Lecture Critique d'Articles
Et
Biostatistiques

Dr E. Chazard, Pr R. Beuscart

CERIM

Faculté de Médecine

Université Lille 2

Introduction
Plan d'un article scientifique

Résumé (abstract)

Introduction

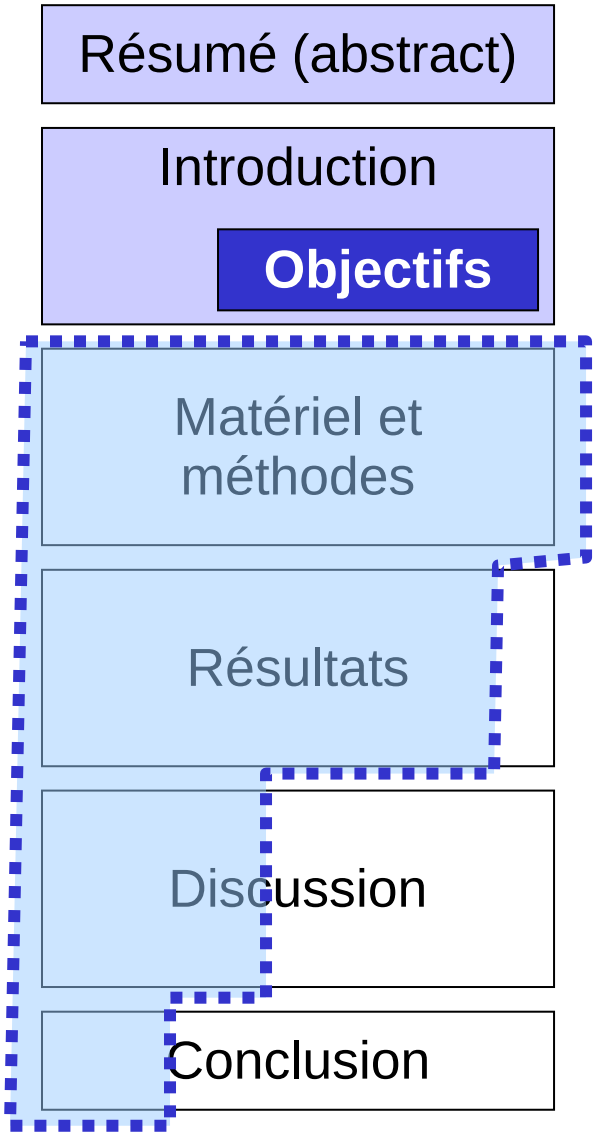
Matériel et méthodes

Résultats

Discussion

Conclusion

Introduction



- Résumé (abstract)
  - Toujours disponible en ligne gratuitement, indexé
- Introduction
  - Contexte, données connues sur la maladie, etc.
  - Question générale, état de l'art pour répondre à cette question (méthodes et résultats des autres chercheurs)
  - Objectif général et objectifs opérationnels *=> la suite doit strictement répondre à l'objectif (point à vérifier)*

Introduction

Résumé (abstract)

Introduction

Objectifs

Matériel et méthodes

Résultats

Discussion

Conclusion

- Matériel et Méthodes :
  - Mode de sélection / inclusion des patients
  - Données étudiées
  - Méthodes statistiques utilisées
- Résultats
  - Toujours :
    - Données traitées : nombre de patients sollicités, nombre de patients inclus, etc.
      => vérifier que ces données sont publiées
    - Description univariée : chaque variable, séparément
    - Statistiques univariées : estimation, intervalles de confiance
  - Pour les études analytiques (≠descriptives) :
    - Description multivariée : représentation de 2 ou plus variables
    - Statistiques multivariées : inférence statistique, association entre ces variables

Introduction

Résumé (abstract)

Intro-duction

Objectifs

Déjà connu

Matériel et méthodes

Résultats

Discussion

Conclusion

- Discussion
  - Discussion des résultats, énumération systématique des biais et réserves
    => doit être systématique
  - Confrontation aux résultats connus
  - Discussion de la généralisation des résultats, perspectives, etc.
    => cf. notamment biais de sélection et adéquation entre les objectifs et matériel et méthodes
- Conclusion
  - Succincte

Introduction
Statistiques et méthodologie

Résumé (abstract)	
Introduction	Objectifs
Matériel et méthodes	Sélection / inclusion
	Méthodes statistiques
Résultats	Données traitées
	Univarié
	Bivarié
Discussion	Résultats, biais
	Comparaison connu
	Généralisation
Conclusion	

Problèmes méthodologiques à discuter en LCA

Problèmes statistiques à discuter en LCA

Introduction
Typologie simplifiée des études

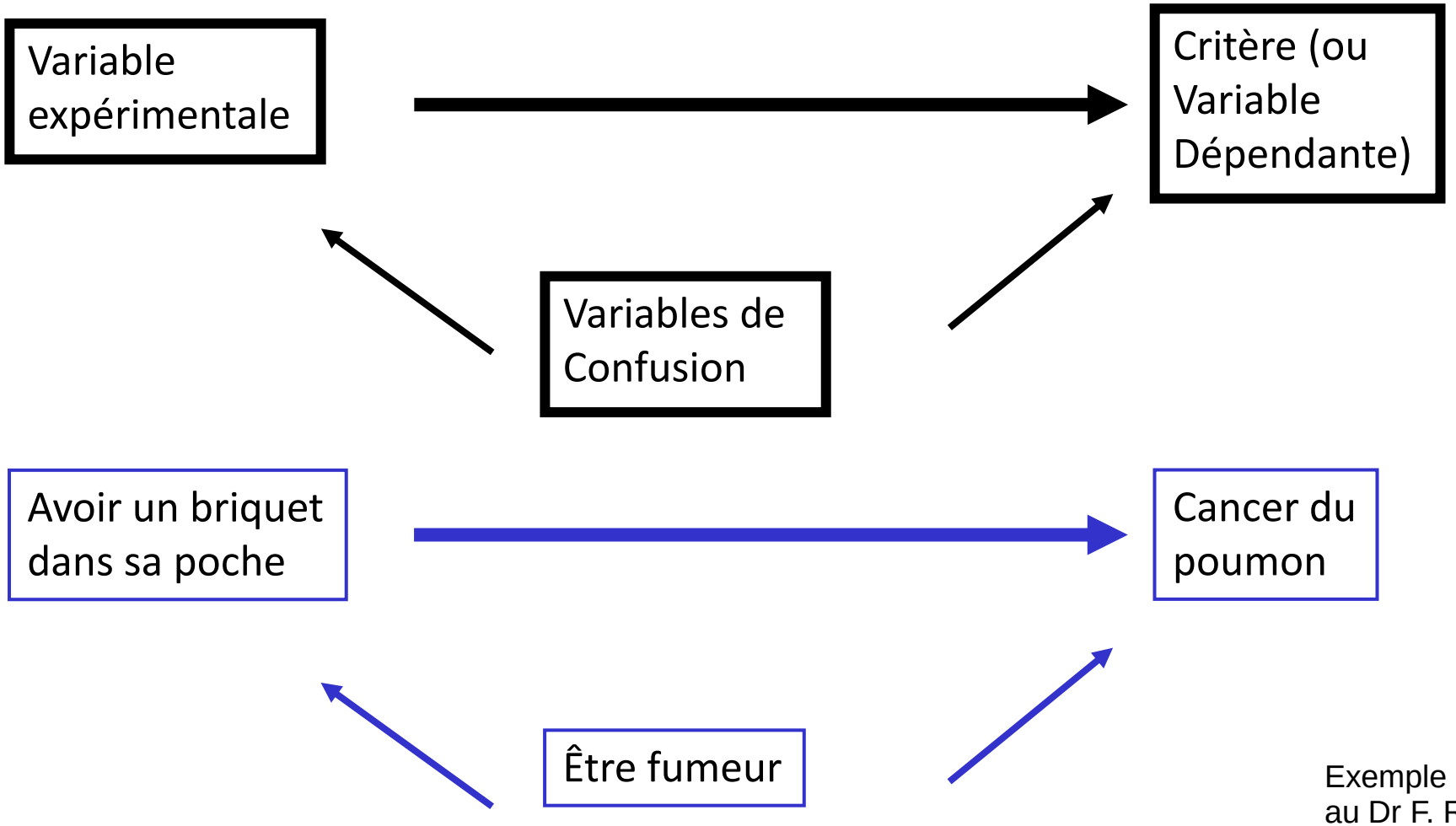
- Études épidémiologiques
  - Descriptives : calculer incidence, prévalence
  - Analytiques : inférence statistique, association entre variables
    - Cas-témoin, rétrospectives
    - Cohortes prospectives (longitudinales)
    - Cohortes historiques / rétrospectives (longitudinales)
- Essais thérapeutiques, épidémiologie interventionnelle

