG	N	0	F	82	155	56	N	N	101	1298	0	B	2	0	44	51	4	16	5	
RG	N	0	M	50	180	67	N	N	88	120	N	A	2	0	27	51	4	8	3	
RD	N	0	M	67	175	100	N	N	73	860	0	A	1	30	8	13	7	17	2	
RD	N	0	M	42	176	75	N	N	74	1180	N	B	3	0	7	0	1	15	5	
RG	N	0	F	54	163	65	N	N	54	990	N	A	1	0	8	18	4	17	2	
RG	N	0	M	31	170	65	N	N	81	1040	N	A	26	31	44	62	11	0	.	
RD	N	0	F	40	165	113	0	N	49	1560	N	A	2	3	44	14	1	4	1	
RD	N	0	M	51	171	83			73	1110	N	A	1	2	37	57	4	0	8	
RD	N	0	F	59	153	70	0	N	63	1470	0	A	3	0	7	27	1	15	5	
RD	N	0	M	36	175	64	N	N	94	880	N	0	2	11	7	44	4	15	1	
RG	N	0	M	38	173	80	N	N	66	1715	N	0	2	0	52	44	4	7	2	
RG	N	0	F	36	162	84	N	N	54	120	N	A	29	9	50	53	13	0	.	
RG	N	0	F	53	165	57	N	N	77	120	N	0	2	29	7	50	7	13	2	
RD	N	0	M	60	180	90	N	N	78	1160	0	0	68	30	7	13	3	7	2	
RG	N	0	F	40	161	57.4	N	N	72	744	N	A	3	32	35	60	15	0	6	
RD	N	0	M	41	175	90	0	N	102	2056	N	0	26	29	35	38	1	11	5	
RG	N	0	F	62	170	70	N	N	48	1740	0	0	2	32	18	61	1	15	.	
RD	N	0	M	24	174	82	N	N	80	1560	N	A	2	24	18	45	8	14	5	
RM	N	1	M	73	170	64	0	N	110	720	0	B	2	0	7	51	4	15	3	
RD	N	0	M	64	175	70	N	N	120	1260	0	A	24	32	14	51	7	15	2	
RG	N	0	M	54	186	90	N	N	71	120	N	0	2	68	14	44	13	14	5	
RD	N	0	M	41	178	77	N	N	69	1460	N	A	3	29	35	38	1	7	1	
RG	N	0	M	45	180	75	0	N	62	930	N	0	2	26	62	0	17	13	.	
RG	N	0	M	48	175	60	N	0	90	650	N	0	2	23	35	58	7	15	2	
RG	N	0	M	56	180	110	0	N	189	1260	0	0	2	24	18	62	4	15	.	
RG	N	0	F	53	169	100	0	N	67	1065	0	A	3	68	44	49	4	7	2	
RD	N	0	F	55	150	120	0	N	115	820	0	A	24	30	13	51	7	13	2	
RG	N	0	M	57	175	82	N	N	98	1180	N	A	2	3	7	44	4	13	1	
RG	N	0	F	69	162	76	0	N	92	120	N	0	2	29	38	44	7	13	2	

PROGRAMME

Vendredi 24 novembre

12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

Comparer mes résultats

✧ **Déviatiion Standard:** racine carrée de la variance

✧ **Variance:** moyenne des différences au carré depuis la moyenne

(<http://scistatcalc.blogspot.fr/2013/12/levenes-test-equality-of-variances.html>)

Sous excel: F.TEST

✧ **Loi normale (distribution/loi Gaussienne):** vérification que les données suivent une loi normale – souvent visuel par histogramme

Sous GraphPad: Analyze → Column statistics → Shapiro-Wilk ou Kolmogorov-Smirnov

✧ **Outils:** Excel, GraphPad Prism ([QuickCalcs](#)) , BiostaTGV



BiostaTGV ne fait pas le test s'il y a des valeurs manquantes

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n_1 , n_2)

Exemple: Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n_1 > 30$ et $n_2 > 30$, égalités des variances: [Test de Student](#)

Sous Excel:

- Mettre des filtres à vos variables
- Trier par la catégorie qui vous intéresse, ici Sexe_Receveur (86 F 129 H)
- [Test d'égalité des variances](#): dans une colonne à part, tapez F.TEST.Ici > 0.05, donc les variances sont égales
- Dans une colonne à part, tapez « =TEST.STUDENT » et fx
- Dans la fenêtre générateur de code qui s'ouvre, matrice 1 correspond aux femmes, matrice 2 aux hommes, uni/bilatéral correspond à la différence attendue du test: si on est sûr que les receveurs hommes sont plus âgés par exemple, c'est unilatéral. En général on choisit bilatéral (pas de certitudes en biologie...), et type correspond au type de test, ici 2 puisque variances égales (3=variances inégales, 1=T-test apparié)
- La p-value du test sur la base complète est de 0,3095

Comparer mes résultats

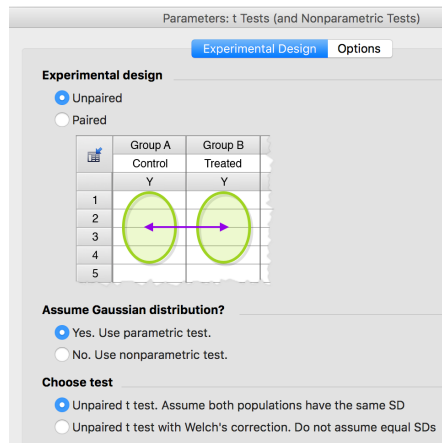
✧ Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)

Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n1 > 30$ et $n2 > 30$, égalités des variances: [Test de Student](#)

Sous GraphPad Prism:

- Choisir column, enter replicate values, stacked into columns → Create
- Entrer Age_Receveur pour les F dans Group A et pour les H dans Group B → Analyze → t tests → cocher les deux colonnes à comparer
- Cocher unpaired, parametric test, unpaired t test



Unpaired t test with equal SD	
Table Analyzed	Data 1
Column B	H
vs.	vs.
Column A	F
Unpaired t test	
P value	0.3095
P value summary	ns
Significantly different? (P < 0.05)	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.019 df=213

→ p-value identique

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)

Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n1 > 30$ et $n2 > 30$, égalités des variances: [Test de Student](#)

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

Tests statistiques

- Réaliser de nombreux tests statistiques
- Choisir le test le mieux adapté à vos données
- Comprendre toute la démarche d'un test



Tableau des tests statistiques d'hypothèse

Un test d'hypothèse est une démarche permettant d'évaluer la validité une hypothèse statistique en fonction d'un échantillon de données dont l'interprétation des résultats nécessite une bonne compréhension. Ce site n'a pas pour objectif de proposer un cours de statistique ni d'expliquer en détail cette démarche mais de permettre la réalisation de certains tests sans logiciel de statistique. Nous vous invitons donc fortement à consulter d'autres ressources pour bien comprendre les tests statistiques:

- Wikipedia Test d'hypothèses et Tests statistiques
- Des ouvrages comme ceux indiqués en page d'accueil

① Besoin d'aide pour choisir votre test ?

Calculez les statistiques de base d'une série de données

Analyse de survie ✨ nouveau !

Type de test à mettre en évidence			Variable de réponse			
① Type de test			Qualitative nominale (2 groupes)	Qualitative nominale (plus de 2 groupes)	Qualitative ordinale	Quantitative
Facteur d'étude	① Qualitatif (deux groupes)	Indépendants	Z de comparaison de proportions.* Chi² (χ².) Test exact de Fisher.	Chi² (χ².)	Test de Cochran-Armitage*	Test de Mann-Whitney. t de Student. Test de Welch.*
		Appariés	Test de McNemar. Test exact de Fisher.	Q de Cochran.*	Tests des signes.* Tests des rangs signés de Wilcoxon.	t de Student pour données appariées. Tests des rangs signés de Wilcoxon.
	① Qualitatif (plus de deux groupes)	Indépendants	Chi² (χ².)	Chi² (χ².)	Test de Kruskal-Wallis. (ordinal)	Analyse de la variance. Test de Kruskal-Wallis. (échelle quanti)
		Appariés	Q de Cochran.*	Q de Cochran.*	Test de Friedman.	Test de Friedman.
	Quantitatif		Régression logistique*	Régression logistique multinomiale*	Corrélation de Spearman. Tau de Kendall.	Corrélation de Pearson. Régression linéaire.*
	* : La réalisation de ces tests n'est actuellement pas disponible sur biostaTGV					

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)

Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n1 > 30$ et $n2 > 30$, égalités des variances: [Test de Student](#)

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

Tests statistiques

- Réaliser de nombreux tests statistiques
- Choisir le test le mieux adapté à vos données
- Comprendre toute la démarche d'un test



Tableau des tests statistiques d'hypothèse

Un test d'hypothèse est une démarche permettant d'évaluer la validité une hypothèse statistique en fonction d'un échantillon de données dont l'interprétation des résultats nécessite une bonne compréhension. Ce site n'a pas pour objectif de proposer un cours de statistique ni d'expliquer en détail cette démarche mais de permettre la réalisation de certains tests sans logiciel de statistique. Nous vous invitons donc fortement à consulter d'autres ressources pour bien comprendre les tests statistiques:

- Wikipedia Test d'hypothèses et Tests statistiques
- Des ouvrages comme ceux indiqués en page d'accueil

① Besoin d'aide pour choisir votre test ?

Calculez les statistiques de base d'une série de données

Analyse de survie ✨ nouveau !

Sexe_Receveur

Age_Receveur

Type de test à mettre en évidence		Variable de réponse			
Type de test		Qualitative nominale (2 groupes)	Qualitative nominale (plus de 2 groupes)	Qualitative ordinale	Quantitative
Facteur d'étude	Indépendants	Z de comparaison de proportions.* Chi² (χ².) Test exact de Fisher.	Chi² (χ².)	Test de Cochran-Armitage*	Test de Mann-Whitney. t de Student. Test de Welch.*
	Appariés	Test de McNemar. Test exact de Fisher.	Q de Cochran.*	Tests des signes.* Tests des rangs signés de Wilcoxon.	t de Student pour données appariées. Tests des rangs signés de Wilcoxon.
	Indépendants	Chi² (χ².) Q de Cochran.*	Chi² (χ².) Q de Cochran.*	Test de Kruskal-Wallis. (ordinal) Test de Friedman.	Analyse de la variance. Test de Kruskal-Wallis. (échelle quanti) Test de Friedman.
	Quantitatif	Régression logistique*	Régression logistique multinomiale*	Corrélation de Spearman. Tau de Kendall.	Corrélation de Pearson. Régression linéaire.*

* : La réalisation de ces tests n'est actuellement pas disponible sur biostaTGV

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n_1 , n_2)

Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n_1 > 30$ et $n_2 > 30$, égalités des variances: [Test de Student](#)

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

Tests statistiques

- Réaliser de nombreux tests statistiques
- Choisir le test le mieux adapté à vos données
- Comprendre toute la démarche d'un test



Quel est le nombre d'observations dans :

le groupe 1 ?

86

le groupe 2 ?

129

Envoyer

- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- Pour un test de Student considérer les variances égales et le test bilatéral: le résultat est le même avec une p-value de 0.3095

42.09171
43.0444
25.3088
50.9103

36,31759
42,091717
43,04449
25,308693
50,910336

Copiez vos données depuis Excel et collez-les ci-dessus

Générer

Retour

Effacer et recommencer

Entrez les valeurs pour chaque observations dans chacun des groupes.

Options du test

☒ Faire le test en considérant les variances des 2 groupes comme égales
(Attention, notre site **ne vérifie pas** cette hypothèse qu'il vous appartient de vérifier si cette case est cochée)

Alternative du test

☒ Bilatérale ☐ Supérieure ☐ Inférieure

Faire le test

ETAPE 4 : Prise de décision, acceptation ou rejet de H_0

```
array(3) { ["lines"]=> int(86) ["columns"]=> int(1) ["vector"]=> string(821)
"c(26.85832,39.151268,52.424366,50.461327,38.48049,49.48118,46.44764,40.8
) }
```

Résultats du test

- Données série 1: 86L x 1C, série 2: 129L x 1C
- Méthode : Two Sample t-test; Alternative : two.sided
- Statistique observée Qobs : -1.0186479018737
- p-value : 0.30952577355136
- T : Array Intervalle de confiance à 95%[-5.2188 ; 1.6626]
- Degrés de liberté : 213
- Moyenne : Groupe 1: 44.829091581395 ; Groupe 2: 46.607155744186

La valeur p (p-value) de votre test est 0.30952577355136.



Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)

Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi normale ou $n1 > 30$ et $n2 > 30$, égalités des variances: [Test de Student](#)

Sous STATA:

```
. ranksum Age_Receveur, by(Sexe_Receveur2)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

Sexe_Recev~2	obs	rank sum	expected
M	129	14370.5	13932
F	86	8849.5	9288
combined	215	23220	23220


```
unadjusted variance    199692.00
adjustment for ties      -0.48
-----
adjusted variance      199691.52
```



```
Ho: Age_Re~r(Sexe_R~2==M) = Age_Re~r(Sexe_R~2==F)
      z = 0.981
      Prob > |z| = 0.3265
```

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n_1, n_2), très petits échantillons ($n < 12$)

Exemple: comparaison de l'âge du receveur en fonction du sexe du receveur: [Test de Mann-Whitney](#)

- Prendre dans la base l'âge des 8 premières femmes et des 11 premiers hommes, les mettre les un en dessous des autres dans une autre colonne
- Les trier par ordre croissant (en gardant l'identification homme/femme) et leur attribuer un rang avec rang 1=plus petite valeur
- Retrier les valeurs suivant leur appartenance au groupe homme ou femme
- Calculer la somme des rangs pour chaque groupe homme ou femme
- Dans le tableau ci-dessous, regarder pour $n_1=8$ et $n_2=11$ la première colonne où le rang se trouve entre les 2 valeurs: la différence est significative à la p-value de la colonne d'avant

532 Appendix B

Table B10 The Mann-Whitney test (Wilcoxon two sample test)

For a comparison of two groups of size n_1 and n_2 , where $n_1 \leq n_2$, if the sum of the ranks in the smaller group is equal to the tabulated value or is outside the range shown (i.e. is not between the tabulated values), the P value of the test is less than the value at the top of the column.

Example: For a rank sum of 29 from a sample of 6 compared with a sample of 11 we look along the row for $n_1 = 6$ and $n_2 = 11$ from the left until we find the last column where the value 29 is not between the tabulated values, which gives $P < 0.02$. A rank sum of 28 would give $P < 0.01$.

n_1	n_2	0.1	Two-tailed probability (P)				0.001
			0.05	0.02	0.01		
3	3	6-15	-	-	-	-	-
3	4	6-18	-	-	-	-	-
4	4	11-25	10-26	-	-	-	-
2	5	3-13	-	-	-	-	-
3	5	7-20	6-21	-	-	-	-
4	5	12-28	11-29	10-30	-	-	-
5	5	19-36	17-38	16-39	15-40	-	-
2	6	3-15	-	-	-	-	-
3	6	8-22	7-23	-	-	-	-
4	6	13-31	12-32	11-33	10-34	-	-
5	6	20-40	18-42	17-43	16-44	-	-
6	6	28-50	26-52	24-54	23-55	-	-
2	7	3-17	-	-	-	-	-
3	7	8-25	7-26	6-27	-	-	-
4	7	14-34	13-35	11-37	10-38	-	-
5	7	21-44	20-45	18-47	16-49	-	-
6	7	29-55	27-57	25-59	24-60	-	-
7	7	39-66	36-69	34-71	32-73	28-77	-
2	8	4-18	3-19	-	-	-	-
3	8	9-27	8-28	6-30	-	-	-
4	8	15-37	14-38	12-40	11-41	-	-
5	8	23-47	21-49	19-51	17-53	-	-
6	8	31-59	29-61	27-63	25-65	21-69	-
7	8	41-71	38-74	35-77	34-78	29-83	-
8	8	51-85	49-87	45-91	43-93	38-98	-

Table B10 533

n_1	n_2	0.1	Two-tailed probability (P)				0.001
			0.05	0.02	0.01		
2	9	4-20	3-21	-	-	-	-
3	9	10-29	8-31	7-32	6-33	-	-
4	9	16-40	14-42	13-43	11-45	-	-
5	9	24-51	22-53	20-55	18-57	15-60	-
6	9	33-63	31-65	28-68	26-70	22-74	-
7	9	43-76	40-79	37-82	35-84	30-89	-
8	9	54-90	51-93	47-97	45-99	40-104	-
9	9	66-105	62-109	59-112	56-115	50-121	-
2	10	4-22	3-23	-	-	-	-
3	10	10-32	9-33	7-35	6-36	-	-
4	10	17-43	15-45	13-47	12-48	-	-
5	10	26-54	23-57	21-59	19-61	15-65	-
6	10	35-67	32-70	29-73	27-75	23-79	-
7	10	45-81	42-84	39-87	37-89	31-95	-
8	10	56-96	53-99	49-103	47-105	41-111	-
9	10	69-111	65-115	61-119	58-122	52-128	-
10	10	82-128	78-132	74-136	71-139	63-147	-
2	11	4-24	3-25	-	-	-	-
3	11	11-34	9-36	7-38	6-39	-	-
4	11	18-46	16-48	14-50	12-52	-	-
5	11	27-58	24-61	22-63	20-65	16-69	-
6	11	37-71	34-74	30-78	28-80	23-85	-
7	11	47-86	44-89	40-93	38-95	32-101	-
8	11	59-101	55-105	51-109	49-111	42-118	-
9	11	72-117	68-121	63-126	61-128	53-136	-
10	11	86-134	81-139	77-143	73-147	65-155	-
11	11	100-153	96-157	91-162	87-166	78-175	-
2	12	5-25	4-26	-	-	-	-
3	12	11-37	10-38	8-40	7-41	-	-
4	12	19-49	17-51	15-53	13-55	-	-
5	12	28-62	26-64	23-67	21-69	16-74	-
6	12	38-76	35-79	32-82	30-84	24-90	-
7	12	49-91	46-94	42-98	40-100	33-107	-
8	12	62-106	58-110	53-115	51-117	43-125	-
9	12	75-123	71-127	66-132	63-135	55-143	-
10	12	89-141	84-146	79-151	76-154	67-163	-
11	12	104-160	99-165	94-170	90-174	81-183	-
12	12	120-180	115-185	109-191	105-195	95-205	-

Exemple: pour une somme des rangs de 110, la 1^{ère} colonne où 110 se trouve entre les valeurs données est pour $p=0.01$. Donc les groupes sont différents avec $p=0.02$

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n_1 , n_2)

Exemple: Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi non normale et $n_1 < 30$ ou $n_2 < 30$, non égalités des variances: [Test de Mann-Whitney](#)

Sous GraphPad Prism:

- Analyse précédente: supprimer dans F et H (group A et B) toutes les valeurs après ligne 8
- Analyze → t tests (and non parametric tests) → cocher les deux colonnes à comparer → ok
- Analyze this table again...
- Cocher Unpaired, non parametric test, Mann-Whitney test → ok
- P-value 0.4331

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)

Exemple: Comparaison de l'âge du receveur en fonction du sexe du receveur

Loi non normale et $n1 < 30$ ou $n2 < 30$, non égalités des variances: **Test de Mann-Whitney**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

Type de test	Test de Mann-Whitney	Test de Fisher	Test de Fisher	Test de Fisher	Test de Fisher
Quantitatif (valeurs continues)	Test de Mann-Whitney	Test de Fisher	Test de Fisher	Test de Fisher	Test de Fisher
Quantitatif (valeurs discrètes)	Test de Mann-Whitney	Test de Fisher	Test de Fisher	Test de Fisher	Test de Fisher
Qualitatif	Test de Mann-Whitney	Test de Fisher	Test de Fisher	Test de Fisher	Test de Fisher

- Rentrer le nombre d'observations: 8 F 8 H
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- P-value de 0.4418 comparé à 0.4331 sous GraphPad

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes appariés (n_1 , n_2)

Exemple: Comparaison du BMI avant et après traitement

$n_1 > 30$ et $n_2 > 30$ ou distribution normale de la différence: [Test de Student apparié](#)

Sous Excel:

- Dans une colonne à part, tapez « =TEST.STUDENT » et fx
- Dans la fenêtre générateur de code qui s'ouvre, matrice 1 correspond aux femmes, matrice 2 au hommes, uni/bilatéral correspond à la différence attendue du test: si on est sûr que les receveurs hommes sont plus âgés par exemple, c'est unilatéral. En général on choisit bilatéral (pas de certitudes en biologie...), et type correspond au type de test, ici 1 puisque on veut un T-test apparié
- P-value 0.0011

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes appariés (n1, n2)

Exemple: Comparaison du BMI avant et après traitement

n1>30 et n2>30 ou distribution normale de la différence: Test de Student apparié

Sous GraphPad Prism:

- Analyse précédente, copier-coller BMI_R_Before dans Group A et BMI_R_After dans Group B
- Analyze → t tests (and non parametric tests) → cocher les deux colonnes à comparer → ok
- Analyze this table again...
- Cocher Paired, Parametric test, Paired t test → ok
- P-value 0.0011 également

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Tests statistiques → t de Student pour données appariées
- Rentrer le nombre d'observations dans chaque échantillon (n=215)
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- P-value de 0.0011 également

Sous STATA:

P-value identique à celles de GraphPad et BiostaTGV

Comparer mes résultats

✧ Comparaison d'une variable continue entre 2 groupes appariés (n_1 , n_2)

Exemple: Comparaison du BMI avant et après traitement

$n_1 < 30$ ou $n_2 < 30$ et distribution non normale de la différence: [Test de Wilcoxon](#)

Sous GraphPad Prism:

- Analyze → t tests (and nonparametric tests) → cocher les deux colonnes à comparer → ok
- Cocher Paired, Use nonparametric test → ok
- P-value 0.0391

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Tests statistiques → Tests des rangs signés de Wilcoxon
- Rentrer le nombre d'observations: 215
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- P-value de 0.0391 également

Comparer mes résultats

✧ Comparaison d'une variable continue entre plus de 2 groupes

Exemple: Comparaison du BMI avant traitement en fonction des incompatibilités HLA

Loi normale et variances égales: **ANOVA**

Sous GraphPad Prism:

- Dans la base Excel, trier les incompatibilités HLA par ordre croissant
- Analyse précédente dans GraphPad, copier-coller le BMI pour chaque groupe d'incompatibilité HLA
- Vérifier la normalité
- Analyze → One-way ANOVA (and nonparametric) → cocher les colonnes à comparer → ok
- Cocher No matching or pairing, Use ANOVA → ok
- P-value 0.0997

Comparer mes résultats

✧ Comparaison d'une variable continue entre plus de 2 groupes

Exemple: Comparaison du BMI avant traitement en fonction des incompatibilités HLA

Loi normale et variances égales: **ANOVA**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

100% 100%

Σ

$\sigma = \frac{1}{n} \sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$

BiosciatuTV

Tests statistiques en ligne

100% 100%

Σ

$\sigma = \frac{1}{n} \sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$

Tableau des tests statistiques d'hypothèse

Le test d'hypothèse est une démarche statistique afin d'évaluer si véritable une hypothèse statistique en fonction d'un échantillon de données peut d'intégration des résultats observés sous forme compréhensible.

Il est en fait une affaire de probabilité, de risque de se tromper et d'analyser ce point sous une démarche mise en perspective la répartition de certains tests sous une logique de distribution.

Voici nos analyses avec l'ensemble d'analyse d'analyse de tests statistiques sous hypothèse :

- **Minimum Test d'hypothèse**
- **Tests statistiques**
- **Les échantillons comme sous-jacent au sujet d'analyse**

❓ Besoin d'aide pour choisir votre test ?

Calculer les statistiques de base d'une série de données

Analyse de survie ➔ **Nouveauté !**

Type de test et de données		Quantitative (symétrique (3 groupes))		Quantitative (asymétrique (plus de 3 groupes))		Qualitative (asymétrique)		Quantitative
Type de test		Type de données		Type de données		Type de données		Type de données
Paramétrique	Indépendants	F (paramétrique) symétrique	ANOVA (3-2)	ANOVA (plus de 3-2)	ANOVA (plus de 3-2)	ANOVA (plus de 3-2)	ANOVA (plus de 3-2)	Tout test de t-Student T-test de t-Student T-test de t-Student
	Appariés	T-test (plus de 3-2)	T-test (plus de 3-2)	T-test (plus de 3-2)	T-test (plus de 3-2)	T-test (plus de 3-2)	T-test (plus de 3-2)	T-test de t-Student T-test de t-Student T-test de t-Student
	Indépendants	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	T-test de t-Student T-test de t-Student T-test de t-Student
	Appariés	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	T-test de t-Student T-test de t-Student T-test de t-Student
Non Paramétrique	Indépendants	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	T-test de t-Student T-test de t-Student T-test de t-Student
	Appariés	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	T-test de t-Student T-test de t-Student T-test de t-Student
Quantitatif		ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	ANOVA (3-2)	T-test de t-Student T-test de t-Student T-test de t-Student

❗ La lecture de la table d'analyse peut dépendre du logiciel.

- Rentrer le nombre d'observations par groupe: 6, 10, 25, 52, 69, 40, 13
 - Copier-coller le BMI pour chaque groupe d'incompatibilité HLA dans les fenêtres prévues et générer le collage
 - P-value de 0.0997 également
-
- L'avantage de BiostaTGV est qu'il vous teste l'homogénéité des variances. Ici $p\text{-value} < 0.05$ donc les variances ne sont pas égales, nous ne sommes pas sensé utiliser l'ANOVA

Comparer mes résultats

✧ Comparaison d'une variable continue entre plus de 2 groupes

Exemple: Comparaison du BMI avant traitement en fonction des incompatibilités HLA

Loi non normale ou variances inégales: **Kruskall-Wallis**

Sous GraphPad Prism:

- Copier-coller le BMI pour chaque groupe d'incompatibilité HLA
- Vérifier la normalité
- Analyze → One-way ANOVA (and nonparametric) → cocher les colonnes à comparer → ok
- Cocher No matching or pairing, Use non parametric → ok
- P-value 0.1009

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Rentrer le nombre d'observations par groupe: 7, 10, 25, 52, 69, 40, 13
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- P-value de 0.1009 également

Tableau des tests statistiques d'hypothèse

Le test d'hypothèse est une démarche permettant d'évaluer la validité d'une hypothèse statistique en fonction d'un échantillon de données dont l'interprétation des résultats nécessite une forme simplifiée. Ce test est basé sur l'analyse de données et permet de conclure sur la validité d'une hypothèse statistique en fonction d'un échantillon de données. Pour cela, il est nécessaire de définir une hypothèse nulle et une hypothèse alternative, puis de choisir un test approprié.

• Choisir le test d'hypothèse à effectuer

• Choisir le type de données (qualitatives ou quantitatives)

• Choisir le type de test (paramétrique ou non paramétrique)

• Choisir le niveau de signification (alpha)

• Calculer les statistiques de base d'une série de données

Analyse de survie - nouveau !

Type de test	Statistique	Distribution	Test de normalité	Test de comparaison
Test de normalité	Test de normalité	Normal	Test de normalité	Test de comparaison
Test de comparaison	Test de comparaison	Normal	Test de normalité	Test de comparaison
Test de normalité	Test de normalité	Normal	Test de normalité	Test de comparaison
Test de comparaison	Test de comparaison	Normal	Test de normalité	Test de comparaison
Test de normalité	Test de normalité	Normal	Test de normalité	Test de comparaison
Test de comparaison	Test de comparaison	Normal	Test de normalité	Test de comparaison

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y est normalement distribuée selon X et la variance est uniforme: **Coefficient de Pearson**

Sous Excel:

- Dans une colonne à part, tapez F.TEST. Ici < 0.05 , donc les variances sont inégales
- Dans une autre colonne tapez COEFFICIENT.CORRELATION, matrice 1 correspond à l'âge du receveur et matrice 2 à l'âge du donneur → ok
- Coefficient $r = 0.3936$

 Ce test ne donne pas la p-value!

Interprétation des coefficients:

Correlation Coefficient
Interpretation

Coefficient Range	Strength of Relationship
0.00 - 0.20	Very Low
0.20 - 0.40	Low
0.40 - 0.60	Moderate
0.60 - 0.80	High Moderate
0.80 - 0.90	High
0.90 - 1.00	Very High

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y est normalement distribuée selon X et la variance est uniforme: **Coefficient de Pearson**

Sous GraphPad Prism:

- Entrer Age_Receveur dans Group A et Age_Donneur dans Group B → Analyze → correlation
→ cocher les deux colonnes à comparer → Compute r for each data set vs a control data set
→ Compute Pearson correlation coefficients → option Two-tailed, confidence interval 95%
→ ok
- Coefficient $r = 0.3936$ (identique à excel), $p\text{-value} < 0.0001$

Comparer mes résultats

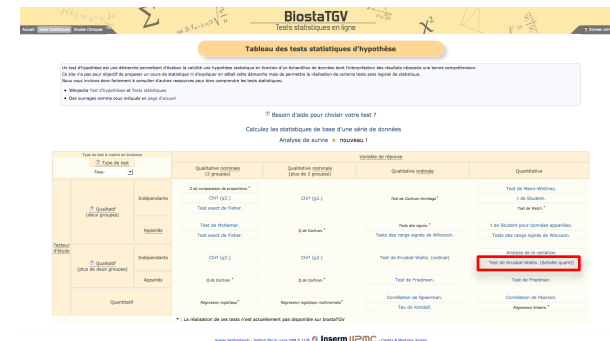
✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y est normalement distribuée selon X et la variance est uniforme: **Coefficient de Pearson**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Rentrer le nombre d'observations: 216
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- Coefficient $r = 0.3936$ également, $p\text{-value} < 0.0001$ également



Résultats du test

- Données série 1: 215L x 2C
- Méthode : Pearson's product-moment correlation; Alternative :two.sided
- Statistique observée Qobs : 6.2496024901807
- p-value : 2.2103062820728E-9
- **p : 0.3936** Intervalle de confiance à 95%[0.2743 ; 0.5011]
- Degrés de liberté : 213

La valeur p (p-value) de votre test est 2.2103062820728E-9.

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y est normalement distribuée selon X et la variance est uniforme: **Coefficient de Pearson**

Sous STATA:

```
. pwcorr Age_Receveur Age_Donneur, sig star(0.05)
```

	Age_Re~r	Age_Do~r
Age_Receveur	1.0000	
Age_Donneur	0.3936* 0.0000	1.0000

L'étoile indique que le test est significatif
Même coefficient r

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y n'est pas normalement distribuée selon X ou la variance n'est pas uniforme: **Coefficient de Spearman**

Sous Excel:

- Dans une colonne à part, tapez F.TEST. Ici < 0.05 , donc les variances sont inégales
- Dans une autre colonne tapez COEFFICIENT.CORRELATION, matrice 1 correspond à l'âge du receveur et matrice 2 à l'âge du donneur → ok
- Coefficient $r = 0.3936$

! Ce test ne donne pas la p-value!

Interprétation des coefficients:



Correlation Coefficient
Interpretation

Coefficient Range	Strength of Relationship
0.00 - 0.20	Very Low
0.20 - 0.40	Low
0.40 - 0.60	Moderate
0.60 - 0.80	High Moderate
0.80 - 0.90	High
0.90 - 1.00	Very High

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y n'est pas normalement distribuée selon X ou la variance n'est pas uniforme: **Coefficient de Spearman**

Sous GraphPad Prism:

- Entrer Age_Receveur dans Group A et Age_Donneur dans Group B → Analyze → correlation → cocher les deux colonnes à comparer → Compute r for each data set vs a control data set → Compute nonparametric Spearman correlation → option Two-tailed, confidence interval 95% → ok
- Coefficient $r = 0.372$, $p\text{-value} < 0.0001$

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Exemple: Comparaison de l'âge du receveur avec l'âge du donneur

Y n'est pas normalement distribuée selon X ou la variance n'est pas uniforme: **Coefficient de Spearman**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Rentrer le nombre d'observations: 216
- Copier-coller les valeurs de chaque groupe dans la fenêtre prévue et générer le collage
- Coefficient $r = 372$ également, p-value < 0.0001 également

BiostaTGV
Tests statistiques en ligne

Tableau des tests statistiques d'hypothèse

Le test d'ajustement est un test statistique permettant d'évaluer la validité d'un modèle statistique en fonction d'un échantillon de données. Les résultats de ce test sont présentés sous forme de tableau.

Calculer les statistiques de base d'une série de données

Analyse de survie

Test	Qualitative catégorielle (2 groupes)	Qualitative catégorielle (plus de 2 groupes)	Qualitative catégorielle (2 groupes)	Qualitative catégorielle (plus de 2 groupes)	Quantitative
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de Mann-Whitney
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared
Test de chi-squared (2x2)	Chi-squared (2x2)	Chi-squared (2x2)	Test de chi-squared	Test de chi-squared	Test de chi-squared

Résultats du test

- Données série 1: 215L x 2C
- Méthode : Spearman's rank correlation rho; Alternative :two.sided
- Statistique observée Qobs : 1039503.1461774
- p-value : 1.7729647073284E-8
- $r_s : 0.3724$

La valeur p (p-value) de votre test est 1.7729647073284E-8.

Comparer mes résultats

✧ Comparaison de 2 variables continues (corrélation)

Sous STATA:

```
. spearman Age_Receveur Age_Donneur
```

```
Number of obs =      215
```

```
Spearman's rho =    0.3724
```

```
Test of Ho: Age_Receveur and Age_Donneur are independent
```

```
Prob > |t| =      0.0000
```

Même coefficient r , p -value < 0.0001

Comparer mes résultats

✧ **Comparaison d'une proportion entre 2 groupes indépendants**

Exemple: Comparaison des décès en fonction du sexe

Effectifs théoriques > 5: **Test du χ^2**

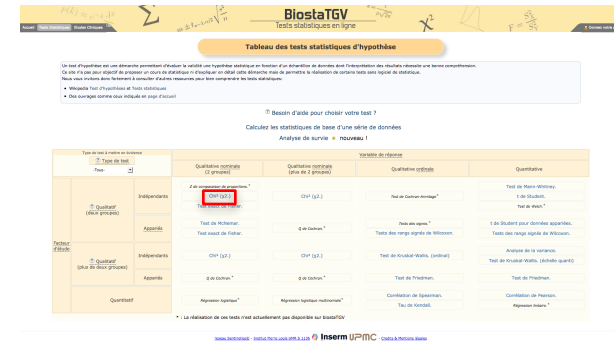
Comparer mes résultats

✧ Comparaison d'une proportion entre 2 groupes indépendants

Exemple: Comparaison des décès en fonction du sexe

Effectifs théoriques > 5: **Test du χ^2**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):



- Sous excel: Données → Trier... → trier par colonne D puis trier par colonne J
- Sous bioStaT: Rentrer le nombre de modalités pour la variable sexe: 2 (F et H) et pour la variable Type de donneur: 2
- Copier-coller le nombre d'occurrence pour chaque type de donneur et pour chaque sexe
- p-value = 0.012

	SME	VIV
	Tableau de contingence 1	
	Y modalité 1	Y modalité 2
F	X modalité 1: 75	X modalité 1: 11
H	X modalité 2: 94	X modalité 2: 35

Résultats du test

- Méthode : Pearson's Chi-squared test
- Statistique observée Qobs : 6.3102435468656
- p-value : **0.012004234844408**
- Paramètre du test : 1
- Tableau des effectifs attendus sous H0 :

La valeur p (p-value) de votre test est 0.012004234844408.

Comparer mes résultats

✧ Comparaison d'une proportion entre 2 groupes indépendants

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques > 5: **Test du χ^2**

Sous STATA:

Sexe_Recev eur	type_donneur		Total
	SME	VIV	
F	75	11	86
	67.6	18.4	86.0
	44.38	23.91	40.00
M	94	35	129
	101.4	27.6	129.0
	55.62	76.09	60.00
Total	169	46	215
	169.0	46.0	215.0
	100.00	100.00	100.00

Pearson chi2(1) = 6.3102 **Pr = 0.012**

Même p-value

Comparer mes résultats

✧ **Comparaison d'une proportion entre 2 groupes indépendants**

Exemple: Comparaison des décès en fonction du sexe

Effectifs théoriques < 5 : **Test exact de Fisher**

Comparer mes résultats

✧ Comparaison d'une proportion entre 2 groupes indépendants

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques < 5: **Test exact de Fisher**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Sous excel: Données → Trier... → trier par colonne D puis trier par colonne J
- Sous bioStaT: Rentrer le nombre de modalités pour la variable sexe: 2 (F et H) et pour la variable Type de donneur: 2
- Copier-coller le nombre d'occurrence pour chaque type de donneur et pour chaque sexe
- p-value = 0.238

SME VIV

Tableau de contingence 1

		Y 1	Y 2	
F H	X 1	74	11	+
	X 2	94	35	-
		+	-	

Résultats du test

- Données série 1: 2L x 2C
- Méthode : Fisher's Exact Test for Count Data; Alternative :two.sided
- p-value : **0.016816592547431**
- Odds Ratio : 2.4947 Intervalle de confiance à 95%[1.1434 ; 5.8298]

La valeur p (p-value) de votre test est 0.016816592547431.

Comparer mes résultats

✧ Comparaison d'une proportion entre 2 groupes indépendants

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques < 5: **Test exact de Fisher**

Sous STATA:

Sexe_Receveur	type_donneur		Total
	SME	VIV	
F	75	11	86
	67.6	18.4	86.0
	44.38	23.91	40.00
M	94	35	129
	101.4	27.6	129.0
	55.62	76.09	60.00
Total	169	46	215
	169.0	46.0	215.0
	100.00	100.00	100.00

Fisher's exact =
1-sided Fisher's exact =

0.017
0.008

Même p-value

Comparer mes résultats

✧ Comparaison d'une proportion entre 2 groupes appariés

Exemple: Comparaison du nombre de récidence de maladie après transplantation

Contrôles: non Tx

Cas: Tx

Nombre de malade (maladie initiale) dans chaque groupe

Test de McNemar: graphpad QuickCalc

[Explications](#)

[Calculator](#)

Comparer mes résultats

✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques > 5 sur 80%: **Test du χ^2**

Comparer mes résultats

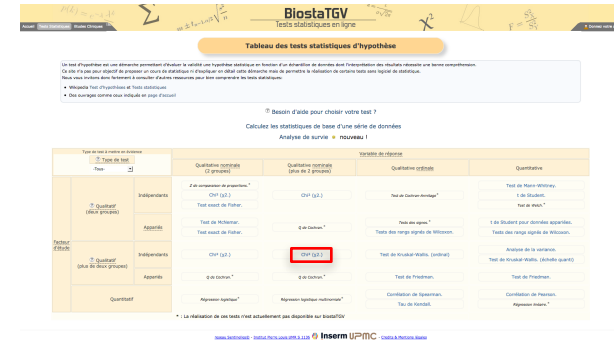
✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques > 5 sur 80%: **Test du χ^2**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):

- Sous excel: Données → Trier... → trier par colonne I puis trier par colonne G
- Sous bioStaT: Rentrer le nombre de modalités pour la variable sexe: 2 (F et H) et pour la variable incompatibilités HLA: 7
- Copier-coller le nombre d'occurrence pour chaque groupe d'incompatibilité et pour chaque sexe
- p-value = 0.238



Résultats du test

- Méthode : Pearson's Chi-squared test
- Statistique observée Qobs : 8.0063159700617
- p-value : **0.23764094666875**
- Paramètre du test : 6
- Tableau des effectifs attendus sous H0 :

Effectif théorique < 5

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
X1	2.411	4.019	10.047	20.897	27.729	16.075	4.822
X2	3.589	5.981	14.953	31.103	41.271	23.925	7.178

Certaines cases des effectifs attendus sont inférieures à 5. Les conditions de validité du Chi2 ne sont pas remplies

La valeur p (p-value) de votre test est 0.23764094666875. Toutefois, étant donné que les conditions de validité du test ne sont pas remplies (voir ci-dessus), nous vous conseillons de vous référer au tableau de classification des tests pour choisir un test plus approprié.

Comparer mes résultats

✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques > 5 sur 80%: **Test du χ^2**

Sous STATA:

Sexe_Receveur	INCOMPATIBILITE_HLA							Total
	0	1	2	3	4	5	6	
F	1	5	10	22	31	10	7	86
	2.4	4.0	10.0	20.9	27.7	16.1	4.8	86.0
	16.67	50.00	40.00	42.31	44.93	25.00	58.33	40.19
M	5	5	15	30	38	30	5	128
	3.6	6.0	15.0	31.1	41.3	23.9	7.2	128.0
	83.33	50.00	60.00	57.69	55.07	75.00	41.67	59.81
Total	6	10	25	52	69	40	12	214
	6.0	10.0	25.0	52.0	69.0	40.0	12.0	214.0
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Pearson chi2(6) = 8.0063 **Pr = 0.238**

Même p-value

Comparer mes résultats

✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques < 5 : **Test de Fisher**

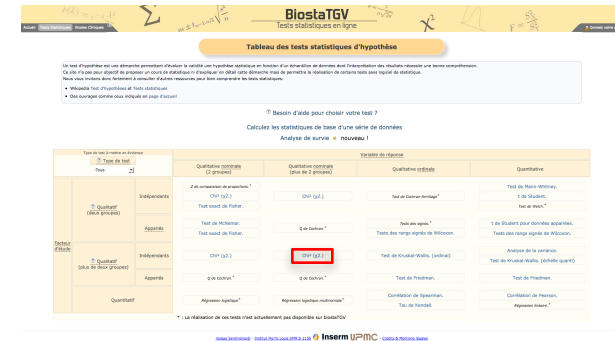
Comparer mes résultats

✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques < 5 : **Test de Fisher**

Sous BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/>):



- Sous excel: Données → Trier... → trier par colonne I puis trier par colonne G
- Sous bioStaT: Rentrer le nombre de modalités pour la variable sexe: 2 (F et H) et pour la variable incompatibilités HLA: 7
- Copier-coller le nombre d'occurrence pour chaque groupe d'incompatibilité et pour chaque sexe
- p-value = 0.234

Résultats du test

- Données série 1: 2L x 7C
- Méthode : Fisher's Exact Test for Count Data; Alternative :two.sided
- p-value : 0.23432239690729

La valeur p (p-value) de votre test est 0.23432239690729.

Comparer mes résultats

✧ Comparaison d'une proportion entre plusieurs groupes

Exemple: Comparaison des incompatibilités HLA en fonction du sexe

Effectifs théoriques < 5: **Test de Fisher**

Sous STATA:

Sexe_Receveur	INCOMPATIBILITE_HLA							Total
	0	1	2	3	4	5	6	
F	1	5	10	22	31	10	7	86
	2.4	4.0	10.0	20.9	27.7	16.1	4.8	86.0
	16.67	50.00	40.00	42.31	44.93	25.00	58.33	40.19
M	5	5	15	30	38	30	5	128
	3.6	6.0	15.0	31.1	41.3	23.9	7.2	128.0
	83.33	50.00	60.00	57.69	55.07	75.00	41.67	59.81
Total	6	10	25	52	69	40	12	214
	6.0	10.0	25.0	52.0	69.0	40.0	12.0	214.0
	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Fisher's exact =

0.234

Même p-value

Comparer mes résultats

Situation	Conditions 1	Conditions 2	Test à utiliser
Comparaison d'une variable continue entre 2 groupes indépendants (n1, n2)	Loi normale ou n1>30 et n2>30	Egalité des variances	Test de Student
	Loi non normale et n1<30 ou n2<30	Non égalité des variances	Test de Student avec variances inégales Test de Mann-Whitney
Comparaison d'une variable continue entre 2 groupes appariés (n1, n2)	n1>30 et n2>30 ou Distribution normale de la différence Distribution non normale de la différence et n1<30 ou n2<30		Test de Student apparié Test de Wilcoxon
Comparaison d'une variable continue entre plus de 2 groupes	Loi normale et variances égales Loi non normale ou variances inégales	Groupes ordonnés Groupes non ordonnés Groupes ordonnés Groupes non ordonnés	Test de tendance de Cuzick Analyse de la variance (ANOVA) Analyse de la variance (ANOVA) Test de tendance de Cuzick Test de Kruskal-Wallis Test de Kruskal-Wallis
Comparaison de 2 variables continues (corrélation)	Y est normalement distribuée selon X et La variance est uniforme Y n'est pas normalement distribuée selon X ou La variance n'est pas uniforme		Coefficient de Pearson Coefficient de Spearman
Comparaison d'une proportion entre 2 groupes indépendants	Effectifs théoriques > 5 Effectifs théoriques < 5		Test du χ^2 Test exact de Fisher
Comparaison d'une proportion entre 2 groupes appariés	Si nombre de paires discordantes > 20 Si nombre de paires discordantes < 20		Test de McNemar Test de McNemar
Comparaison d'une proportion entre plusieurs groupes	Effectifs théoriques > 5 sur 80% Effectifs théoriques < 5 Groupes ordonnés		Test du χ^2 Test exact de Fisher Test de tendance

PROGRAMME

Vendredi 24 novembre

12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

Introduction à la régression logistique

- Régression logistique: modèle qui permet d'exprimer la relation entre une variable Y qualitative à deux classes (oui/non) et des variables X_j qui peuvent être quantitatives ou qualitatives
- Variable Y : variable dépendante ou « maladie »
- Variables X_j : variables indépendantes ou « expositions »
- ex: association entre cancer du poumon (maladie) et tabac (exposition)

Introduction à la régression logistique

Comment exprimer la relation entre Y et les variables indépendantes?

- Exemple de l'association entre:
 - la présence d'anticorps anti-VHC (variable dépendante qualitative à deux classes: présents/absents) et
 - l'âge (variable indépendante)

Odds-ratio

L'odds-ratio est la mesure de choix pour exprimer le risque relatif dans les études cas-témoins, c'est à dire l'augmentation (la diminution) du risque d'être malade pour les exposés comparés aux non exposés.

On obtient en faisant le ratio entre l'odds d'être exposé pour les cas divisé par l'odds d'être exposé pour les témoins.

Odds-ratio

Dans les études transversales, il est fréquent de faire une analyse dite « type cas-témoins »:

- Les « malades » de l'enquête sont considérés comme les cas, et les « non malades » comme témoins
- On compare les « odds » d'exposition entre cas et témoins
- Attention, l'odds-ratio ainsi obtenu est le plus souvent biaisé, notamment:
 - Surestimé si la maladie est fréquente dans la population
 - Ininterprétable si on peut garantir que l'exposition mesurée est antérieure au développement de la maladie

Test d'hypothèse

Sous l'hypothèse nulle (absence d'association):

$$\hat{\beta} = 0 \Leftrightarrow \log OR = 0 \Leftrightarrow OR = 1$$

$$\hat{\beta} \approx N(0, \text{var}(\hat{\beta}))$$

$$Z = \frac{\hat{\beta} - 0}{\sqrt{\text{var}(\hat{\beta})}} = \frac{\hat{\beta}}{s(\hat{\beta})} \approx N(0,1) \quad (\text{test de Wald})$$

$$I.C. \text{ à } 95\% \text{ de } OR : e^{\hat{\beta} \pm z_{1-\alpha/2} * s(\hat{\beta})}$$

Tableau de contingence

```
. tab Sexe_Receveur PERTE_GREFFON, exp chi2
```

Pourcentage en colonne: proportion d'exposés

Key
<i>frequency</i>
<i>expected frequency</i>

SexeReceveur	PERTE_GREFFON_Update		Total
	0	1	
F	674 670.2	121 124.8	795 795.0
M	1,012 1,015.8	193 189.2	1,205 1,205.0
Total	1,686 1,686.0	314 314.0	2,000 2,000.0

OR: Produit en croix
 $= 674 \times 193 / 121 \times 1012$
 $= 1.06231$

Valeur P: test du chi2

Pearson chi2(1) = 0.2296 Pr = 0.632

Régression logistique univariée

```
. xi:logistic PERTE_GREFFON i.Sexe_Receveur  
i.Sexe_Receveur    _ISexe_Rece_1-2      (_ISexe_Rece_1 for Sexe_Rec~r==F omitted)
```