1. Les VARIABLES

Objectif Pédagogique 1: Les Variables

- Etudes descriptives
- Etudes analytiques :
  - Variables à expliquer :
    critère principal, critères secondaires. Exemples :
    - Binaire (malade/sain)
      - Cas particulier : avec notion de temps (survie)
    - Quantitative : diamètre d'une tumeur, score…
      - Cas particulier : nombre d'événements (Poisson)
    - Qualitative (type de tumeur) éventuellement ordonnée (stade d'une maladie)
  - Variables explicatives :
    variables associées au critère, censées l'expliquer. Les mêmes types que précédemment (sauf survie).
  - Variables de confusion

Variables de confusion



Exemple emprunté au Dr F. Richard

Variables de confusion

- Définition :
  - Facteur de confusion = facteur qui perturbe l'association entre l'exposition étudiée et la maladie
  - Si facteur ignoré, on observe à tort une association entre l'exposition et la maladie
  - Survient notamment lorsque la variable explicative (ou l'exposition) est corrélée à une variable latente
- Etudes analytiques (cas témoin, cohortes)
  - L'exposition n'est pas fixée pour l'expérience => risque élevé
  - Solutions : Ajustement (= Analyses multivariées) OU Stratification OU Appariement
- Essais thérapeutiques et interventionnels
  - L'exposition est imposée au sujet => risque moindre
  - Illustration : on peut forcer les sujets à avoir un briquet dans la poche ou le leur interdire, cela ne devrait pas influencer le risque de cancer du poumon. L'exposition (briquet) étant imposée aléatoirement, la corrélation avec la variable de confusion (le tabagisme, vraie cause) disparait.

Objectifs. – Nous avons recherché si les patients asthmatiques qui se plaignent d'effets indésirables (EI) sous traitement anti-asthmatique différaient des autres patients en termes de caractéristiques personnelles, de perception de leur maladie ou en termes de traitements.

Méthodes. – L'étude a inclus des patients asthmatiques âgés de 18 à 50 ans, clients réguliers des pharmacies. Les patients remplissaient un autoquestionnaire qui était complété par des données informatisées de délivrance concernant les traitements anti-asthmatiques. Les patients mentionnaient tout EI ressenti et qu'ils attribuaient à leur traitement anti-asthmatique. Le contrôle de l'asthme était mesuré par l'Asthma Control Test. Les facteurs associés au fait de rapporter au moins deux EI ont été identifiés au moyen d'une régression logistique.

Variable expliquée: effets indésirables

Variable explicatives:
- **Caractéristiques des malades**
- **Contrôle de l'asthme**
- **Traitements**

Les variables aléatoires (VA)

- Rappel: une VA comporte une partie « loi du phénomène » et une partie « variabilité »
- Sources de la variabilité:
  - Variabilité inter groupe
    en général celle qu'on étudie, traduit l'effet de la variable de groupe
    ex : taille garçons > taille filles : taille influencée par le sexe
  - Variabilité intra groupe
    ex : taille variable entre les garçons
  - Variabilité individuelle
    ex : la taille de Monsieur X varie dans la journée
  - Incertitude de mesure

Les variables aléatoires

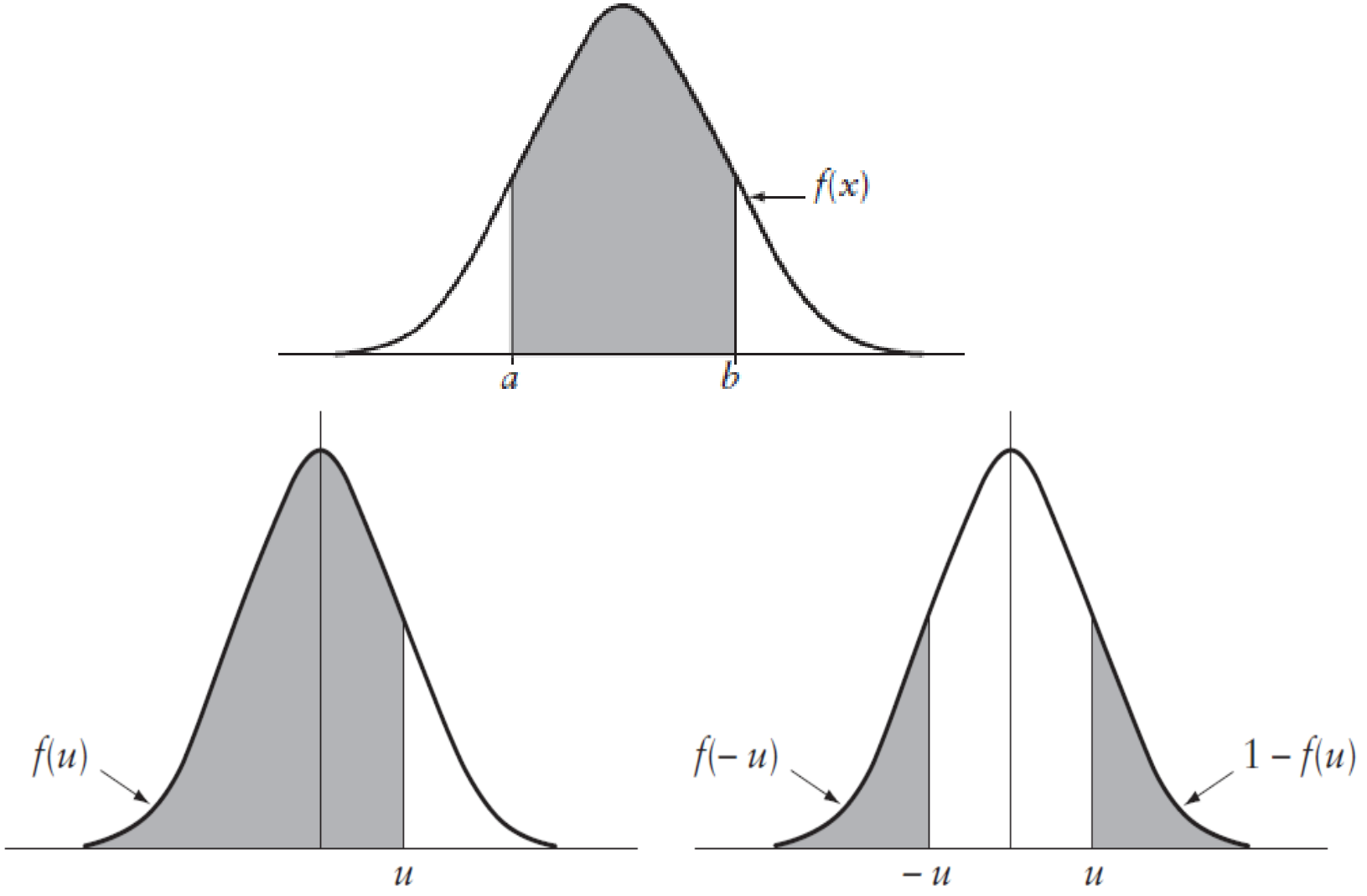
Types de Variables Aléatoires:

- Quantitatives discrètes finies ou infinies X= {1.. N}
  Ex: Nombre d'enfants d'une famille, nombre de métastases. En pratique, nombreux ex-æquo.
- Quantitatives continues X ε R ou X ε {50, 200}
  Ex: Poids, taille. En théorie, jamais d'ex-æquo.
- Catégorielles / qualitatives
  Ex: Types de fractures, types histologiques
- Ordinales (catégorielles ordonnées)
  Ex: Stades d'une maladie (NYHA : I, II, III, IV)
- Binaires : qualitatives à 2 modalités, quantitatives
  Ex : Sexe (H, F), Malade (1, 0)

Les variables suivent une distribution

- Distribution = loi de densité de probabilité
- Bon nombre de variables aléatoires :
  - suivent une loi Normale (Laplace-Gauss)
  - ou peuvent être transformées en VA qui suivent une loi Normale.
  - à défaut, certains paramètres comme la Moyenne suivent souvent une loi Normale (cf. théorème de la limite centrale et théorème des grands nombres)
- Intérêt : loi connue, tables, intervalles de confiance, …

Exemples de calculs de probabilités sur une loi tabulée :