```
Logistic regression              Number of obs      =      2,000  
                                LR chi2(1)          =      0.23  
                                Prob > chi2         =      0.6313  
Log likelihood = -869.20796      Pseudo R2       =      0.0001
```

PERTE_GREFFON	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_ISexe_Rece_2	1.06231	.1340265	0.48	0.632	.8295821	1.360327
_cons	.1795252	.017725	-17.39	0.000	.1479395	.2178547

Régression logistique multivariée

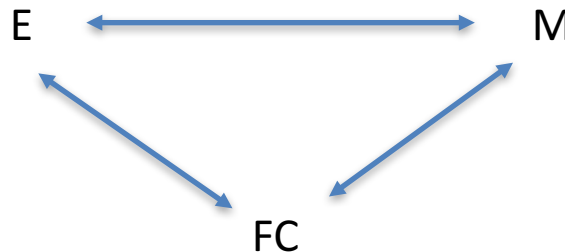
Choix des variables

- 1/ Choix d'une variable d'exposition privilégiée, autour de laquelle le modèle est construit, ou analyse de toutes les expositions sur le même plan
- 2/ Connaissance scientifiques préalables sur les facteurs de risque
- 3/ Résultats de l'analyse univariée (on inclut les variables avec une p-value $<0,25$ en univariée)
- 4/ Certaines variables seront forcées dans le modèle (facteurs de risque reconnus de la maladie).

Ajustement des biais de confusion

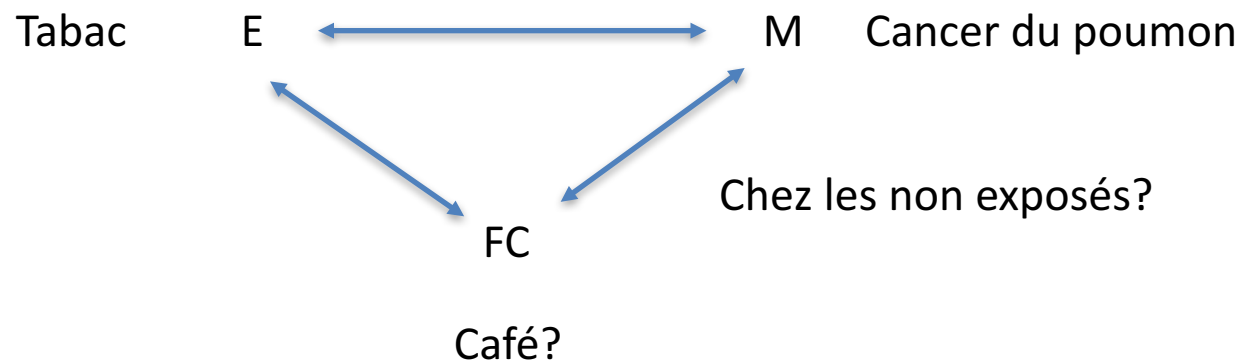
Un biais de confusion est dû à un 3^{ème} facteur, dit de confusion, associé simultanément à l'exposition et à la maladie

Par association, on entend une proportion différente du FC chez les exposés et les non-exposés, et une proportion différente du FC chez les malades et les non-malades



Ajustement des biais de confusion

Exemple: Etude sur le cancer du poumon. Association entre le tabac et le cancer mais aussi le café et le cancer



Ajustement des biais de confusion

Attention, le facteur de confusion ne doit pas être dans la chaîne causale entre exposition et maladies

Nourriture riche en graisses → hypercholestérolémie → Infarctus

Ici l'hypercholestérolémie n'est pas un facteur de confusion même si elle est liée à une nourriture riche en graisses, et à l'infarctus.

Variables pouvant déséquilibrer le modèle

- Les variables pour lesquelles les effectifs sont déséquilibrés après croisement avec la maladie
- Les variables corrélées (ayant la même information)
- Les variables avec des valeurs manquantes: observation éliminées du modèle → Perte de puissance ou biais de sélection

Interactions

- Il y a une interaction entre l'exposition et une autre variable si le risque de maladie associé à l'exposition varie selon les catégories de la variable d'interaction
- Exemple: le risque de mésothéliome pleural associé à l'amiante est élevé chez les fumeurs comparés aux non-fumeurs. Le tabac « modifie » l'effet de l'amiante sur le risque de mésothéliome.
- L'interaction n'est pas un biais mais une réalité et doit être rapportée en tant que telle, en décrivant l'effet de l'exposition selon les catégories de la variable d'interaction.
- A détecter en analyse bivariée
- A rapporter seulement si importantes, car alourdissent l'interprétation du modèle
- Préférer un modèle stratifié si nombreuses interactions

Détecter une interaction

- L'interaction, comme l'effet de confusion, se détecte en faisant une analyse stratifiée:
 - Si l'OR entre l'exposition et la maladie est différent selon les strates d'une troisième variable, il s'agit d'une interaction.
 - Si l'OR entre l'exposition et la maladie est le même selon les strates d'une troisième variable, mais différent de l'OR avant analyse stratifiée, il s'agit d'un effet de confusion
- La démarche d'analyse est donc la même (analyse stratifiée) pour détecter les deux effets (confusion et interaction). Il s'agit néanmoins de deux phénomènes différents appelant des réponses différentes:
 - L'interaction est rapportée telle quelle, avec des Ors par strates.
 - Le biais de confusion est corrigé par l'analyse, avec un OR ajusté.

Construction du modèle

- Automatique: pas-à-pas montant ou descendant (par le logiciel)
 - Avantages:
 - Exploration systématique de toutes les variables
 - Inconvénients:
 - Ne permet pas de prendre en compte des considérations épidémiologiques dans le choix es variables
 - Se fait piéger par les variables très corrélées entre elles
 - Ne prend en compte que les observation pour lesquelles toutes les variables du modèle de départ sont renseignées
 - Ne permet pas de modéliser les interactions
- Manuelle: l'investigateur teste lui-même les variables candidates dans le modèle:
 - Avantages:
 - Facile à réaliser avec les logiciels et meilleur contrôle de la construction du modèle
 - Inconvénients:
 - Fastidieux quand on a beaucoup de variables à tester

En pratique

- Commencer par une analyse univariée en visualisant le tableau de contingence et en faisant une estimation de l'OR par régression logistique univariée
- Si une variable exerce un effet de confusion très important sur l'ensemble des autres variables, tester avec un modèle logistique réduit la valeur P de chaque exposition après ajustement sur la variable confondante
- Tester les interactions qui vous paraissent épidémiologiquement pertinentes avec la variable d'exposition principale
- Commencer avec un modèle contenant les variables forcées et les variables trouvées associées à un $p < 0,2$ en analyse univariée. Tester l'effet du retrait des variables une à une sur l'ensemble du modèle
- Ne garder au final que les variables dont le degré de significativité est $< 0,05$, et les variables forcées

Introduction à l'analyse de survie

- *Descriptif*: “survie” d’une population
- *Comparatif*: survie entre un groupe traité et un groupe placebo, entre homme et femme, selon le niveau de calcémie ...
- *Prédictif*:
 - relation entre variables explicatives et survie
 - prédiction individuelle du risque

Introduction à l'analyse de survie

Exemple: comparaison incidence hépatite B à 1 an chez sujets vaccinés vs. placebo

Essai thérapeutique double aveugle

2 bras randomisés de 184 sujets chacun

Date de l'injection variable

Critère de jugement= nbre hépatite B 1 an après injection

Résultats: après 1 an de suivi, 6 hép B observées dans le groupe vacciné, mais 7 sujets n'ont pas eu une surveillance complète de 1 an (perdus de vue).

Introduction à l'analyse de survie

1^{ère} approche= comparaison des taux d'hépatite B dans les 2 groupes

Mais ne tient pas compte du DELAI de survenue des évts

Poser un taux à 6/184 dans le groupe vacciné suppose que les 7 sujets à surveillance incomplète sont restés indemnes d'hép B

Exclure ces 7 sujets: le taux est $6/(184-7)$ ne tient pas compte du fait que les 7 sujets ont été suivis une partie du temps et ont contribué au risque de développer une hép B pendant un certain temps (jusqu'à leur censure);
risque d'introduction de biais (sous estimation)

L'approche par la survie permet de comptabiliser les durées de suivi de l'**ensemble** des sujets (y compris ceux dont la durée de suivi est incomplète), permettant d'avoir un test plus puissant que la simple comparaison des %

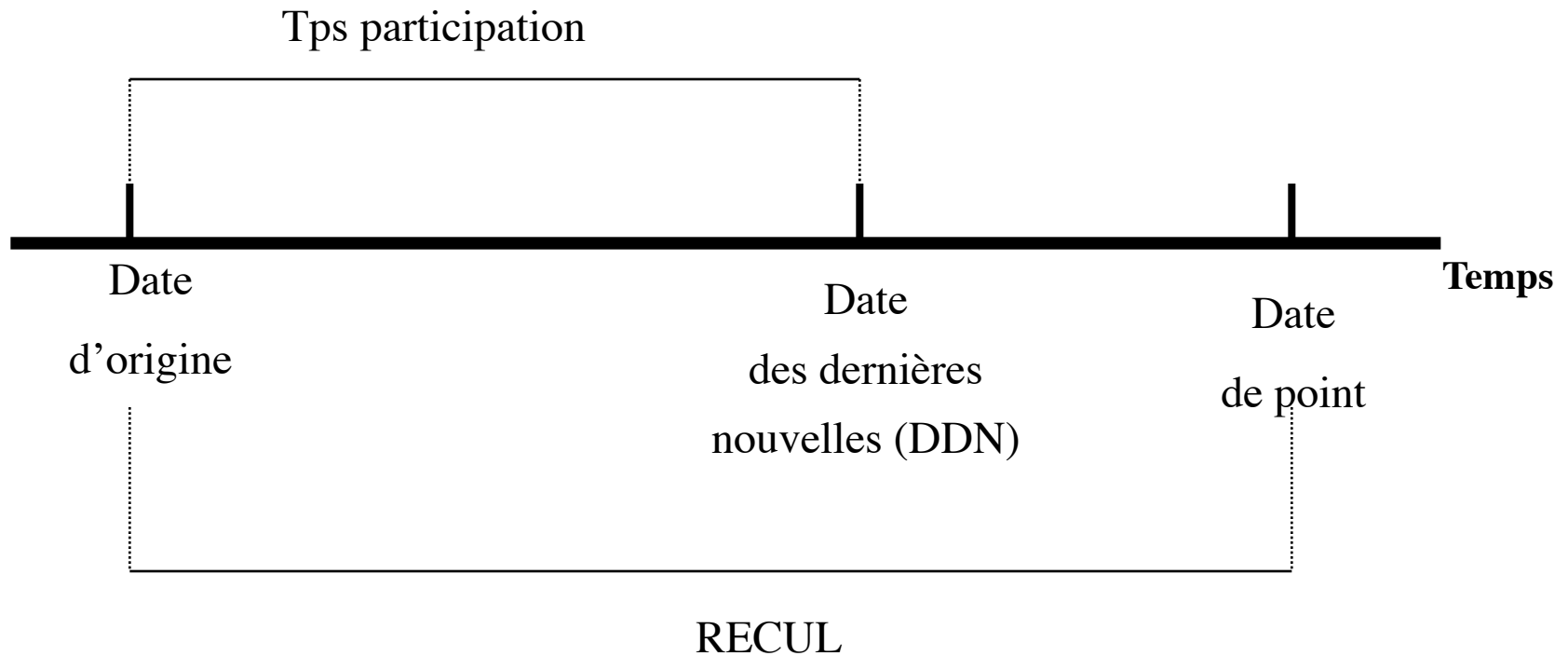
Introduction à l'analyse de survie

- DEFINITION

- Décrire un phénomène (biologique) dans lequel le TEMPS (vb aléatoire) intervient
- Survenue d'un évènement au cours du temps: tenir compte
 - la durée de surveillance **propre à chaque sujet** de l'étude
 - du **moment de survenue de l'évènement** pour chaque individu de l'étude

Introduction à l'analyse de survie

- VOCABULAIRE



Introduction à l'analyse de survie

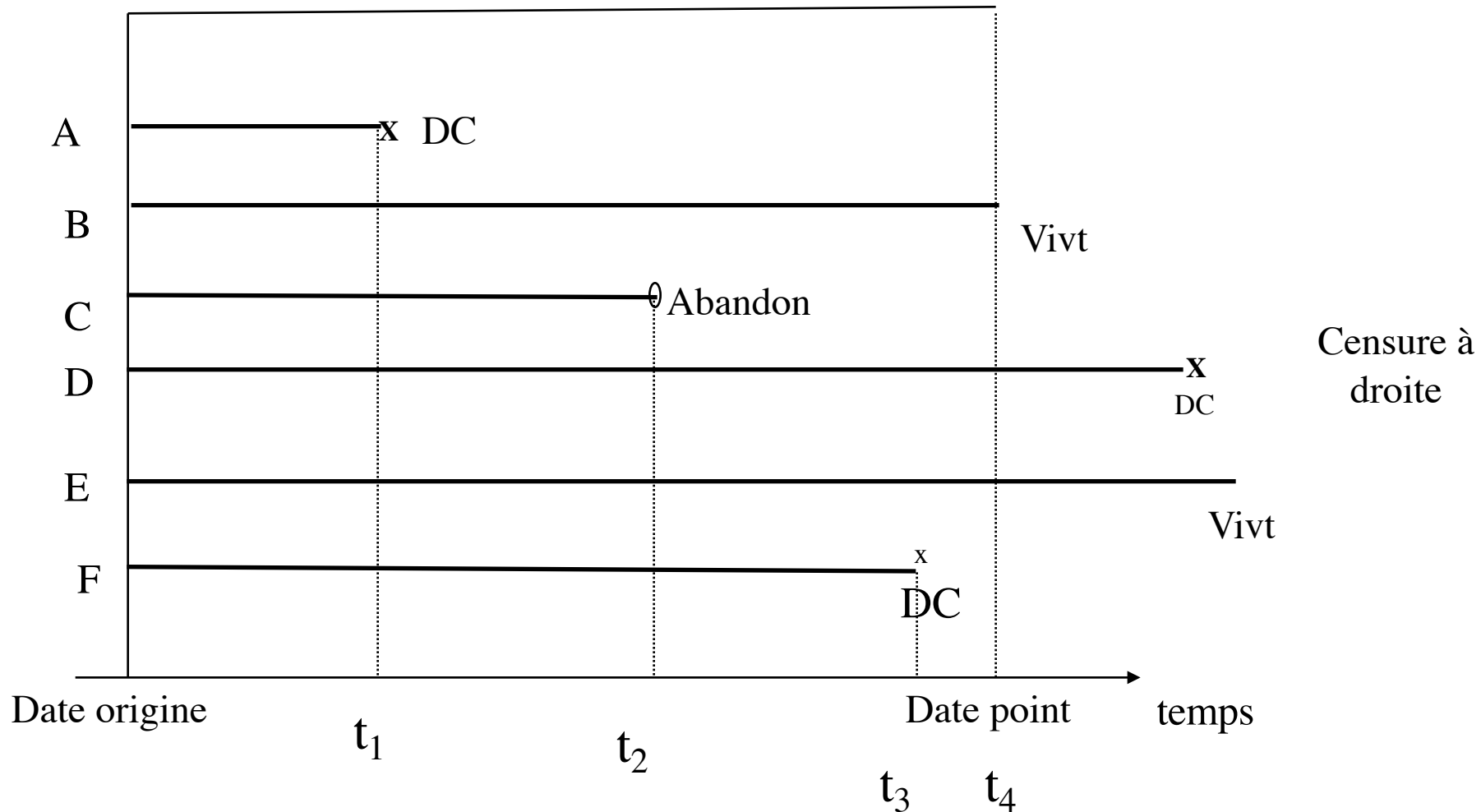
- Délai de censure à droite
- Type I et II: **fixées a priori par l'investigateur**
 - Type I: fin de l'étude après k mois de suivi
 - Type II: attente de k évts parmi n sujets pour arrêter l'étude
- **Type III**: aléatoire: survenue des censures de manière aléatoire au cours du suivi

Introduction à l'analyse de survie

- Type III
- Ex: étude de cohorte sur l'incidence de la dépression : des sujets en cours de suivi peuvent refuser de participer et sortir de l'étude
 - Censures **informatives** ? Sorties en lien avec l'exposition (difficultés psychologiques) ?

Introduction à l'analyse de survie

- Exemple

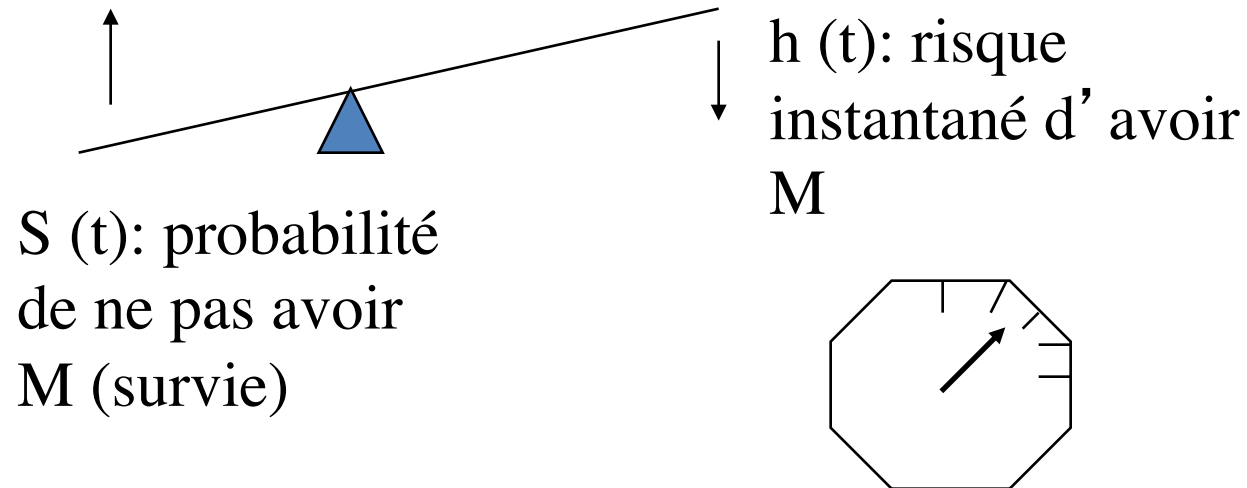


Introduction à l'analyse de survie

- Perdus de vue
- Limiter leur nb au maximum (cf procédures de surveillance)
- Conséquences possibles
 - Perte de puissance
 - Biais de sélection possible
- Décrire leurs caractéristiques initiales par rapport aux non perdus de vue, afin d'évaluer les conséquences possibles sur le sens et le niveau des associations observées

Introduction à l'analyse de survie

- Fonction de survie



Introduction à l'analyse de survie

- Fonction de survie
- Approche probabiliste: les évènements surviennent de manière aléatoire dans le temps.
- La variable aléatoire durée de “vie” (tps de suivi) peut être décrite par des fonctions de probabilité
- Il existe des liens entre ces fonctions

Méthode de Kaplan-Meier

Méthode actuarielle ou Life table method

Estimation de la fonction de survie $S(t)$

- Intérêt descriptif de la survie entre 2 ou plusieurs groupes
- Etape préliminaire avant la mise en route d'un modèle de régression de type Cox
- Permet de tester des hypothèses de validité du modèle du Cox
- Mais ne quantifie pas la relation entre E et M

Principes généraux

- L'estimation de la survie instantannée ne se fait qu'aux instants où surviennent au moins un évènement
- Les censures n'interviennent que dans le calcul du nb de sujets exposés au risque en début d'intervalle

Notation

Tps de décès	t_{i-1}	t_i
Nb de sujets exposés au risque	n_{i-1}	n_i
Nb de censures dans $[t_{i-1}; t_i[$	c_{i-1}	
Nb de décès	m_{i-1}	m_i

Notation

$n_i = n_{i-1} - m_{i-1} - c_{i-1}$ (nb sujets à risque en début d' [i])

$$S(t) = \prod (n_i - m_i) / n_i$$

En l'absence de censure avant t, $S(t)$ = % de sujets encore vivants en t

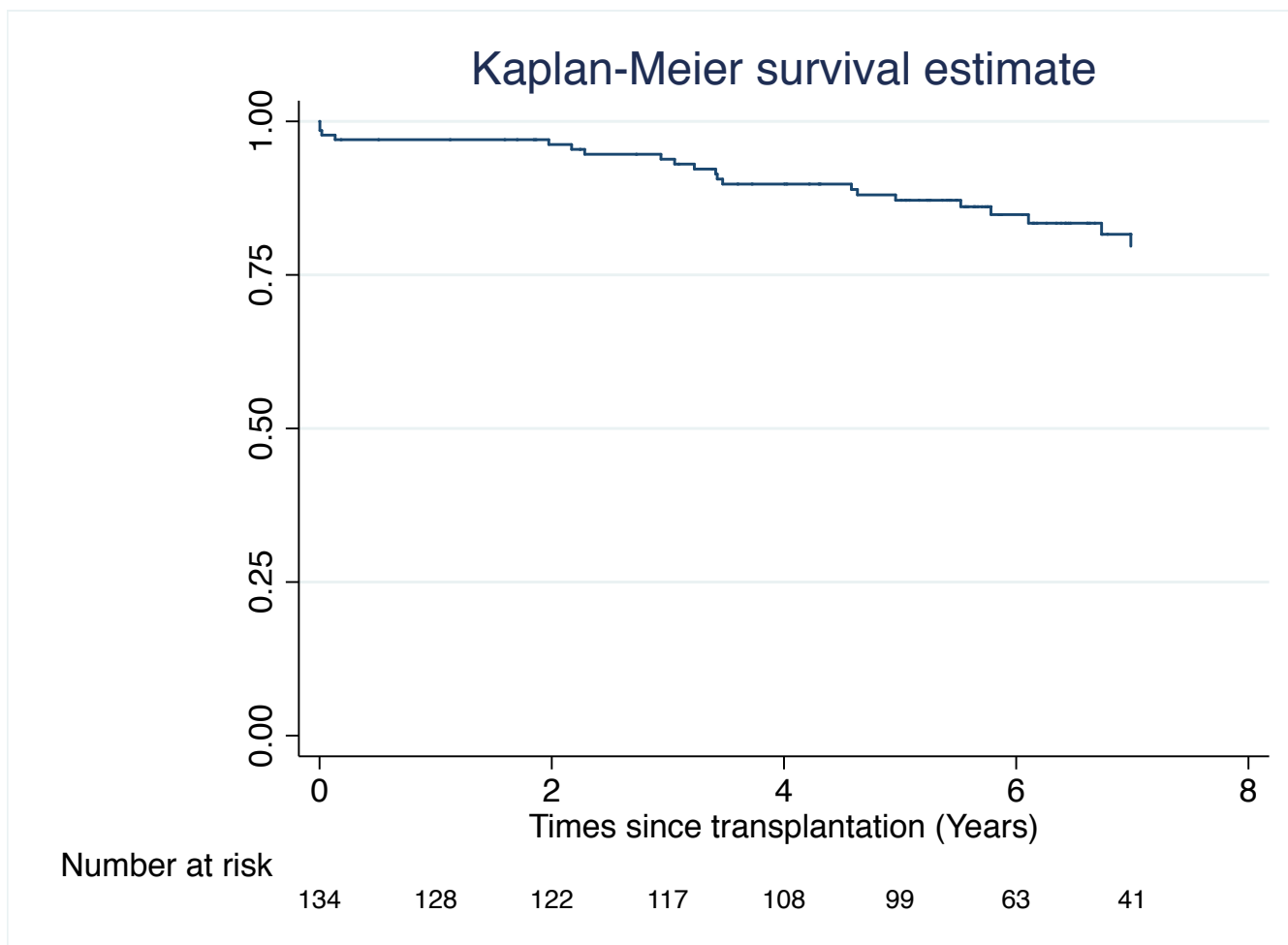
failure _d: **graft_lost_status_new == 1**
analysis time _t: **delai_en_a**

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
.003	2763	8	0	0.9971	0.0010	0.9942	0.9986
.0055	2755	0	1	0.9971	0.0010	0.9942	0.9986
.0137	2754	0	1	0.9971	0.0010	0.9942	0.9986
.0164	2753	1	0	0.9967	0.0011	0.9937	0.9983
.0219	2752	1	0	0.9964	0.0011	0.9933	0.9981
.0246	2751	1	0	0.9960	0.0012	0.9928	0.9978
.0274	2750	1	0	0.9957	0.0013	0.9924	0.9975
.0301	2749	1	0	0.9953	0.0013	0.9919	0.9973
.0385	2748	1	0	0.9949	0.0014	0.9915	0.9970
.044	2747	1	0	0.9946	0.0014	0.9910	0.9967
.0495	2746	1	0	0.9942	0.0014	0.9906	0.9964
.0522	2745	1	0	0.9938	0.0015	0.9901	0.9962
.0687	2744	1	0	0.9935	0.0015	0.9897	0.9959
.0714	2743	1	0	0.9931	0.0016	0.9892	0.9956
.0769	2742	0	1	0.9931	0.0016	0.9892	0.9956
.0797	2741	2	0	0.9924	0.0017	0.9884	0.9950
.0824	2739	2	0	0.9917	0.0017	0.9875	0.9945
.0931	2737	0	1	0.9917	0.0017	0.9875	0.9945
.0934	2736	1	0	0.9913	0.0018	0.9871	0.9942

Représentation graphique

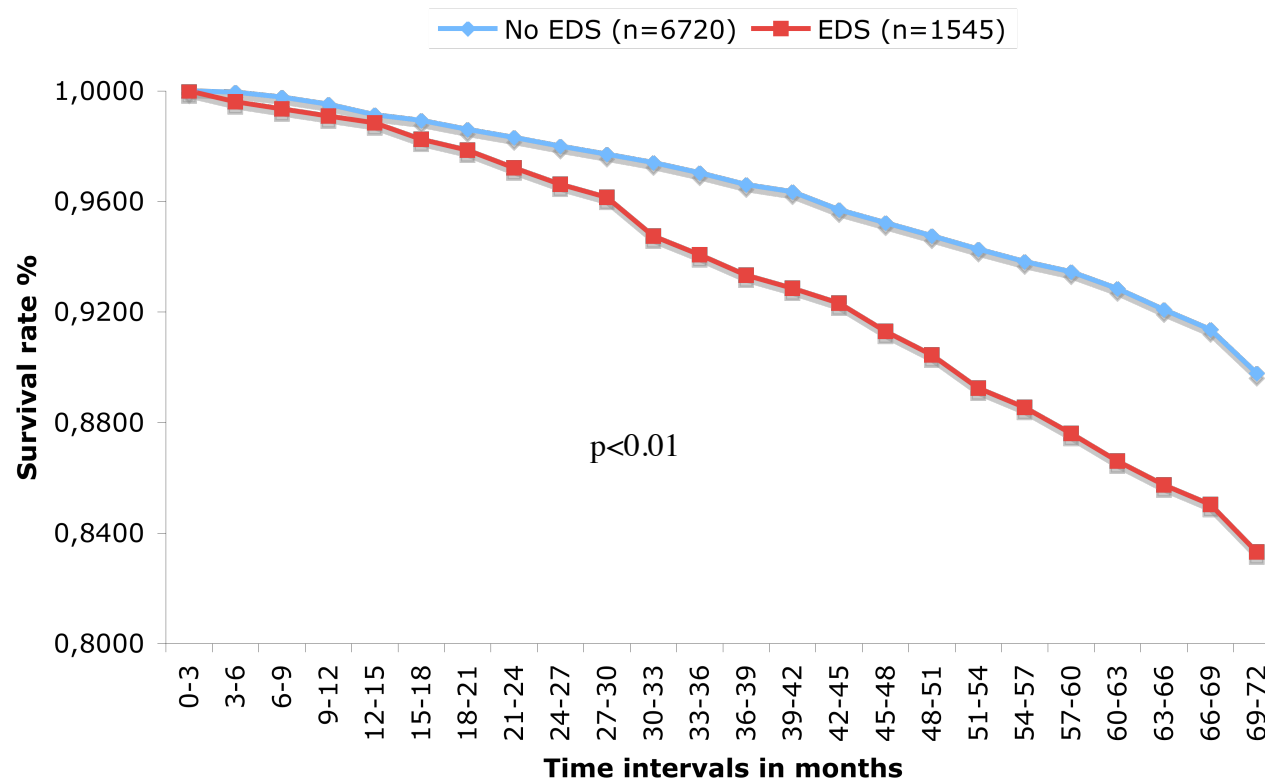
- Kaplan-Meier (product limit method): estimation de $S(t)$ à chaque instant où survient un évènement.
 - Adapté aux “petits échantillons” avec temps de survenue des evts précis
- Actuarielle: définition à priori des intervalles de temps d’observation des évènements et de l’estimation de $S(t)$.
 - Plus adapté aux grands échantillons ou à des délais de survenue d’évts imprécis

Représentation graphique

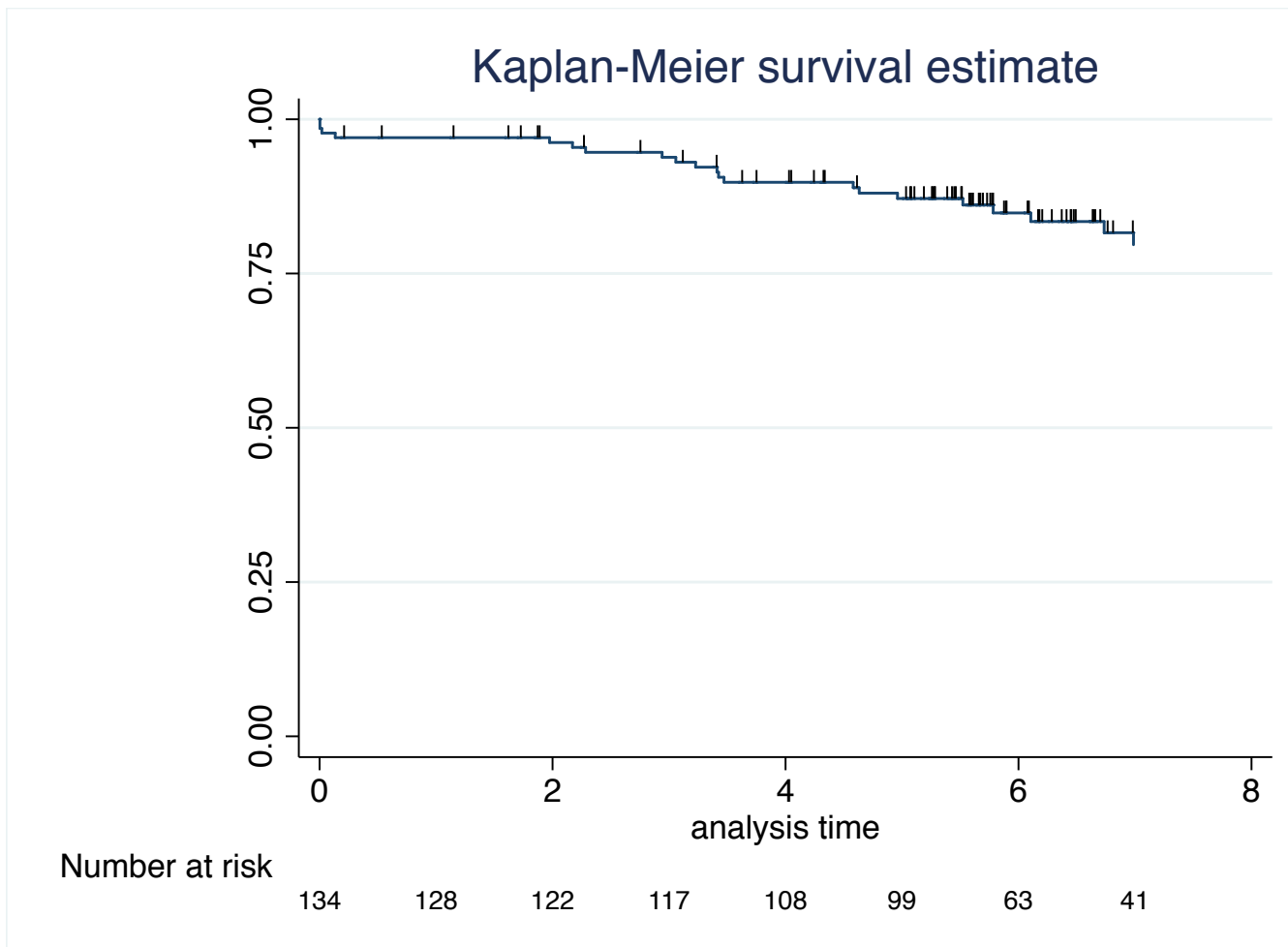