2. Les Populations

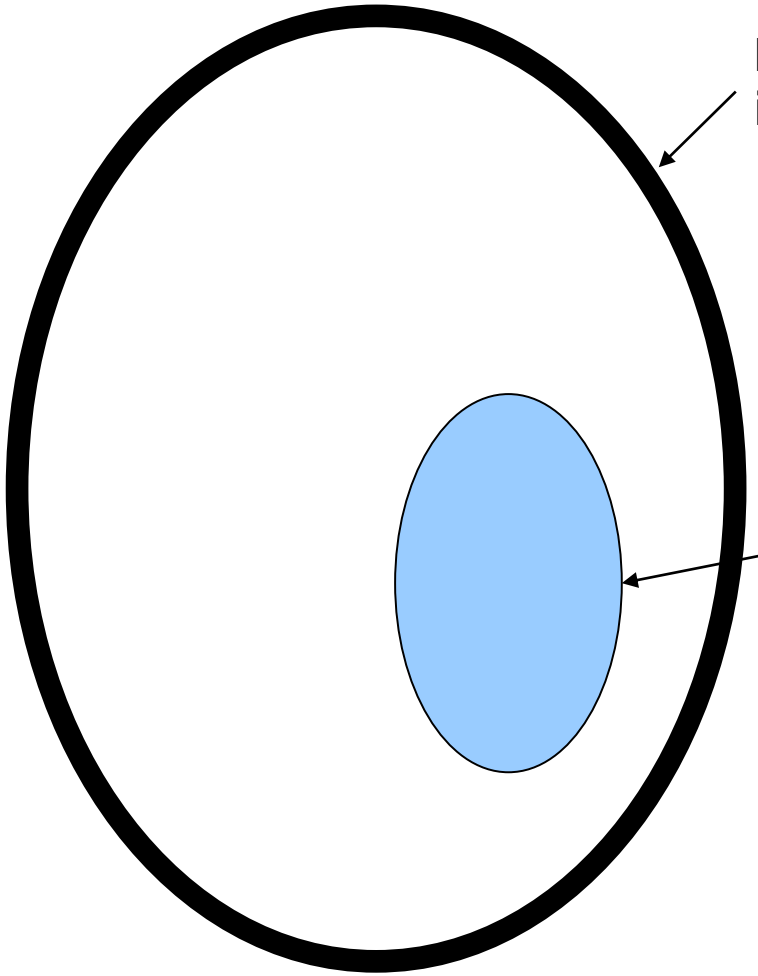
Objectif pédagogique 2 : Populations

- **Population**
  - Identifier les caractéristiques de la population étudiée
  - Analyser les modes de sélection des sujets
  - Technique de randomisation (tirage au sort)
  - Discuter la comparabilité des groupes
  - Nombre de sujets nécessaires

Population

- Les caractéristiques de la population sont importantes :
  - Démarche de la « Médecine Basée sur des Preuves »
  - Application des conclusions d'un article scientifique à la pratique médicale (décision de faire un test, de traiter, etc.)
- Pré-requis pour généraliser les résultats :
  - Connaître les caractéristiques de l’échantillon utilisé dans l’étude
  - Connaître les caractéristiques de la population visée dans la pratique médicale
  - => ces populations sont-elles comparables ?

Population et Echantillon



Population (infinie ou pratiquement infinie)

Echantillon de taille n

Echantillon: représentativité

- Il faut que l'échantillon soit représentatif de la population étudiée
- Pas de biais de recrutement (les caractéristiques de l'échantillon doivent être les mêmes que celles de la population
- Eviter tous les biais de sélection

Population : Critères d'Inclusion et d' Exclusion

- Des biais de recrutement explicites : les critères d'inclusion et de non-inclusion :
  - Age, sexe (gender), facteurs de risque
  - Pathologies associées
  - Formes de la maladie
  - Sévérité, stades de la maladie (les patients sévèrement atteints sont généralement exclus de fait, ainsi que… les décédés !)
  - Traitements associés
  - Morbidités associées
  - Personnes généralement exclues : grossesse, âge <18 ans, âge > 70 ans, sous tutelle, prisonniers, etc.
  - Etc…

Population: modalités de sélection des sujets

- Il existe aussi des critères d'inclusion « latents » :
  - Lieux de recrutement:
    - Étude mono-centrique ou multicentrique
    - Patients hospitalisés / médecine de ville / médecine ambulatoire / patients hors circuit de soins
  - Mode de recrutement:
    - Volontariat : biais de sélection (intérêt à participer, caractéristiques socioculturelles) +++
    - Sondage par tirage au sort : sur quelle base ? (ex : liste électorale, annuaire des abonnés téléphoniques, etc.)
    - Recrutement exhaustif
    - Attention aux sondages sur Internet ou, plus généralement, aux sollicitations de masse

Biais de sélection et types d'études…

- Études épidémiologiques
  - Descriptives
  - Analytiques
    - Cohortes prospectives ou historiques
    - Cas-témoin rétrospectives, transversales
- Essais thérapeutiques, épidémiologie interventionnelle

Risque que l'échantillon ne soit pas représentatif de la population ?

Oui, éventuellement

Oui, éventuellement

Oui, presque toujours :

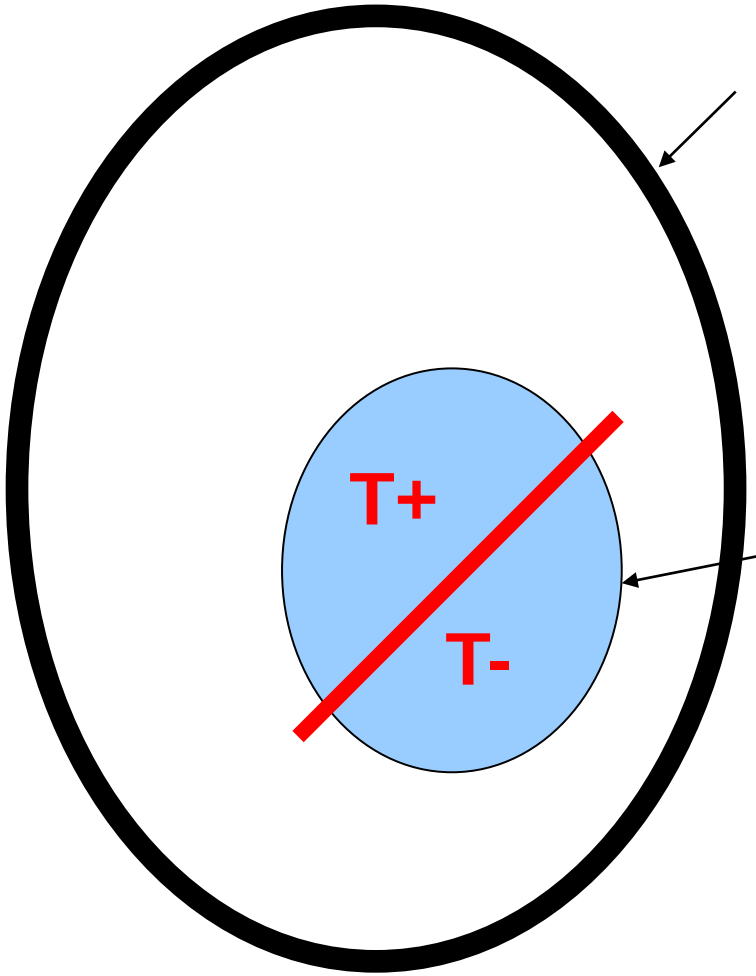
- Taux de prévalence de la maladie faussé
- Chez les malades en particulier, il manque les morts !

Oui, presque toujours* :

- Nombreux exclus
- Dont polypathologiques et polymédiqués

** Mais ce risque peut être acceptable, en outre ces études présentent de nombreuses autres qualités*

Population et Echantillon : cas des essais thérapeutiques



Population (infinie ou pratiquement infinie)

Echantillon de taille n

Deux sous-groupes :
-Patients exposés au traitement
-Patients non exposés

1) échantillon et population comparables ?
Problème de sélection des patients
2) T+ et T- sont-ils comparables ?
Problème de la méthode d'affectation

Randomisation
Essais thérapeutiques

- Tirage au sort
- Meilleure technique d'échantillonnage
- Définition :
  - Allocation aléatoire des patients dans les groupes d'étude: chaque patient a la même probabilité d'être intégré dans chacun des groupes
- Toutes les études ne peuvent pas être randomisées (ex : études observationnelles, épidémiologie analytique)
- Le groupe assigné ne doit jamais être prévisible

Population et randomisation: Points-clés à vérifier

- Vérifier les caractéristiques de la population
- Vérifier les méthodes d'échantillonnage
- Vérifier la séquence de randomisation
- Secret de l'assignation :
  - Imprévisibilité
  - Randomisation centralisée (internet, téléphone)
- Organisation générale du processus:
  - Qui a généré la séquence ?
  - Qui a assigné les patients ?