Représentation graphique

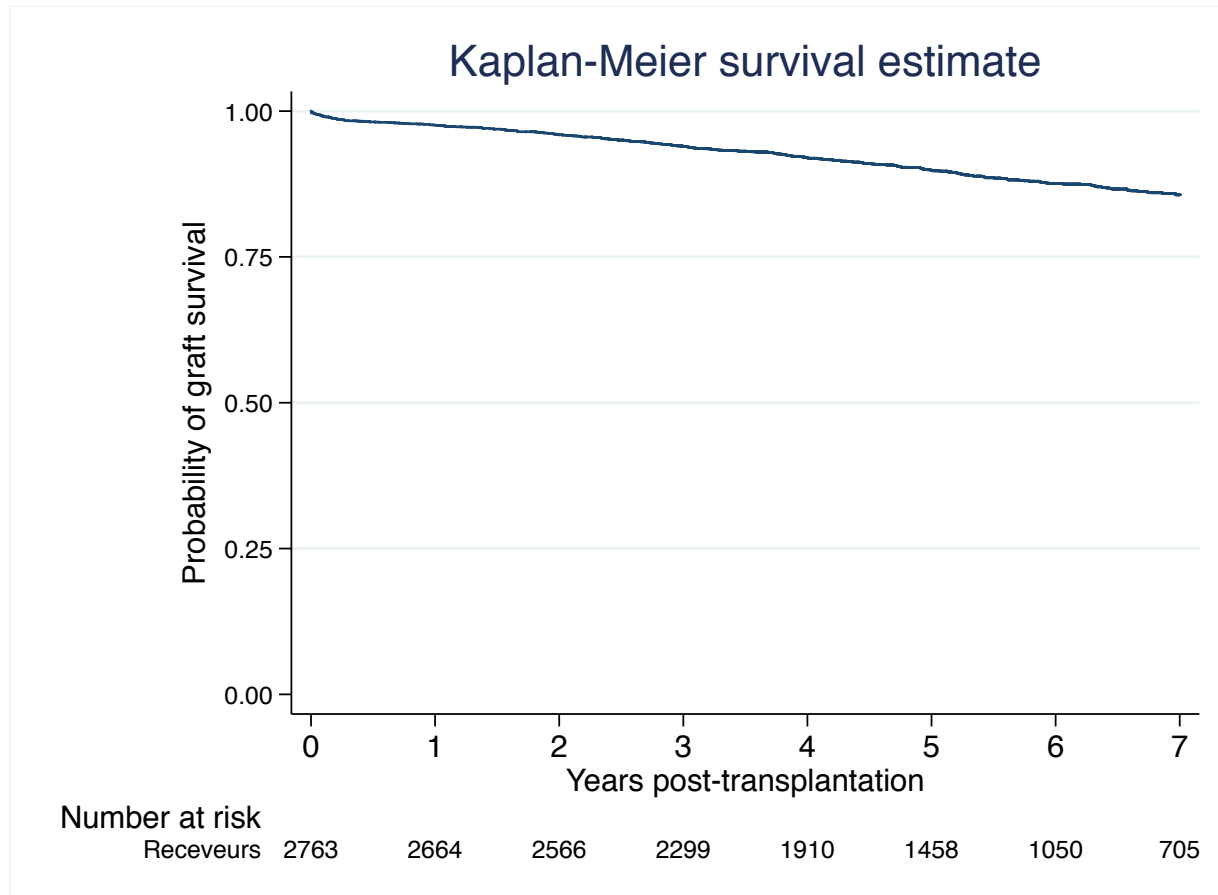
Life table analysis (méth. actuarielle) showing association between excessive daytime sleepiness (EDS) and mortality in the elderly. The 3C Study. *Empana et al. Stroke 2009;40:1219-24*



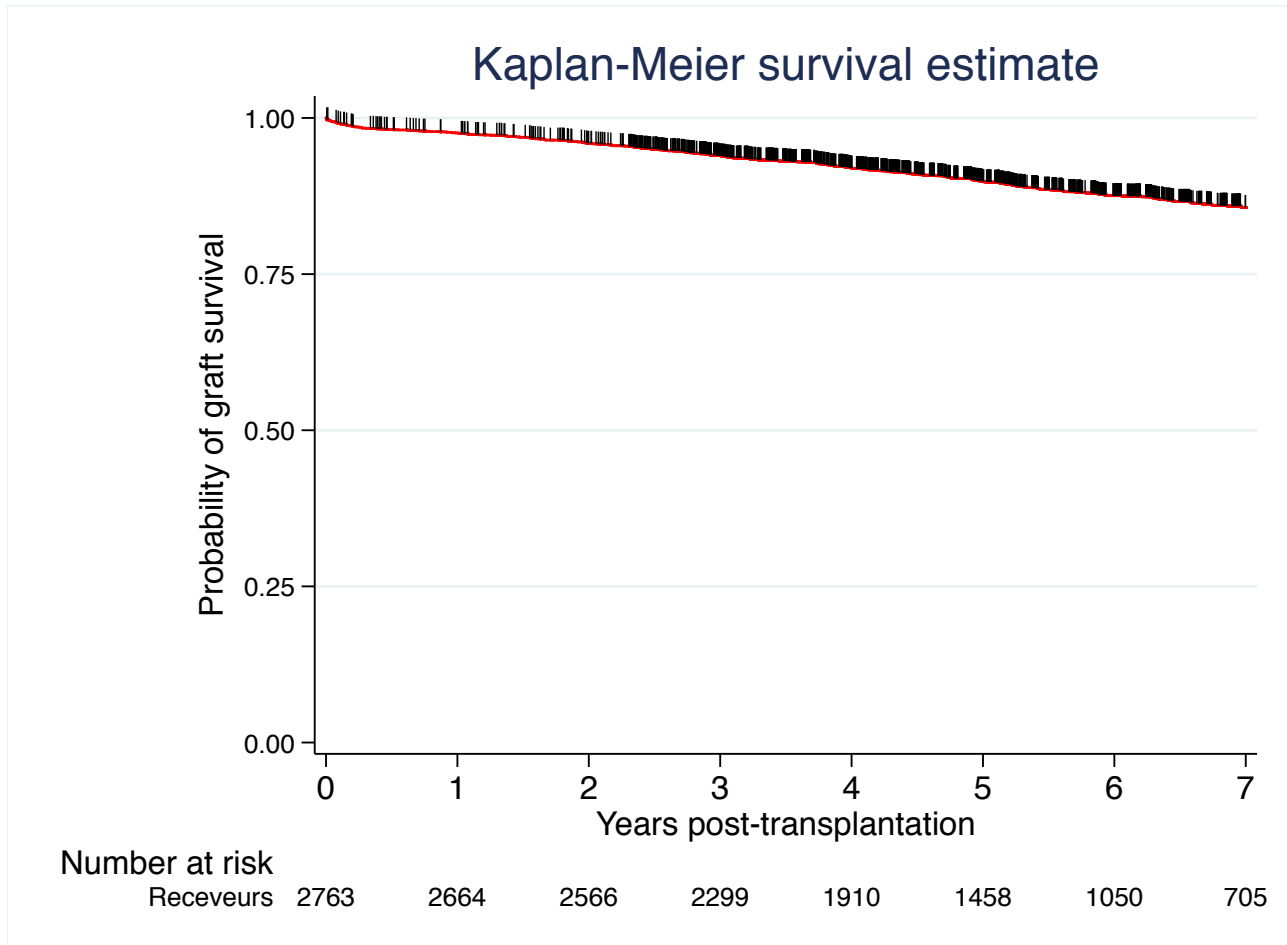
Représentation des censures



Représentation des censures → limite



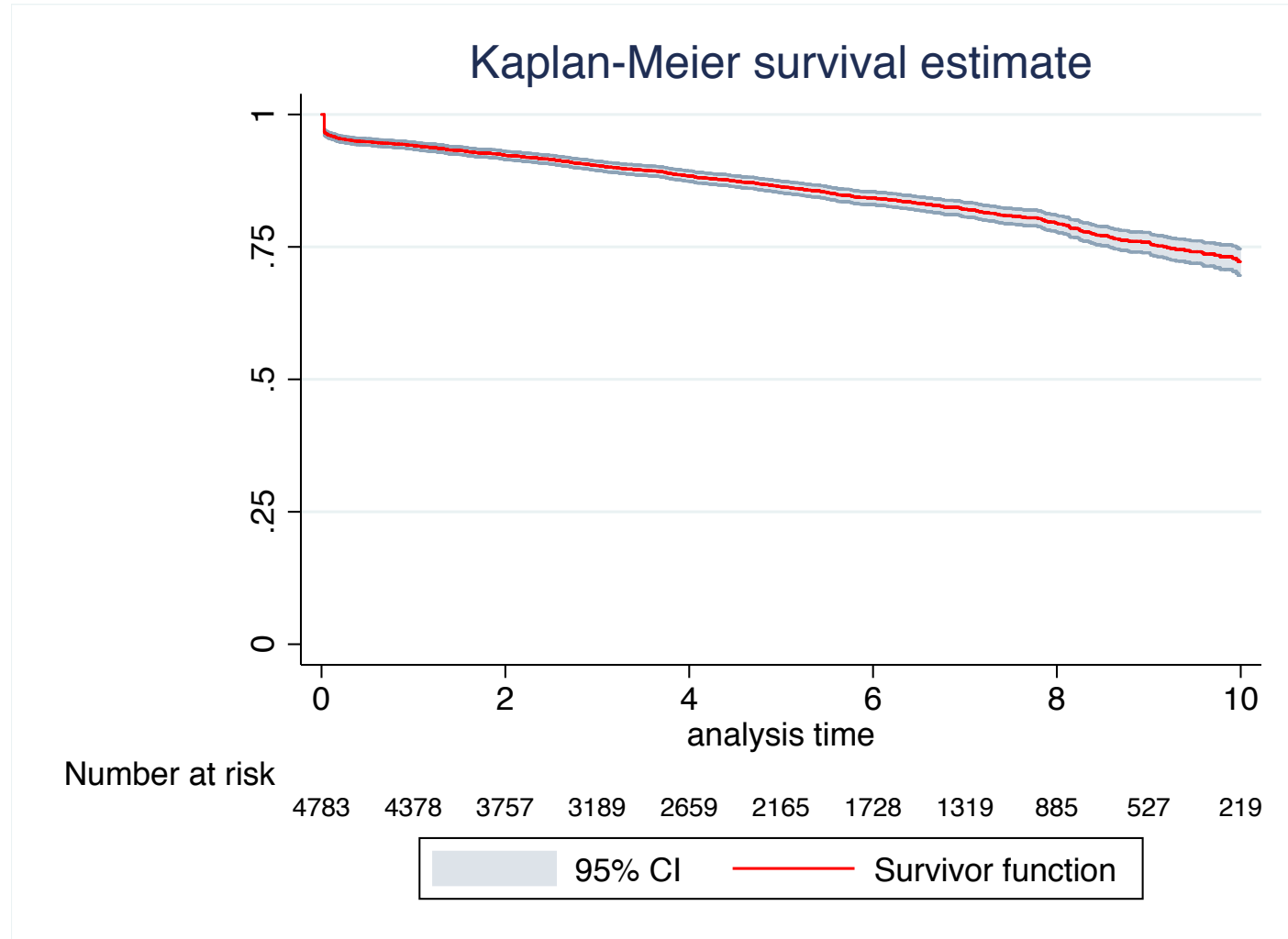
Représentation des censures → limite



Précision de l'estimation

- La précision de l'estimation d'une courbe de survie est représentée par un intervalle de confiance (le plus souvent à 95%)
- Cet intervalle prend la forme de deux courbes correspondant à la limite supérieure et à la limite inférieure
- Cet intervalle a une probabilité de 95% d'inclure la vraie courbe de survie

Précision de l'estimation : exemple



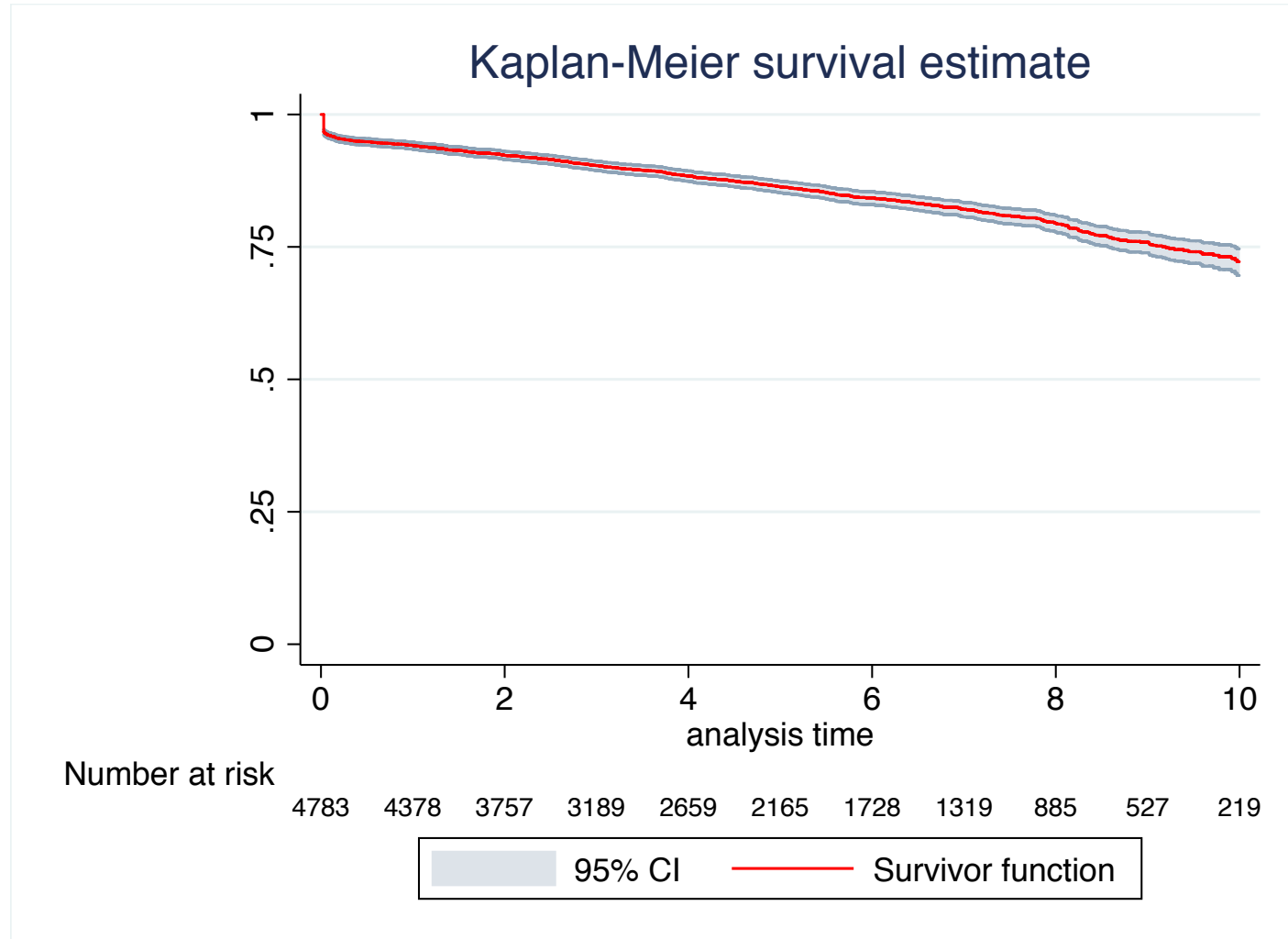
Intervalle de confiance des KM

- Dans quels cas la précision de l'estimation est réduite?
- Petit échantillon
- La survenue de censure

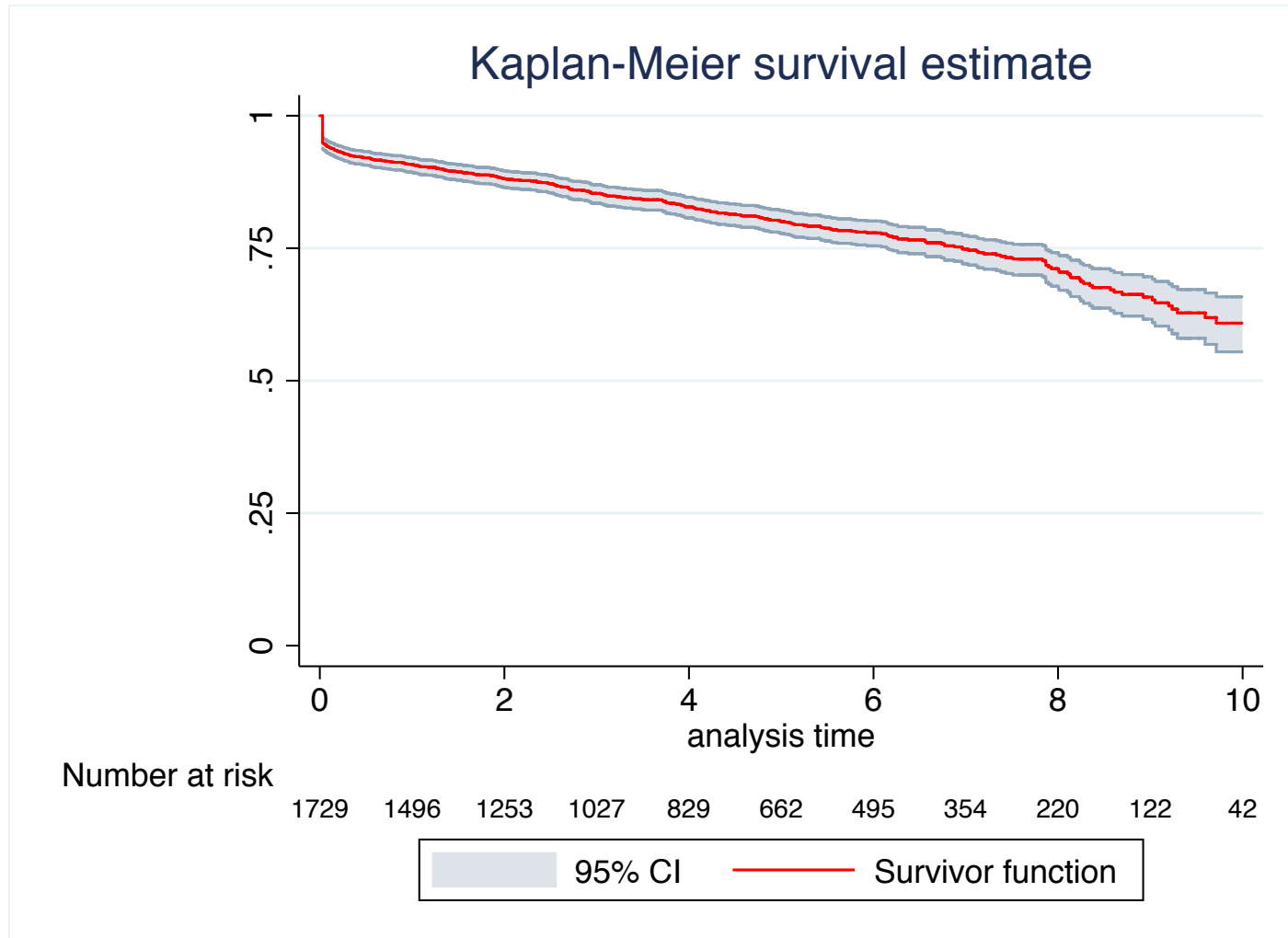
Rôle de la taille des échantillons

- Exemple de modification de la taille de l'échantillon
 - Données de survie de patients transplantés rénaux dans 4 centres français
 - KM représentant le survie des greffons rénaux depuis la transplantation

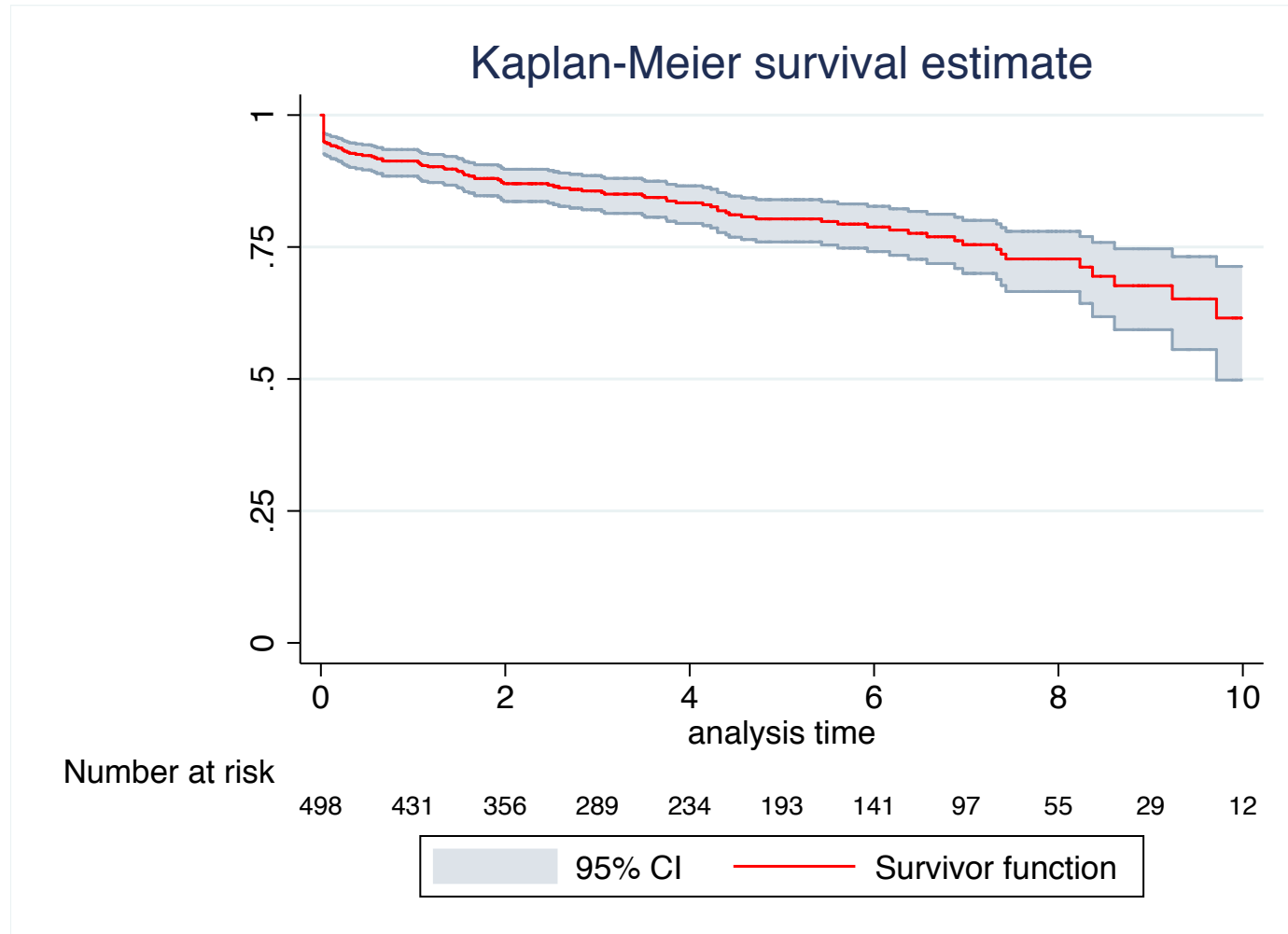
Rôle de la taille des échantillons : exemple



Rôle de la taille des échantillons : exemple



Rôle de la taille des échantillons : exemple



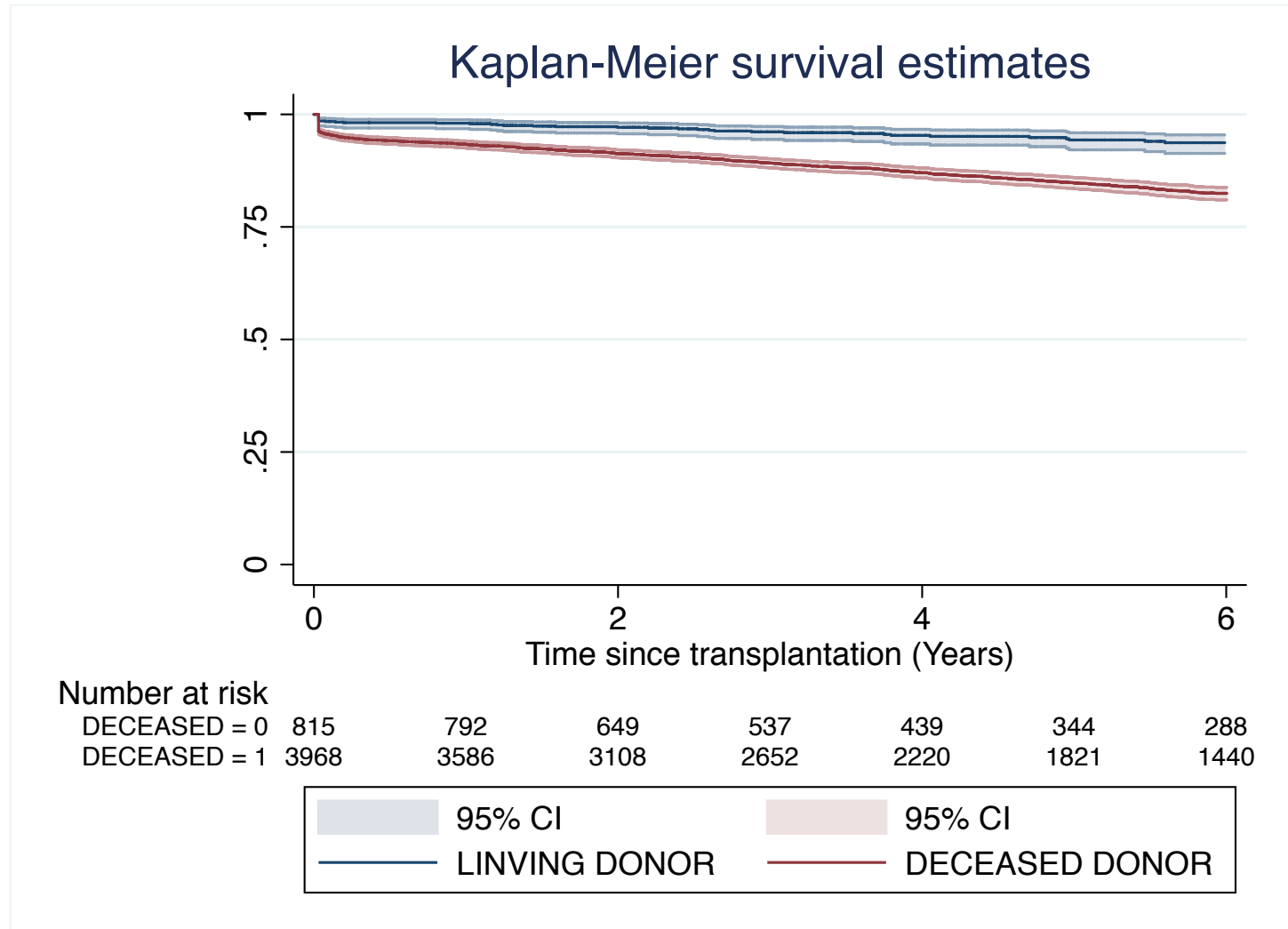
Rôle de la durée de suivi

- Attention à l'évolution du nombre de patients exposés au risque ainsi que le nombre d'évènements.
- Intérêt du nombre de sujets à risque.+++
- Permet d'apprécier la quantité d'information disponible à un temps donné, et ainsi apprécier la précision et la fiabilité de l'estimation de la survie à ce temps.

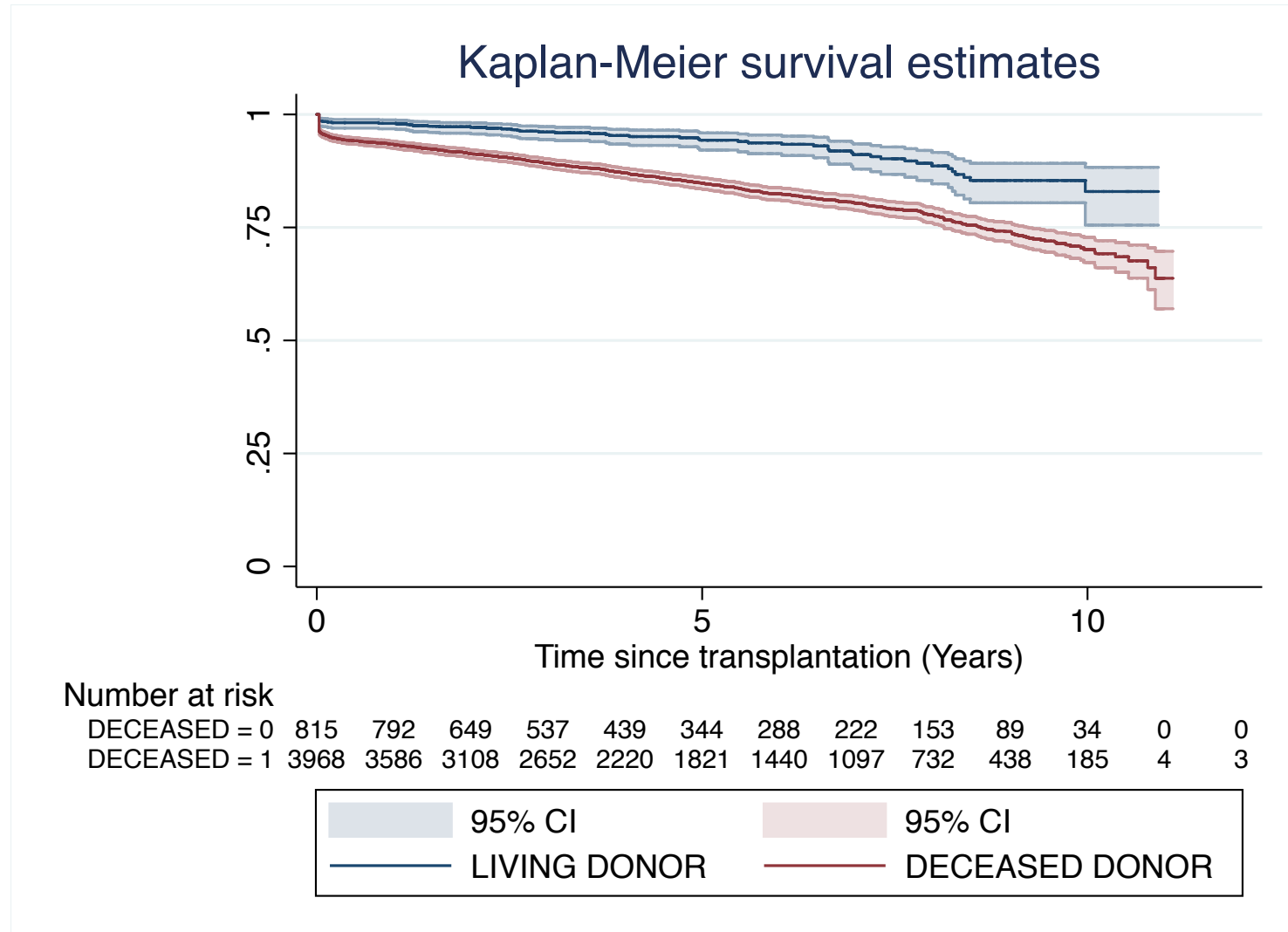
Rôle de la durée de suivi

- Données de survie de patients transplantés rénaux dans 4 centres de transplantation français
- KM représentant le survie des greffons rénaux depuis la transplantation
- 2 groupes:
 - Patients ayant reçus un rein issu de donneur vivant
 - Patients ayant reçus un rein issu de donneur décédé

Rôle de la durée de suivi : exemple de la censure



Rôle de la durée de suivi : exemple de la censure



Comparaison de la survie entre deux groupes

- Comparer la survie d'un groupe avec et sans tt actif; ou selon le genre (hommes vs. femmes) ...
- 2 séries de tables dans chacun des 2 groupes
- 2 courbes de survie
- Une série de tests statistiques évaluant la différence de survie entre les 2 groupes

Objectifs

- Hypothèse nulle H_0 : identité des distributions de survie dans les 2 groupes
- 2 approches
 - *Paramétrique*: distribution connue des fonctions de survie
 - *Non-paramétrique*: peu d'hypothèse sur la distribution des fonctions de survie

Log-rank test for equality of survivor functions

ECD_FINAL	Events observed	Events expected
0	187	234.61
1	146	98.39
Total	333	333.00

chi2(1) = 32.88

Pr>chi2 = 0.0000

. sts test ECD_FINAL, wilcoxon

failure _d: graft_lost_status_new == 1
analysis time _t: delai_en_a

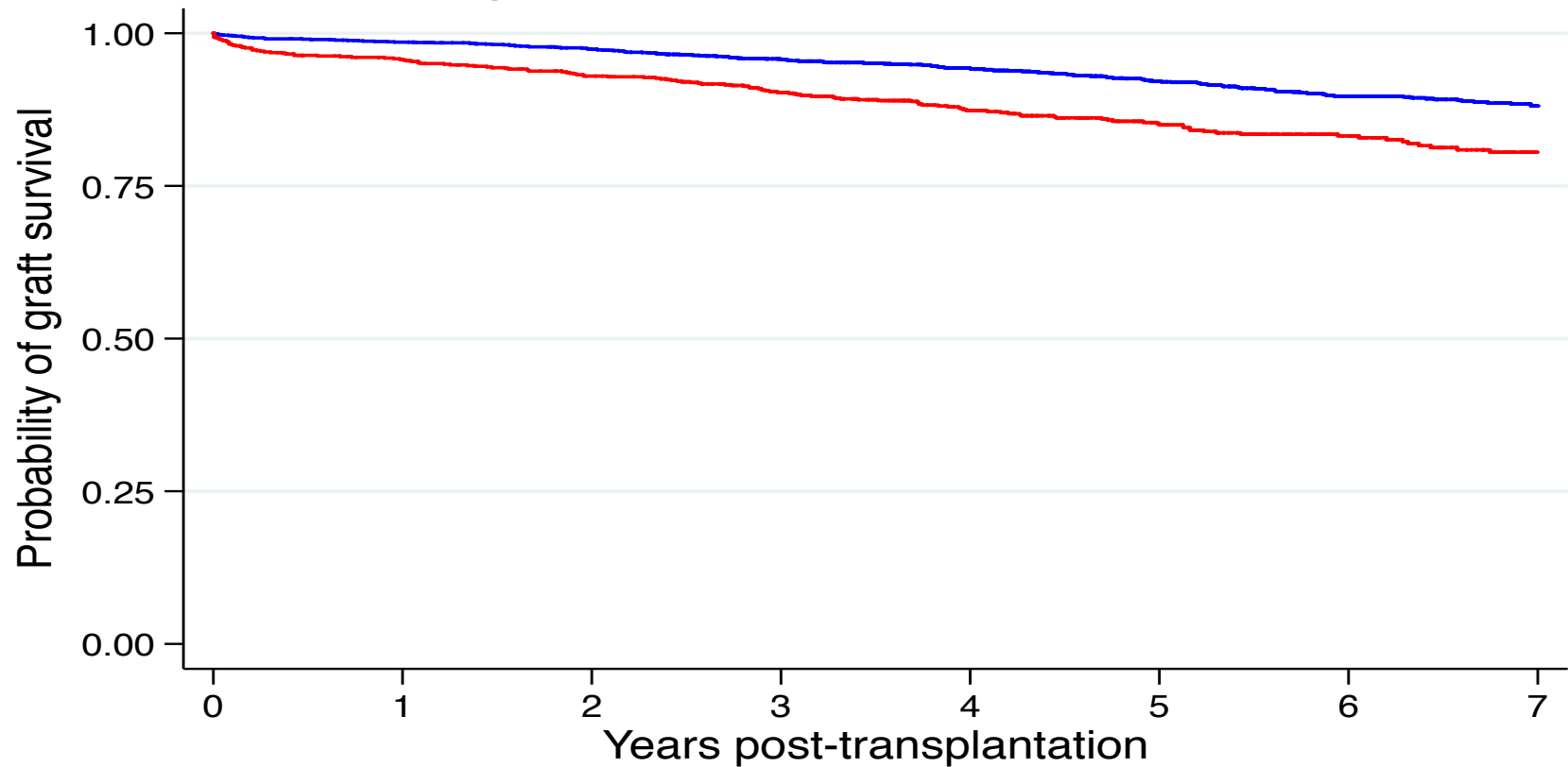
Wilcoxon (Breslow) test for equality of survivor functions

ECD_FINAL	Events observed	Events expected	Sum of ranks
0	187	234.61	-109077
1	146	98.39	109077
Total	333	333.00	0

chi2(1) = 37.30

Pr>chi2 = 0.0000

Kaplan-Meier survival estimates



Number at risk

Standard	1847	1810	1760	1594	1352	1039	758	526
Expanded	916	854	806	705	558	419	292	179



Modèle de Cox

Objectif du modèle de Cox

- **Quantifier** la relation entre un facteur et la survenue d'un évènement
- Prise en compte simultanée de **plusieurs covariables** (ajustement multivarié)
- Adapté aux situations où les évènements surviennent à des moments différents
- Très utilisé en épidémiologie

Définition

- Variable à expliquer QUALITATIVE (à 2 classes dans le cas le + simple: 0/1)
- Variables explicatives qualitatives/quantitatives avec ≥ 2 classes
- Les fonctions de survie nont pas de forme paramétrique précise (modèle semi-paramétrique)

Notation

Le risque instantané à l'instant t pour un sujet de caractéristiques Z (tabac, age, sexe, ...) s'écrit :

$$h(t, Z) = h_0(t) c(bZ)$$

fonction dépendante du tps

fonction INDEPENDANTE du tps

$$= \exp(\sum b_i Z_i)$$

permet une fonction
de risque positive

b_i = coeff de régression à estimer

Z_i = vb explicatives (tabac, age, sexe, ...)

Notation

- $h_0(t)$
 - “baseline function”: toutes les covb sont à 0
 - Dépend du temps
 - n' est pas spécifiée: modèle non paramétrique
 - En pratique rarement estimée sauf dans le domaine de la prédiction individuelle

Estimation du rapport de risque

- Soit la vb Z (statut fumeur) codée en 1 chez les fumeurs et en 0 chez les non fumeurs
- Le risque instantané $h(t) = h_0(t) \exp(bZ)$ s'écrit
 - $h_0(t) (\exp b)$ chez les fumeurs
 - $h_0(t)$ chez les non fumeurs
- Le RR (rapport des risques instantanés) s'écrit à chaque t
 $h_0(t) (\exp b) / h_0(t) = \exp(b)$
-  $\exp(b)$ mesure donc le RR associé à un facteur donné (ici en 2 classes) et ne dépend pas du temps

Estimation de β

- Maximisation de la vraisemblance partielle
 - Vraisemblance du modèle à chaque temps d'évt

Interprétation

- 3 statistiques de test
 - test du rapport de vraisemblance
 - test de Wald
 - test du score
- Ces 3 tests sont asymptotiquement (quand n grand) équivalents et suivent sous H_0 un X^2 à 1ddl (dans le cadre d'un modèle à une covariable)
- On peut aussi se baser sur l'estimation de l'IC 95% du RR

Exemple sous Stata