Population: calcul de sujets nécessaires

- Deux contraintes contradictoires :
  - Inclure plus de patients :
    - pour garantir une puissance suffisante à l'étude : un test non significatif de faible puissance: pas de conclusion possible
  - Inclure moins de patients :
    - Diminuer le coût de l'étude
    - Diminuer la durée de l'étude (recrutement au fil de l'eau)
    - Exigence éthique d'inclure le minimum nécessaire (avis du CPP)
- Contexte :
  - Obligatoire pour essais thérapeutiques ou études interventionnelles
  - (parfois fait pour études analytiques)

Population: calcul de sujets nécessaires

- Situation :
  - On sait que les patients sous médicament M ont une cholestérolémie plus faible que les patients sous placebo P
  - Combien de patients faut-il inclure dans un essai thérapeutique pour observer une différence significative ?
- Bases du calcul:
  - On connait (estimation) la différence attendue entre les deux groupes
    ex : une étude antérieure a montré que m_P=1,8g/l et m_M=1,6g/l
    Faute de mieux, on supposera que μ_P=1,8g/l et μ_M=1,6g/l et on espèrera observer de telles valeurs dans notre nouvel échantillon
    => exige une étude préliminaire fiable !
  - On connait la statistique de test qui sera utilisée, et sa loi de distribution
  - On fixe les risques d'erreurs, alpha=0,05 et généralement bêta =0,80 à 0,95
  - On calcule l'effectif minimal
- Nombre de Sujets Nécessaires : « matériel et méthodes »

3. Les Méthodes

Objectif Pédagogique 3. Méthodes

Deux questions:

1. Le schéma de l'étude correspond-il à l'objectif principal ?

2. Les points-clés de l'étude sont-ils respectés ?

Il faut s'assurer que la méthode employée est cohérente avec le projet du travail et qu'elle est effectivement susceptible d'apporter une réponse à la question posée dans l'introduction

Méthodes

- Estimer une prévalence, une incidence, une valeur moyenne
  - Etudes transversales
- Rechercher un facteur de risque
  - Etude de cohorte (longitudinale, exposé non exposé)
  - Etude cas-témoin
- Rechercher un facteur pronostique
  - Etude de cohorte
- Rechercher l'effet d'une action
  - Essai thérapeutique
  - Epidémiologie interventionnelle (mode de vie, alimentation…)
- Validité d'un test diagnostique
  - Sensibilité, spécificité, valeurs prédictives, courbe ROC

Méthodes

Etude expérimentale:

- Contrôle du déroulement de l'étude
- Par définition prospective
- Vivement souhaités : en aveugle, avec randomisation
- Imputation causale possible
- Ex: Essai contrôle randomisé

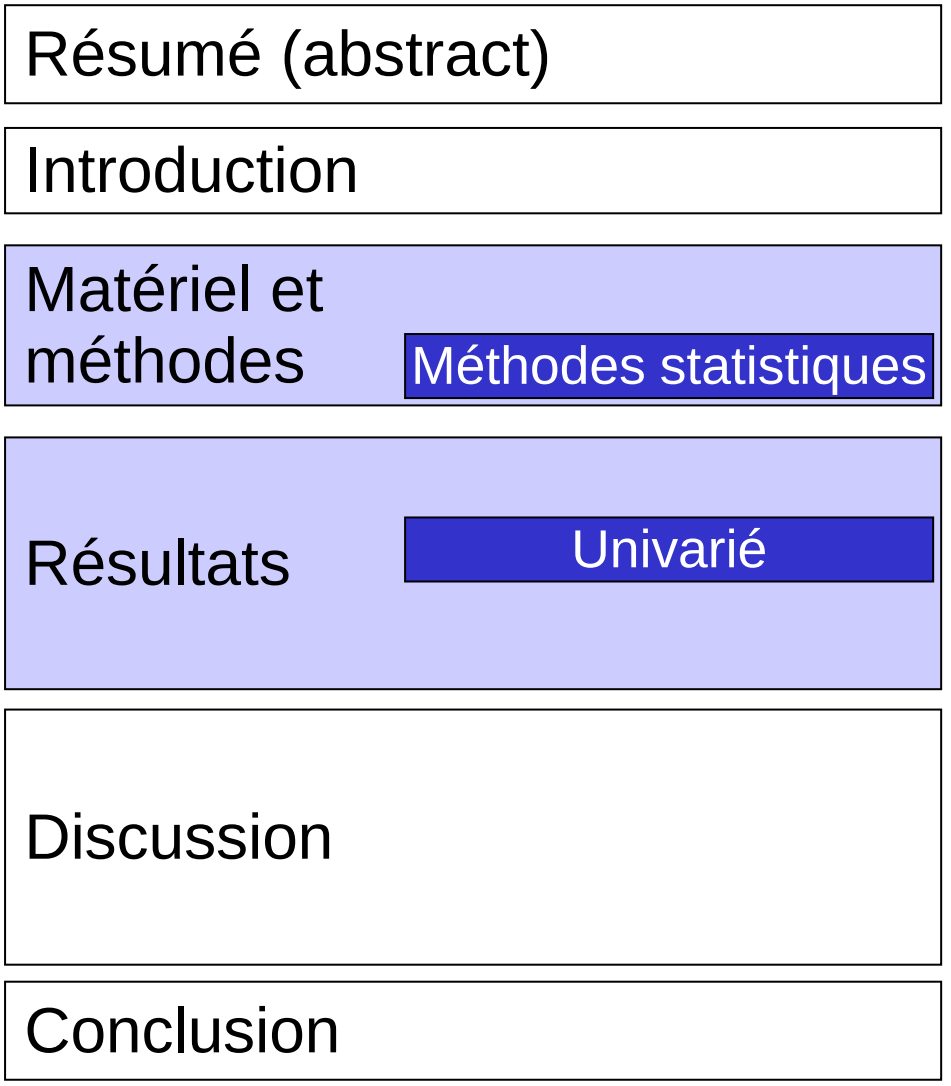
Etude Observationnelle

- Pas de Contrôle du déroulement de l'étude
- Prospective ou Rétrospective
- Aveugle et randomisation impossibles
- Risque de Biais important
- Imputation causale difficile
- Ex: Etude cas-témoins

4. Analyses univariées

a) Position
b) Représentations graphiques
c) Paramètres calculés sur l'échantillon
d) Extrapolation : estimation et IC