```
. xi:stcox i.ECD_FINAL
i.ECD_FINAL      _IECD_FINAL_0-1      (naturally coded; _IECD_FINAL_0 omitted)

      failure _d:  graft_lost_status_new == 1
      analysis time _t:  delai_en_a
```

```
Iteration 0:  log likelihood = -2480.6638
Iteration 1:  log likelihood = -2465.5412
Iteration 2:  log likelihood = -2465.3677
Iteration 3:  log likelihood = -2465.3677
Refining estimates:
Iteration 0:  log likelihood = -2465.3677
```

Cox regression -- Breslow method for ties

```
No. of subjects =          2,763      Number of obs      =          2,763
No. of failures =           333
Time at risk    =    15324.1404
```

Log likelihood = -2465.3677

```
LR chi2(1)      =          30.59
Prob > chi2     =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_IECD_FINAL_1	1.868754	.2070832	5.64	0.000	1.503929	2.322077

.
end of do-file = exp (B)

Exemple sous R

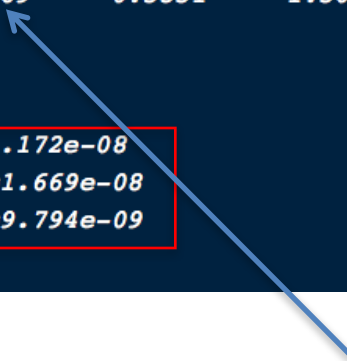
```
> summary(coxph(Surv(BASE_BMJ_ECD$delai_en_a, BASE_BMJ_ECD$graft_lost_status_new) ~ BASE_BMJ_ECD$ecd_final_ecd_final))
Call:
coxph(formula = Surv(BASE_BMJ_ECD$delai_en_a, BASE_BMJ_ECD$graft_lost_status_new) ~
      BASE_BMJ_ECD$ecd_final_ecd_final)

n= 2763, number of events= 333

              coef exp(coef) se(coef)      z Pr(>|z|)
BASE_BMJ_ECD$ecd_final_ecd_finalECD 0.6253    1.8689    0.1108 5.643 1.67e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
BASE_BMJ_ECD$ecd_final_ecd_finalECD    1.869    0.5351    1.504    2.322

Concordance= 0.583 (se = 0.014 )
Rsquare= 0.011 (max possible= 0.834 )
Likelihood ratio test= 30.6 on 1 df,  p=3.172e-08
Wald test               = 31.85 on 1 df,  p=1.669e-08
Score (logrank) test = 32.88 on 1 df,  p=9.794e-09
```



= exp (B)

Hypothèse principale à vérifier: risques
proportionnels

Définition

- Le rapport des risques instantanés pour 2 sujets de caractéristiques Z_1 et Z_2 (fum/non fum) vaut
$$\frac{h_0(t) \exp(b Z_1)}{h_0(t) \exp(b Z_2)}$$
$$= \exp(b Z_1) / \exp(b Z_2)$$
- le rapport des risques instantanés (RR) est **indépendante du temps** et est **constant** au cours du tps: **hypothèse des risques proportionnels**
- dit autrement, “l’effet” d’une covariable sur la survenue de l’évènement est constant au cours du tps

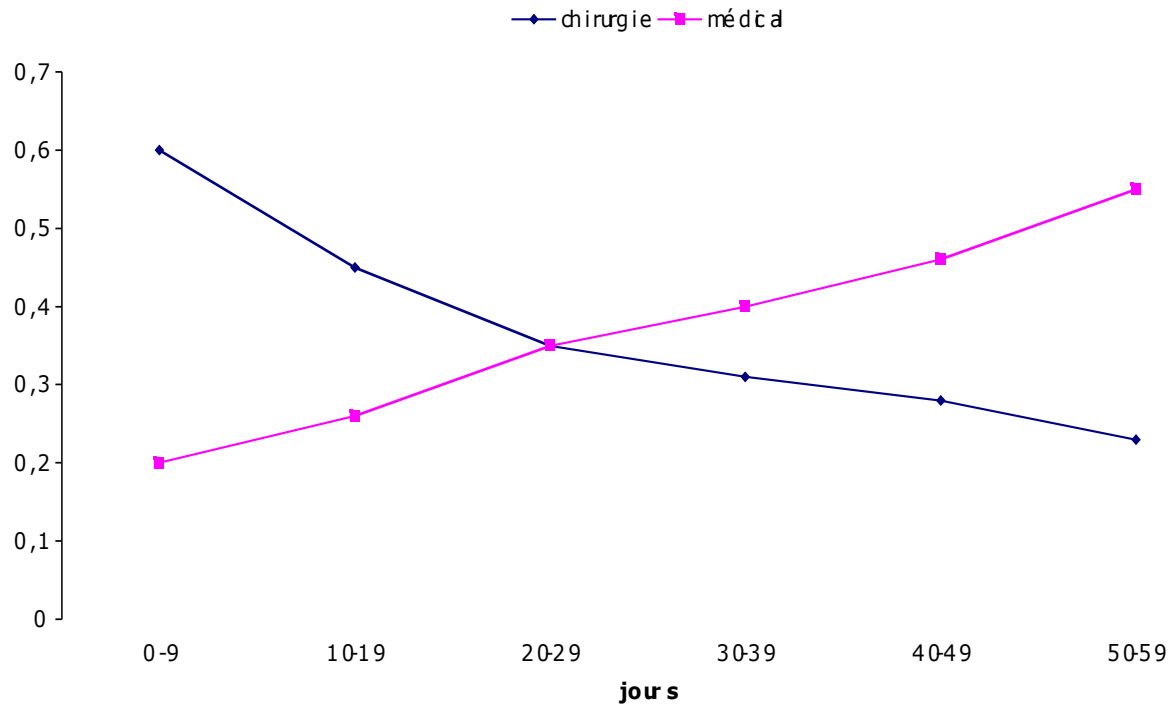
Méthode d'évaluation

- Graphiques
 - Parallélisme des courbes de survie observées
 - Courbes de survie observées vs. prédites
- “Goodness of fit”: valeurs prédites vs. valeurs observées
- Interaction avec le temps
- A tester (en théorie) pour chaque covariable du modèle; tester en pratique ++ sur la vb d'exposition principale

Parallelisme des courbes de survie **observées**

- Les courbes de survie des 2 groupes restent “parallèles” au cours du suivi

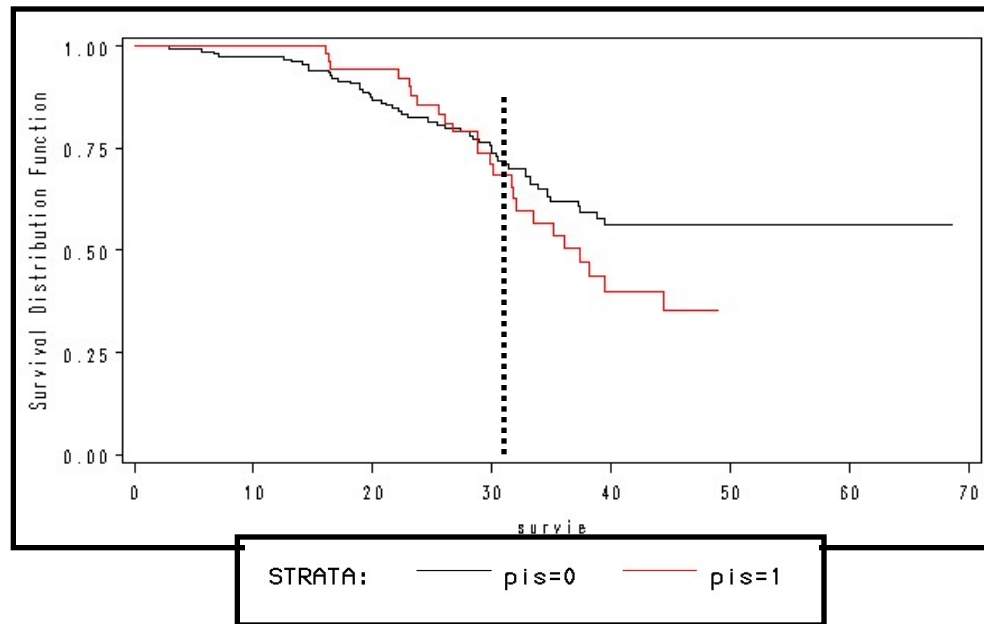
Méthodes graphiques



Risque de mortalité après tt antiKC médical vs. chirurgical

Les courbes des risques instantanés se croisent

Méthodes graphiques



On peut isoler 2 périodes de part et d'autre de 30j

Méthode graphique

- Simplicité des méth. graphiques,
- Mais
 - **subjectivité**=pas de test statistique
 - nbre de catégories à constituer pour une vb continue ? Résultats possiblement différents en fonction du nbre de catégories.
 - Choisir des cut off pertinents
- Adopter une attitude **conservatrice**, sauf si forte évidence de non proportionnalité

- Résidut de Schoenfeld
 - Test analytique basé sur les résidus de Schoenfeld: On calcule la corrélation entre les résidus et une fonction du temps basée sur l'estimateur de KM
 - Test graphique des résidus de Schoenfeld: il doit y avoir absence de tendance temporelle.