4.A) Position des analyses univariées

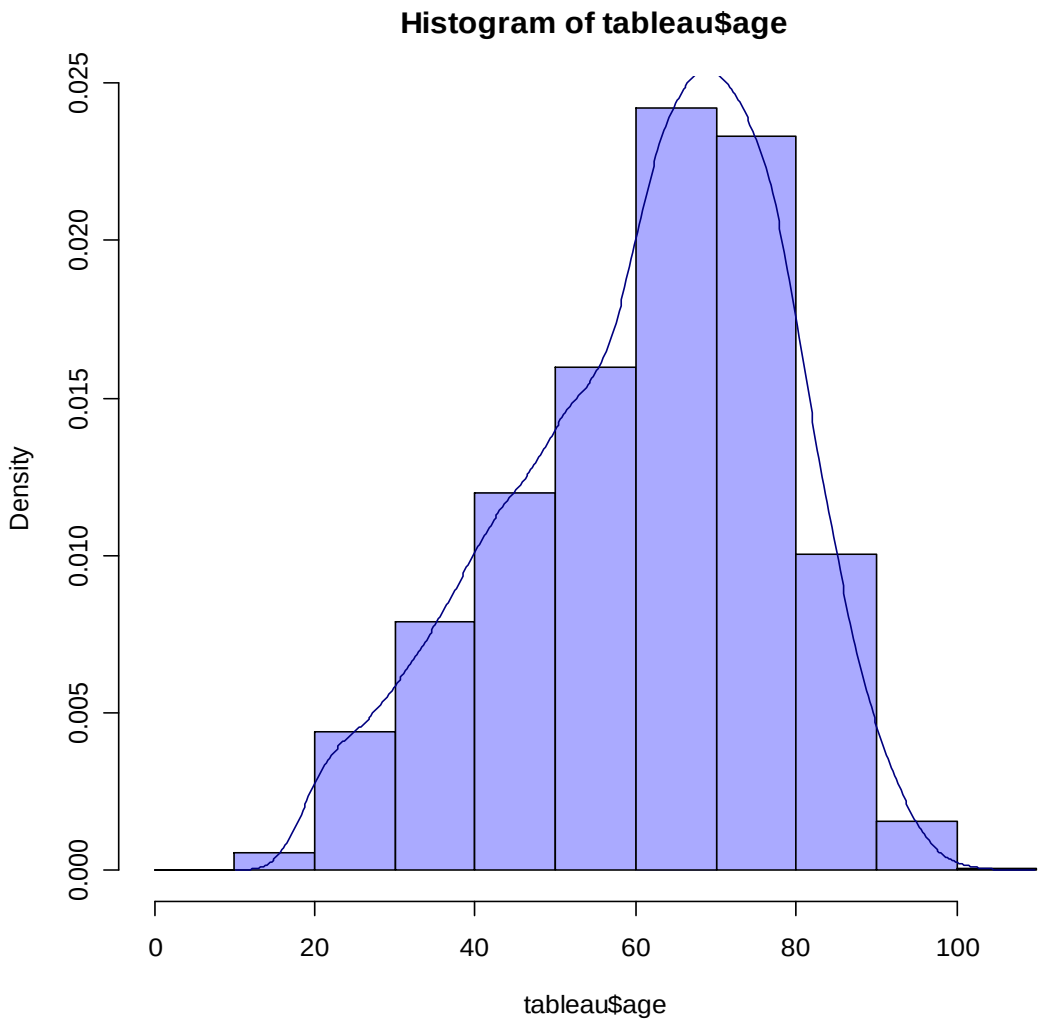


4.B.1) Représentation univariée Variables quantitatives continues

Histogramme

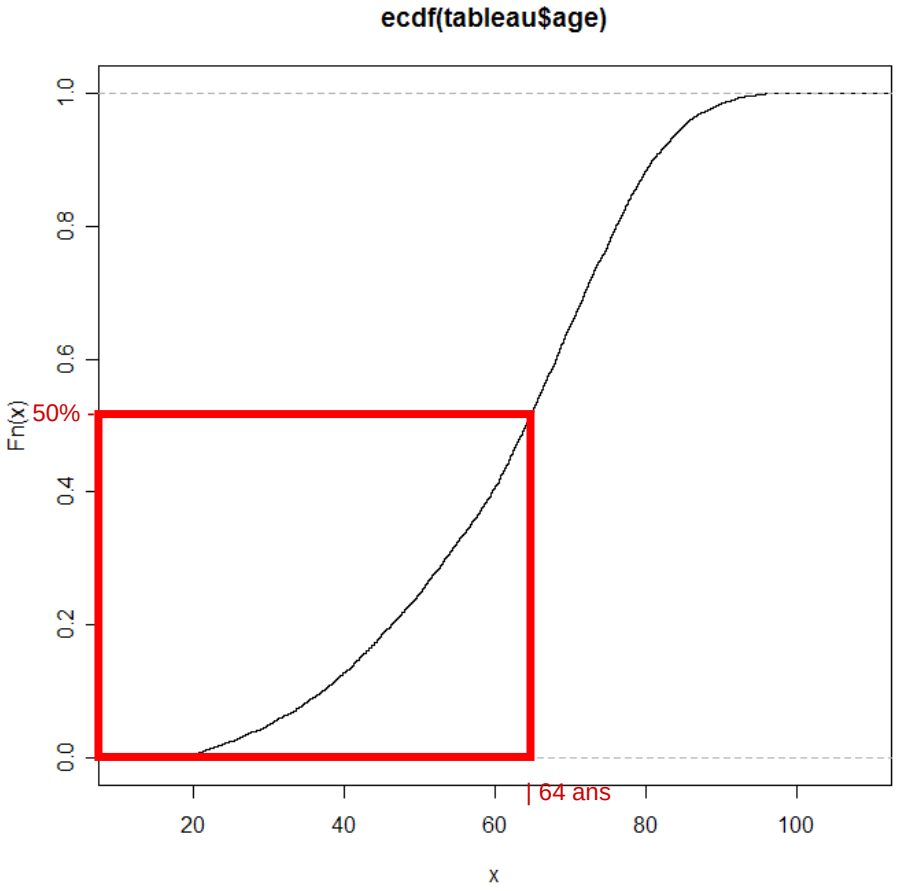
Estimation graphique de la densité de probabilité, aire=1

=> Allure normale ou non (utile pour les tests statistiques)



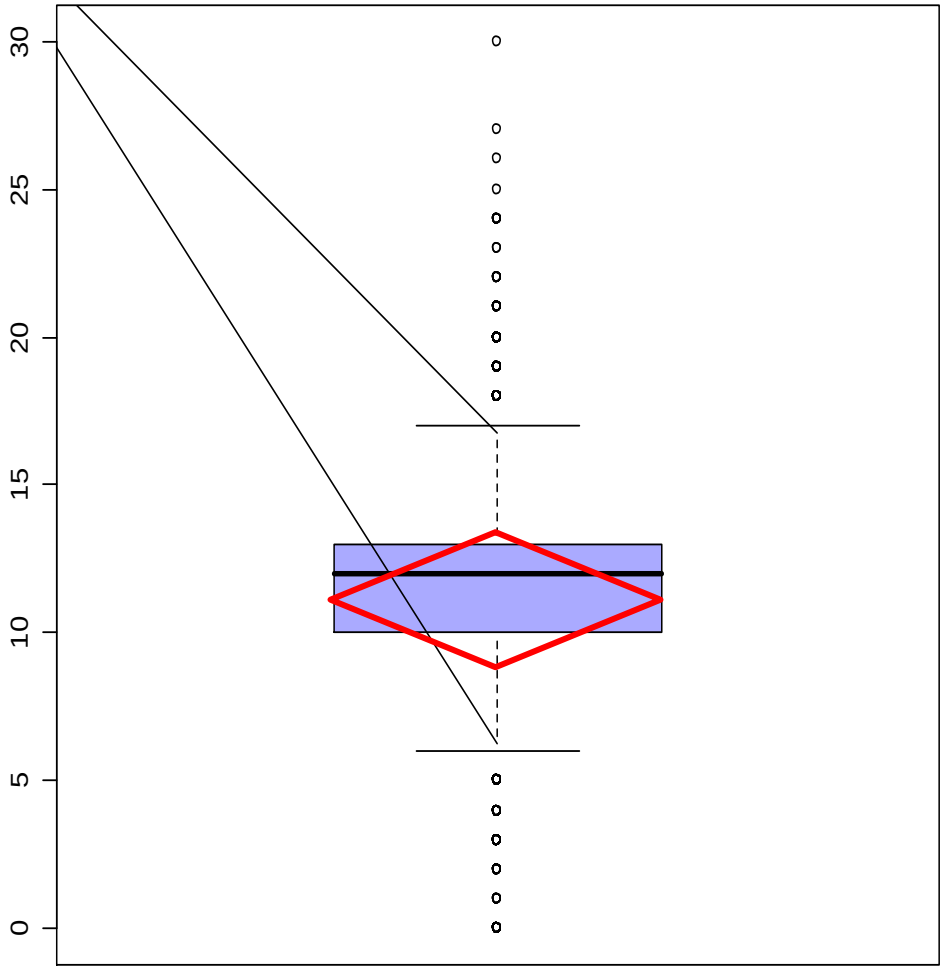
Représentation univariée
Variables quantitatives continues

- Courbe des fréquences cumulées
- Estimation graphique de la fonction de répartition
- => Permet de lire la médiane

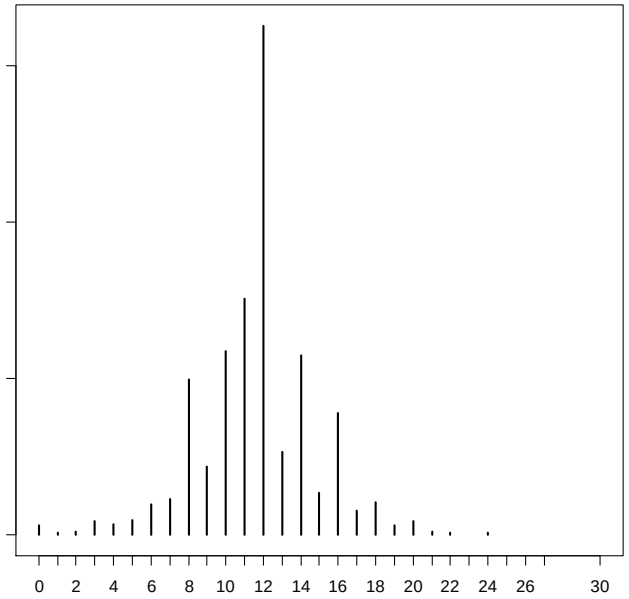


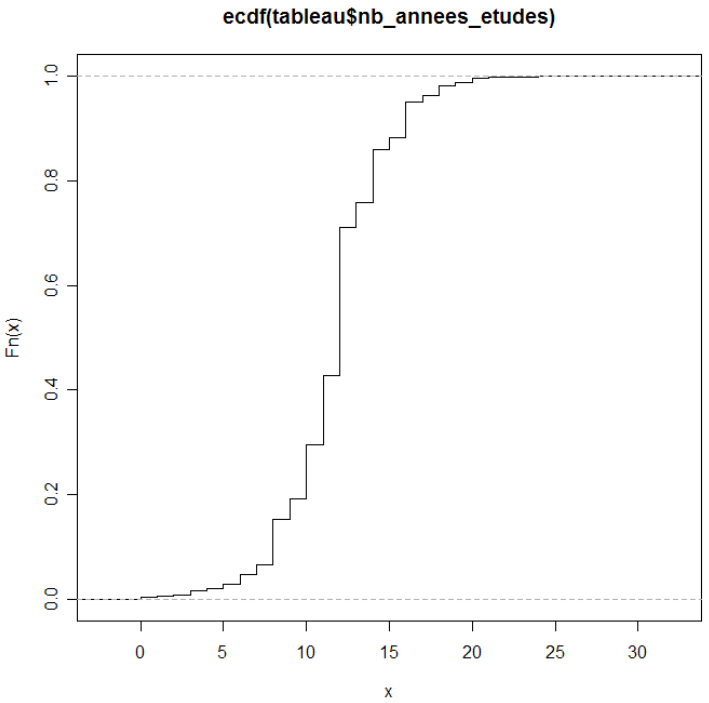
Représentation univariée
Variables quantitatives continues

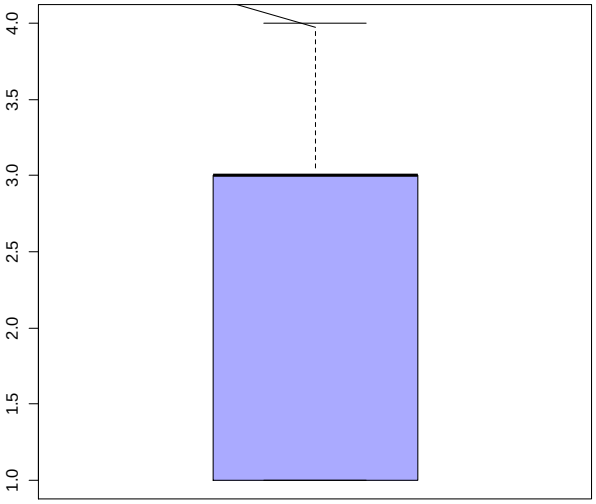
- Boxplot ou boîte à moustaches
- Représente les valeurs extrêmes (°) et les quartiles. Parfois moyenne et écart type (lozange)
- /!\ la variable est en ordonnée



4.B.2) Représentation univariée Variables quantitatives discrètes







Interdiction d'utiliser un histogramme : densité de probabilité estimée par un diagramme en bâtons

Fréquences cumulées : aspect « moins lisse »

Boxplot possible mais parfois peu lisible

4.B.3) Variables censurées à droite : la survie

- Variable observée est caractérisée par :
  - Un événement « irréversible ». Exemples : décès, prochaine hospitalisation, apparition d'un symptôme, guérison, etc.
  - Une durée d'observation :
    - Si événement survient, clôt la période d'observation.
    - Si événement ne survient pas au terme de l'observation, on ne sait pas combien de temps il aurait fallu attendre pour l'observer : on parle de « censure à droite »
  - Motifs de « censure » (durée inconnue) :
    - Patient perdu de vue avec la fin de l'étude
    - Patient encore sans événement à la fin de l'étude
- => On s'intéresse à la probabilité cumulée de survie sans événement

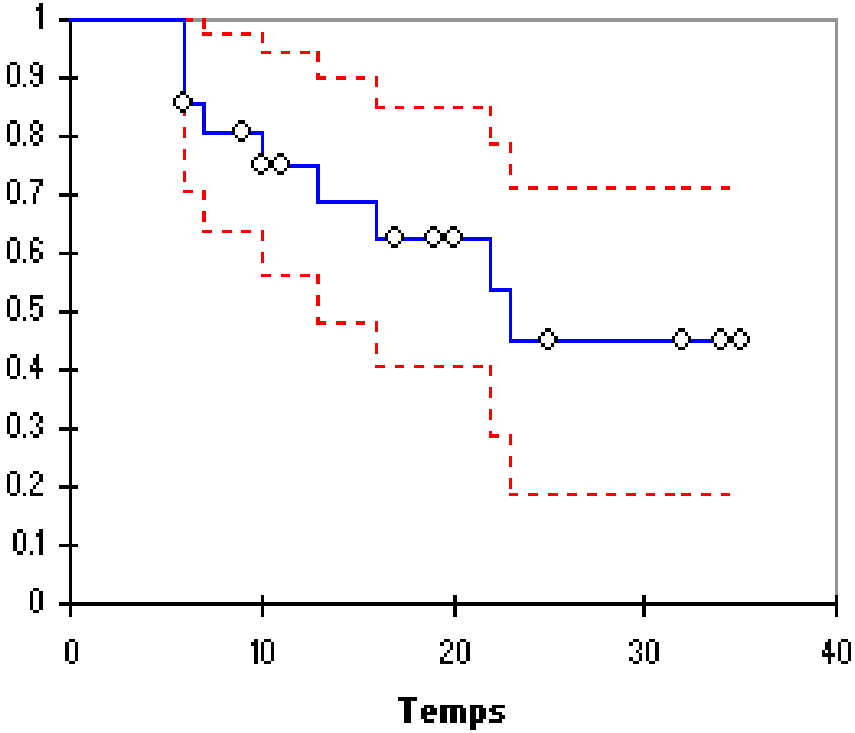
Variables censurées à droite : la survie

Courbe de survie : estime graphiquement la probabilité cumulée de survie. Part de 1 et décroît dans le temps.

Survenue de l'événement => marche d'escalier

Censures : pas d'événement dont courbe stable. Représentées par « ° ».

Eventuellement intervalle de confiance en pointillés (s'élargit au fur et à mesure que l'effectif diminue)



4.B.4) Variables catégorielles / qualitatives

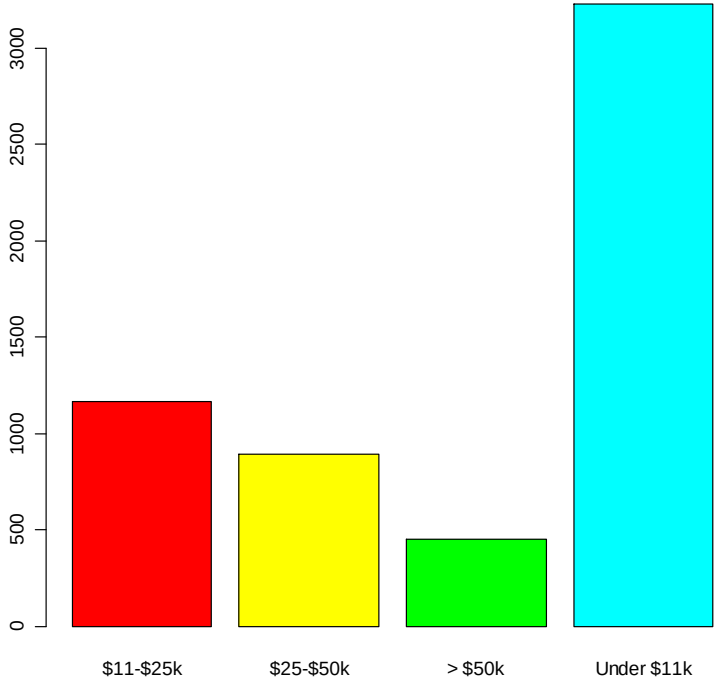


Diagramme en barres
/!\ ce n'est pas un histogramme : échelle X non quantitative, barres non jointives