Exemple sous R

On teste la variable DSA (anticorps le jour de la greffe):

```
> cox.mydata <- coxph(Surv(BASE_BMJ_ECD$delai_en_a, BASE_BMJ_ECD$graft_lost_status_new) ~ BASE_BMJ_ECD$j0_dsa_v3_2)
> cox.zph (cox.mydata)
```

	rho	chisq	p
BASE_BMJ_ECD\$j0_dsa_v3_2	0.0101	0.0337	0.854

Le test de validité de l'hypothèse des risques proportionnels conduit à rejeter cette hypothèse: $p > 0.05$

En cas d'interaction avec le temps

- Rejet hypothèse des risques proportionnels
- 2 solutions:
 - Conserver dans le modèle le terme d'interaction
 - STRATIFIER le modèle
 - sur la(les) covariable(s) ne vérifiant pas l'hypothèse de proportionalité des risques
 - sur une période de rupture

Résumé

- Pas de méthode d'adéquation parfaite
- Tester plusieurs méthodes
- Cox suffisamment robuste, ++ quand effectifs importants

Utilité du RR

- Confusion fréquente entre RR et risque absolu
- Le RR est un rapport de probabilités entre 2 groupes de populations
- Le risque absolu est une probabilité, le + souvent estimée au niveau individuel

RR et R absolu

- RR: mesure d'association
pas d'intérêt sur le plan de la prédiction
des RR élevés ou très bas peuvent en fait
concerner des evts rares: attention à
l'interprétation qu'on en fait.
- R Absolu: probabilité, intérêt dans la
prédiction individuelle du risque

En pratique : similaire à la régression logistique

- Commencer par une analyse univariée en visualisant le tableau de contingence et en faisant une estimation de le RR par Cox univarié
- Si une variable exerce un effet de confusion très important sur l'ensemble des autres variables, tester avec un modèle logistique réduit la valeur P de chaque exposition après ajustement sur la variable confondante
- Tester les interactions qui vous paraissent épidémiologiquement pertinentes avec la variable d'exposition principale
- Commencer avec un modèle contenant les variables forcées et les variables trouvées associées à un $p < 0,2$ en analyse univariée. Tester l'effet du retrait des variables une à une sur l'ensemble du modèle
- Ne garder au final que les variables dont le degré de significativité est $< 0,05$, et les variables forcées

PROGRAMME

Vendredi 24 novembre

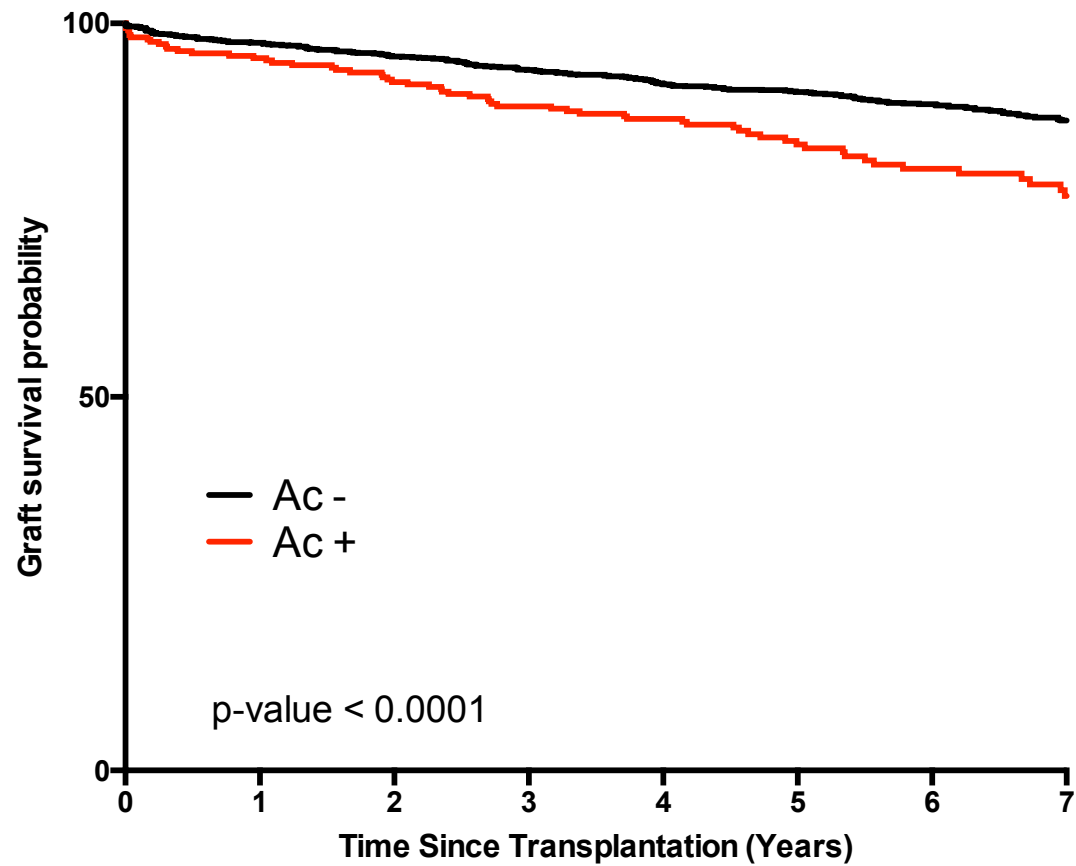
12h00-14h00	Repas
14h00-14h30	Les statistiques : pourquoi faire ? (introduction générale)
14h30-16h00	Créer une base de données : le B-A BA
16h00-16h30	Intermède - café
16h30-19h30	Comparer mes résultats : comment faire avec des logiciels simples (<i>en ligne, excel et Prism</i>) ou avec d'autres plus expérimentés
20h30	Dîner

Samedi 25 novembre

8h30-10h30	Les statistiques : comment aller plus loin ? (<i>Introduction à la régression logistique et aux analyses de survie</i>)
10h30-11h00	Intermède - café
11h00-12h30	Mise en pratique : Courbes de Kaplan-Meier et analyses univariées
12h30-13h00	Mon kit de survie pratique en (aux !) statistiques

Réaliser une courbe de survie en
fonction de la présence
d'anticorps au moment de la
greffe

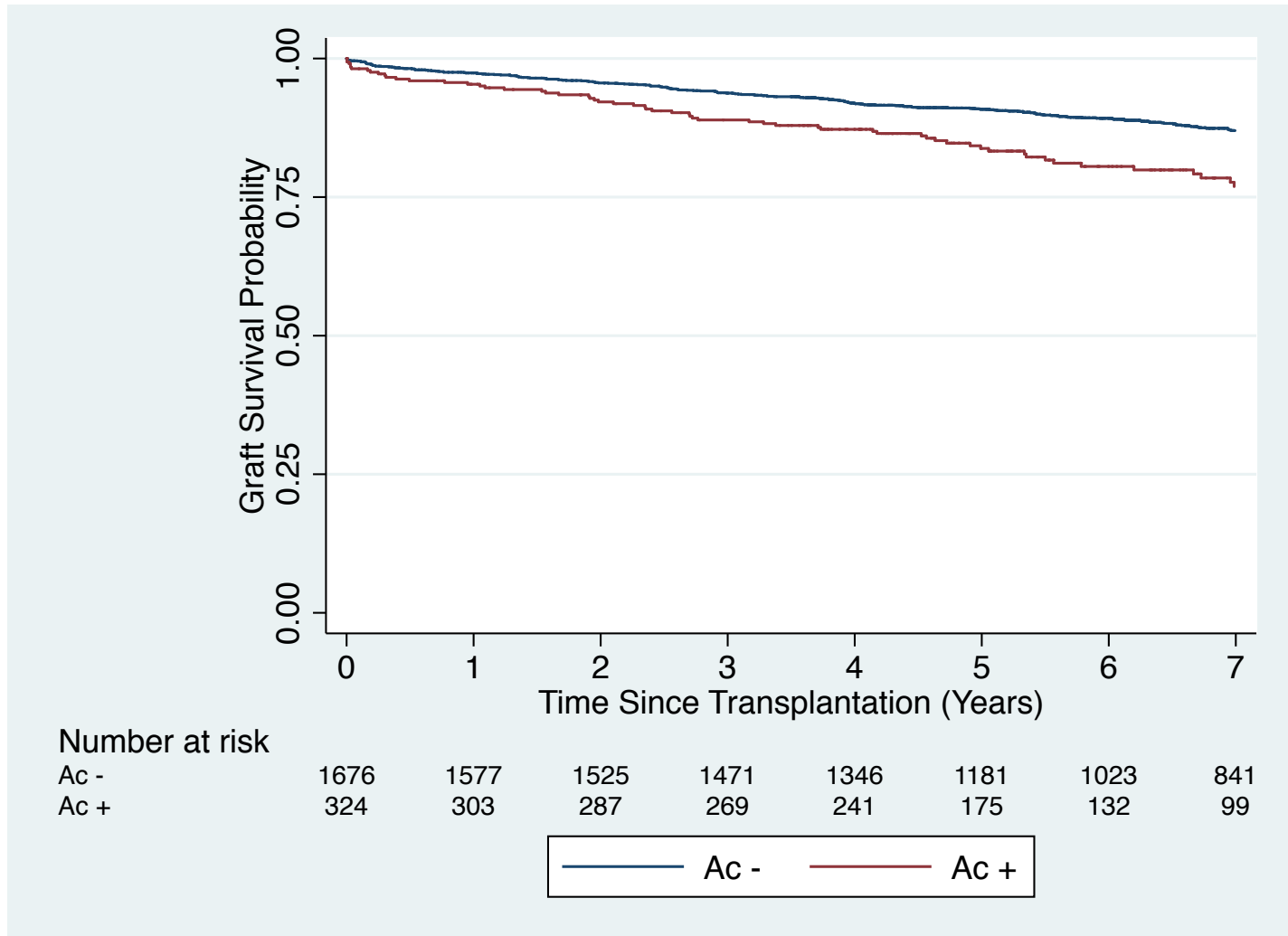
Sous Prism



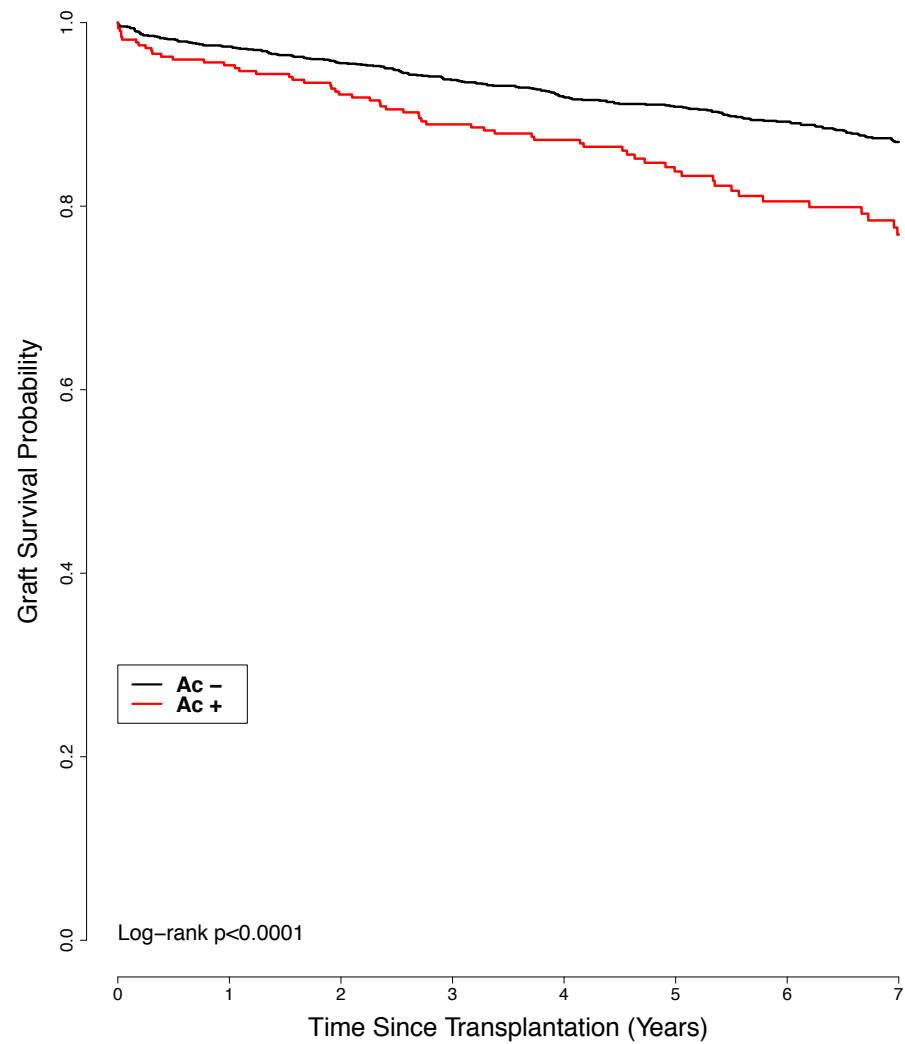
N at risk

Ac -	1676	1577	1525	1471	1346	1181	1023	841
Ac +	324	303	287	269	241	175	132	99

Sous Stata



Sous R



N at Risk

Ac -	1676	1577	1525	1471	1346	1181	1023	841
Ac +	324	303	287	269	241	175	132	99

Analyse univariée : Log-rank

Survival Curve comparison			
1	Comparison of Survival Curves		
2			
3	Log-rank (Mantel-Cox) test		
4	Chi square	17.30	
5	df	1	
6	P value	< 0.0001	
7	P value summary	****	
8	Are the survival curves sig different?	Yes	
9			
10	Gehan-Breslow-Wilcoxon test		
11	Chi square	16.55	
12	df	1	
13	P value	< 0.0001	
14	P value summary	****	
15	Are the survival curves sig different?	Yes	
16			
17	Median survival		
18	AC -	Undefined	
19	Ac +	Undefined	
20			
21	Hazard Ratio (Mantel-Haenszel)	A/B	B/A
22	Ratio (and its reciprocal)	0.4938	2.025
23	95% CI of ratio	0.3541 to 0.6886	1.452 to 2.824
24			
25	Hazard Ratio (logrank)	A/B	B/A
26	Ratio (and its reciprocal)	0.5669	1.764
27	95% CI of ratio	0.3550 to 0.6870	1.456 to 2.817
28			

Analyse univariée : Cox univarié sous R

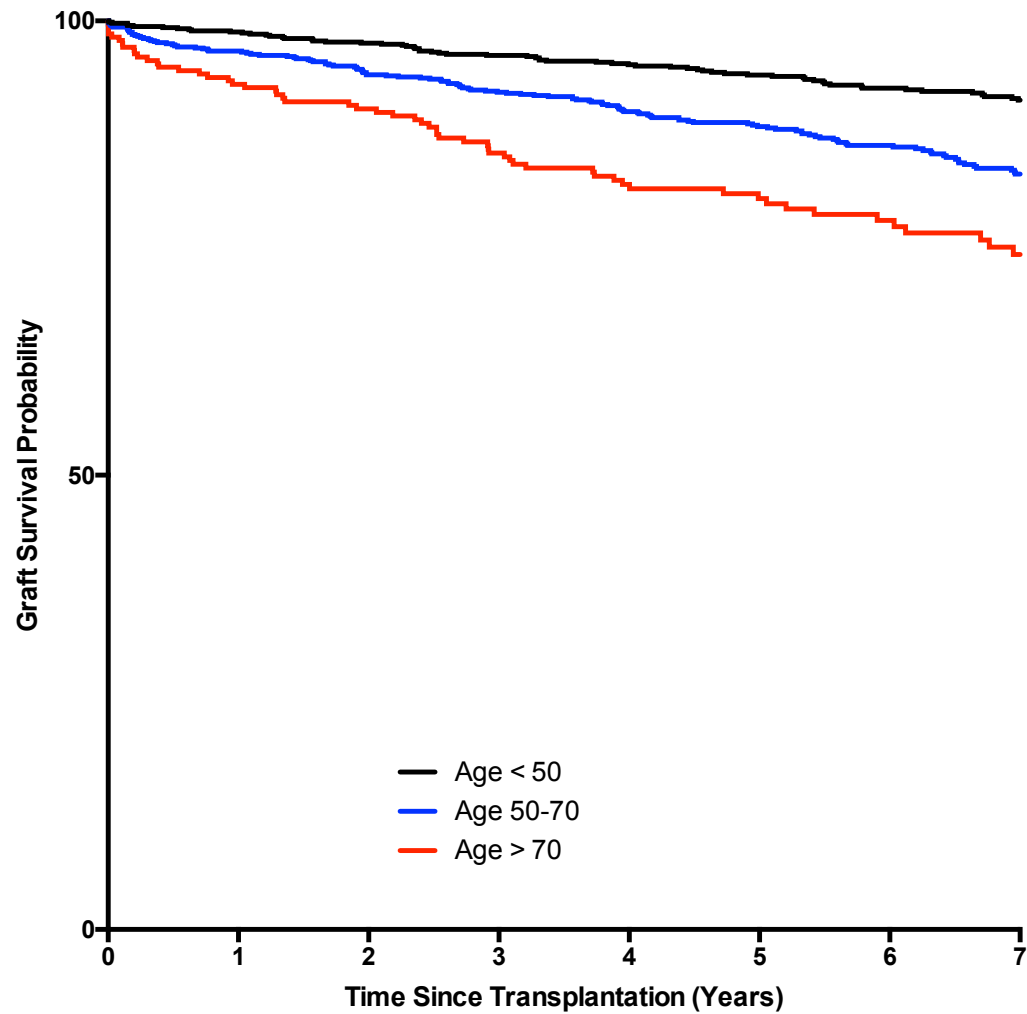
```
              coef exp(coef) se(coef)      z Pr(>|z|)
ANTICORPS_J01 0.5781    1.7826  0.1409 4.103 4.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	<code>exp(coef)</code>	<code>exp(-coef)</code>	lower .95	upper .95
ANTICORPS_J01	1.783	0.561	1.352	2.35

```
Concordance= 0.543 (se = 0.011 )
Rsquare= 0.007 (max possible= 0.895 )
Likelihood ratio test= 15.05 on 1 df, p=0.0001047
Wald test               = 16.84 on 1 df, p=4.076e-05
Score (logrank) test = 17.3 on 1 df, p=3.19e-05
```

Estimer une variable continue
avec KM

Sous Prism



Sous Prism

Survival		
Curve comparison		
1	Comparison of Survival Curves	
2		
3	Log-rank (Mantel-Cox) test (recommended)	
4	Chi square	45.94
5	df	2
6	P value	< 0.0001
7	P value summary	****
8	Are the survival curves sig different?	Yes
9		
10	Logrank test for trend (recommended)	
11	Chi square	45.45
12	df	1
13	P value	< 0.0001
14	P value summary	****
15	Sig. trend?	Yes
16		
17	Gehan-Breslow-Wilcoxon test	
18	Chi square	52.23
19	df	2
20	P value	< 0.0001
21	P value summary	****
22	Are the survival curves sig different?	Yes

Analyse univariée : Cox univarié sous R avec variable en continue

```
n= 2000, number of events= 314

              coef exp(coef) se(coef)      z Pr(>|z|)
Age_Donneur 0.023203  1.023474 0.003679  6.307 2.85e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
Age_Donneur      1.023      0.9771      1.016      1.031

Concordance= 0.618  (se = 0.017 )
Rsquare= 0.02  (max possible= 0.895 )
Likelihood ratio test= 41.39  on 1 df,   p=1.248e-10
Wald test               = 39.78  on 1 df,   p=2.848e-10
Score (logrank) test = 40.08  on 1 df,   p=2.433e-10
```

Kit de Survie

Le projet/la base de données:

- Une hypothèse de travail claire
- Une base 'propre': une colonne par variable, un type de donnée par colonne, des données homogènes

Les tests simples:

- Paramétriques = grands échantillons = la distribution vérifie la loi normale et l'égalité des variances
- Non paramétrique = petits échantillons, indépendant de la distribution, à privilégier

La régression logistique:

- relation entre une variable Y qualitative à deux classes (oui/non) à temps fixe et des variables Xj qui peuvent être quantitatives ou qualitatives
- Odds-Ratio (OR) = risque relatif = augmentation (la diminution) du risque d'être malade pour les exposés comparés aux non exposés

Le modèle de Cox:

- Etude de la survenue d'un événement au cours du temps
- Attention: décès, abandon, perdus de vue // facteurs d'interaction, de confusion, colinéarité
- Attention: le Kaplan-Meier représente une analyse **univariée**!

Les outils:

- Excel
- GraphPad prism (QuickCalcs, <https://www.graphpad.com/quickcalcs/>) (permet de réaliser une courbe de Kaplan-Meier en plus des tests stats simples)
- BiostaTGV (<http://marne.u707.jussieu.fr/biostatgv/?module=tests>)
- script R

Outils d'aide à la compréhension, forums

- GraphPad Statistics Guide (<https://www.graphpad.com/guides/prism/7/statistics/>)
- Stack overflow (<https://stackoverflow.com/>)



paristransplantgroup.org

@ParisTxGroup 