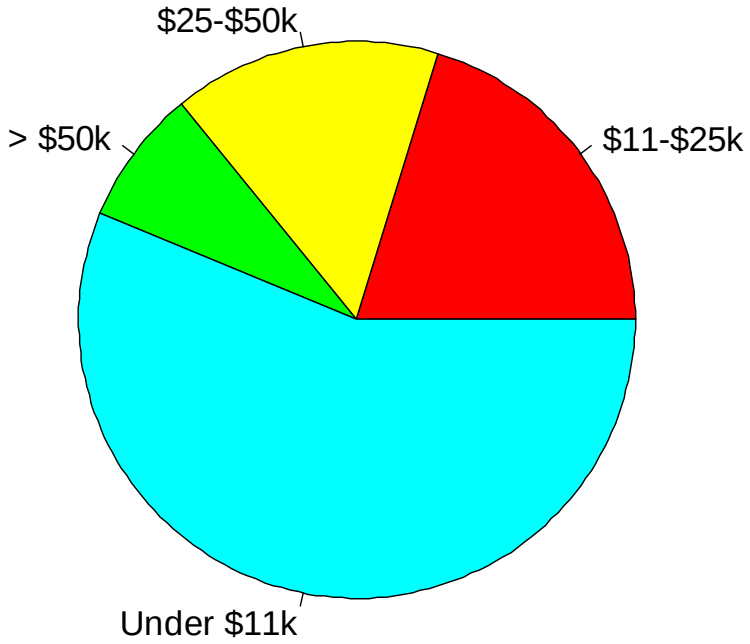


Diagramme en secteur / camembert / pie chart

4.C) Calcul de paramètres sur l'échantillon

- Variable qualitative :
  - Proportion de chaque classe
  - Classe modale
- Variable quantitative discrète :
  - Paramètres centraux : moyenne, médiane, mode
  - Paramètres de dispersion : écart-type, variance, quantiles (percentiles, quartiles)
  - Eventuellement, ajustement sur une loi binomiale, Poisson ou autre
- Variable quantitative continue :
  - Paramètres centraux : idem mais classe modale à la place du mode
  - Paramètres de dispersion : idem
  - Eventuellement, ajustement sur une loi Normale, Log-normale ou autre.
- Paramètres extrapolables sur la population : estimation et IC

4.D) L'estimation ponctuelle

Soit une variable X *(ex : taille)*

Soit θ un paramètre statistique : il reste INCONNU
(ex : taille moyenne des humains)

L'estimation consiste à approcher
la valeur de θ en utilisant
un échantillon aléatoire de taille n.

t (ou t_n ou $\hat{\theta}$) est une estimation
ponctuelle de θ.

La valeur de t fluctuera autour de θ mais sera
fonction de l'échantillon

Les intervalles de confiance

Lorsque l'estimateur *t* d'un paramètre *θ* suit une loi connue, on peut savoir dans quelle mesure *t* s'écarte du paramètre initial *θ*.

On peut parier, au risque *α* de se tromper, que l'erreur entre *t* et *θ* ne dépasse pas *a* *(a≥0)*.

$$P(|t-\theta| > a) = \alpha$$

Lorsqu'on dispose d'une estimation ponctuelle *t*, on peut calculer l'intervalle de confiance de *θ* au risque *α : $IC_{1-\alpha}$*

$$IC_{1-\alpha} = [t-a; t+a]$$

a est le demi-intervalle, c'est la précision. Lorsque *n* croit, *a* décroit.

Estimation et intervalles de confiance : illustration

- On veut connaître la cholestérolémie des adultes de 50 à 60 ans. On mesure la cholestérolémie sur un échantillon de 100 personnes, on calcule la moyenne m=1,4, les quantiles $p_{0,025}$=1,2 et $p_{0,0975}$=1,7
- Que représente l'intervalle [1,2 ; 1,7] ?

  C'est un intervalle contenant 95% des valeurs de l'échantillon
- Quelle est la cholestérolémie moyenne sur l'échantillon ?

  m=1,4
- Quelle est la cholestérolémie moyenne dans la population des adultes de 50 à 60 ans ?

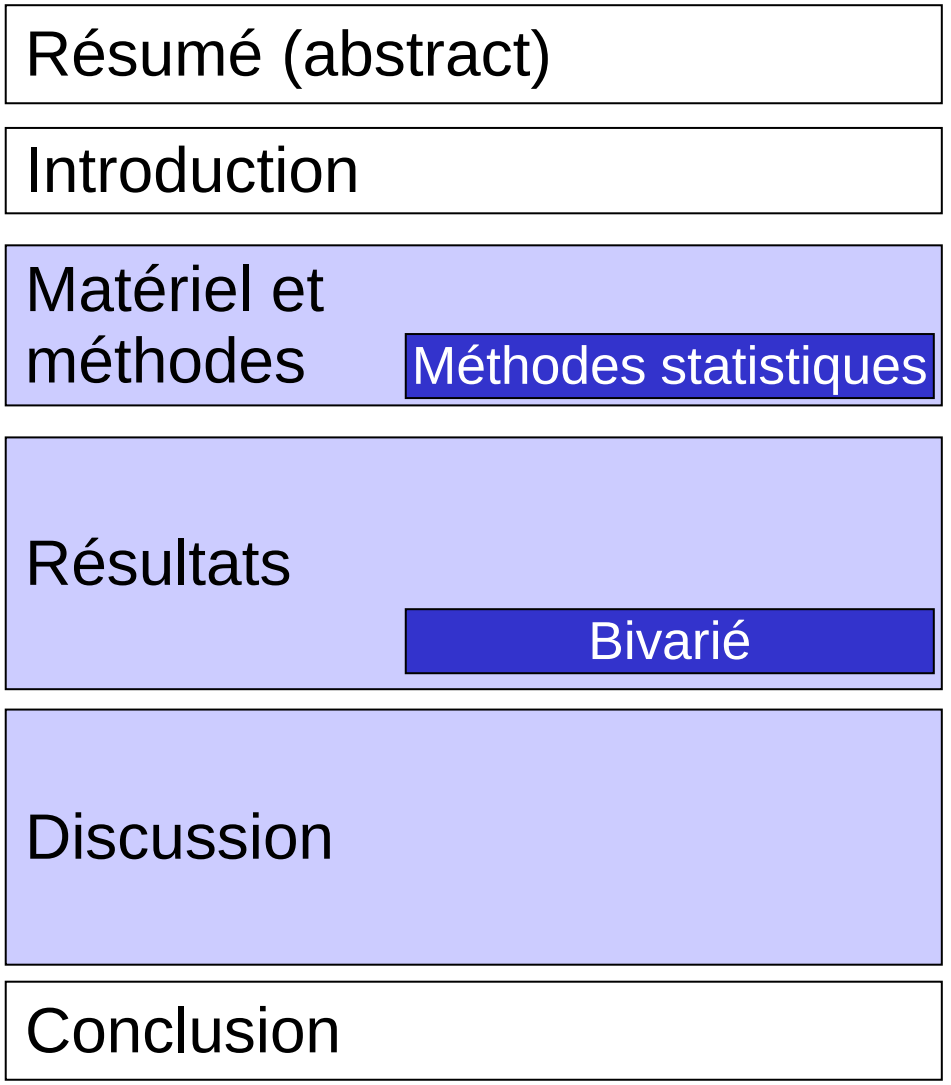
  Elle est inconnue. Cependant, si l'échantillon est représentatif, le « plus raisonnable » est de penser que μ=1,4

  Puisque les moyennes suivent un loi normale, on peut même calculer l'intervalle de confiance de μ. D'autant plus étroit que n est élevé (la moyenne a un estimateur convergent).

5. Analyses bivariées

a) Position
b) Représentations graphiques
c) Paramètres bivariés
d) Rappel sur les tests statistiques
e) Tests statistiques bivariés

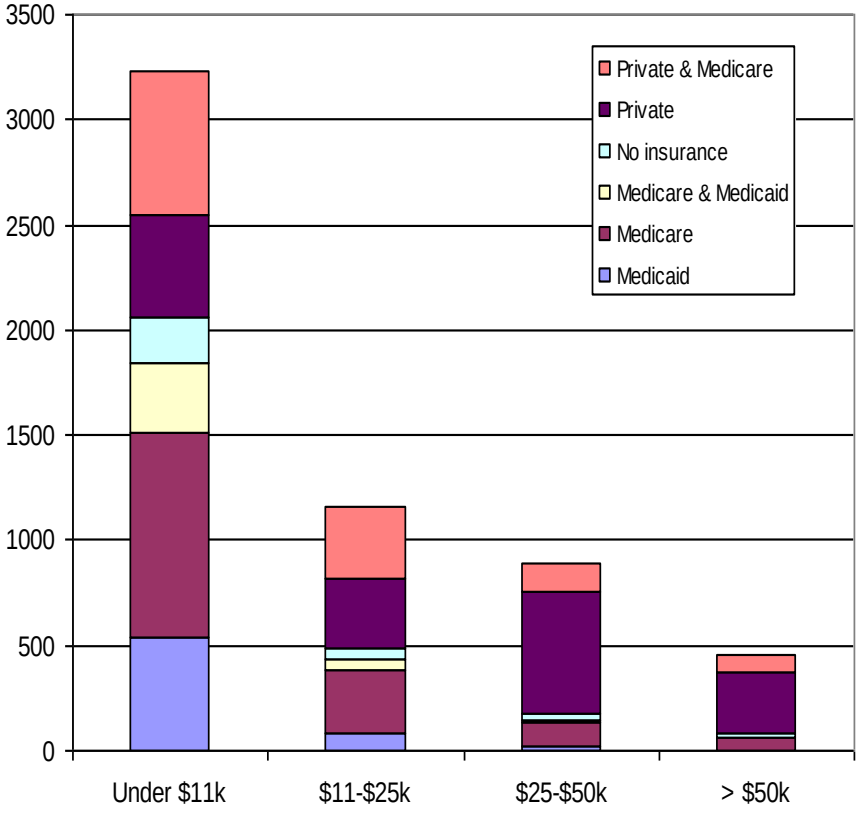
5.A) Position des analyses bivariées



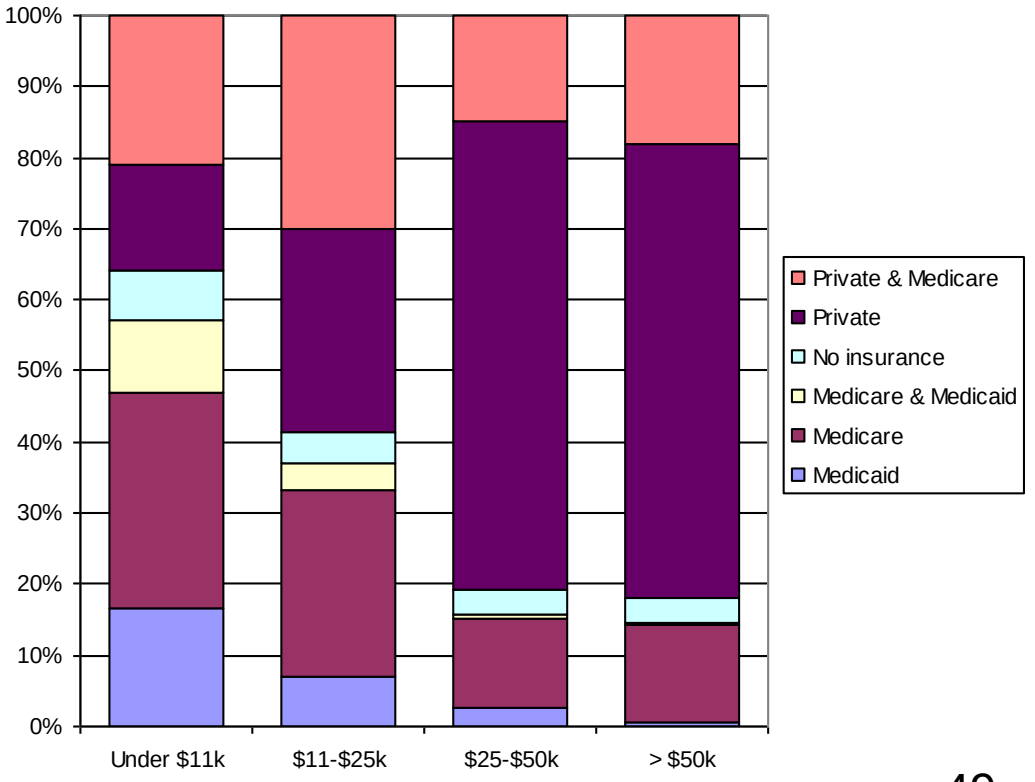
5.B.1) Graphique qualitatif * qualitatif

Effectifs représentées par des barres empilées :

Avec respect de l'effectif total…



…Ou sans (vérifier les sous-totaux)



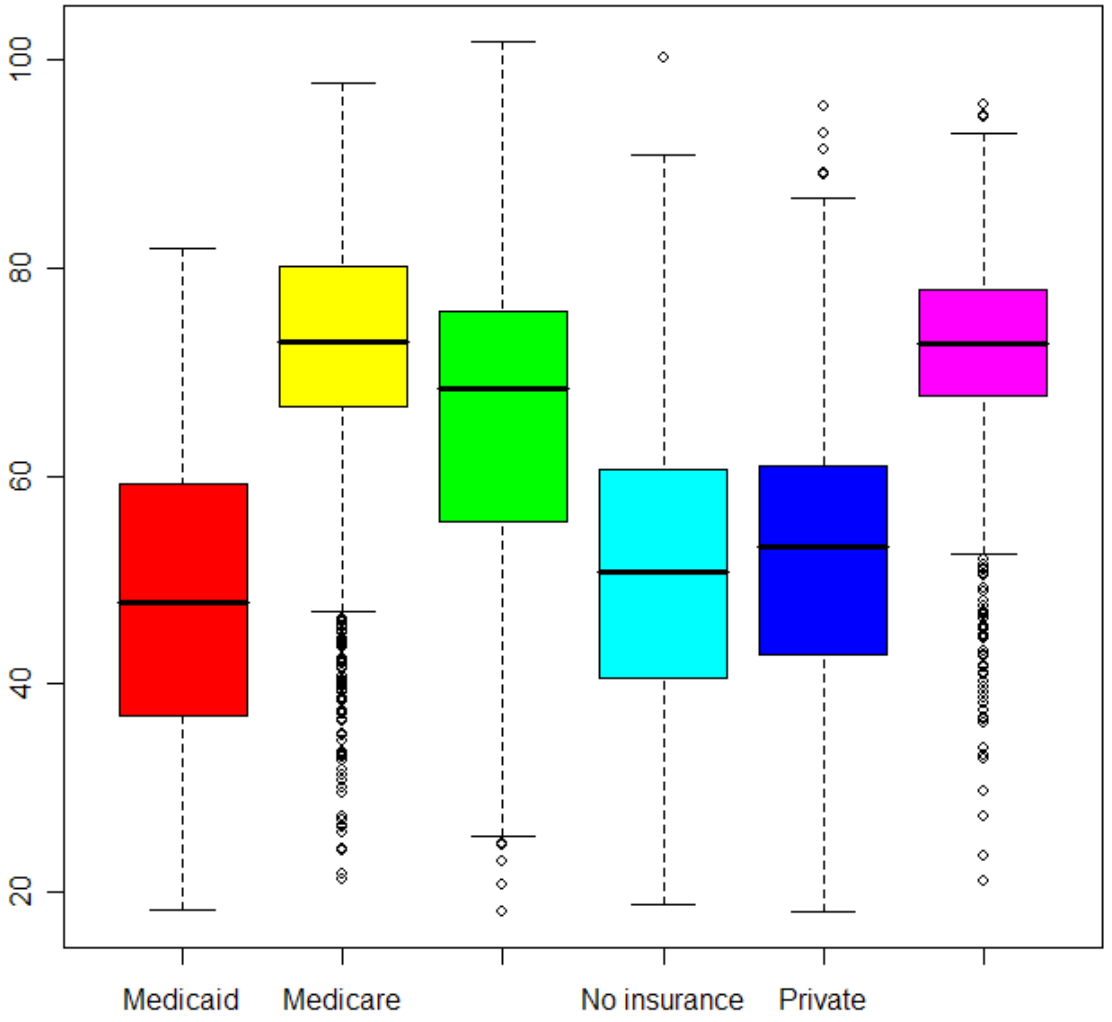
5.B.2) Graphique qualitatif * quantitatif

Représentation côte à côte d'un boxplot de la variable quantitative pour chaque modalité de la variable qualitative

Ici :

Y=âge,

X=type_assurance



Dans quelle(s) catégorie(s) trouve-t-on les patients les plus âgés ?

5.B.3) Graphique qualitatif * quantitatif

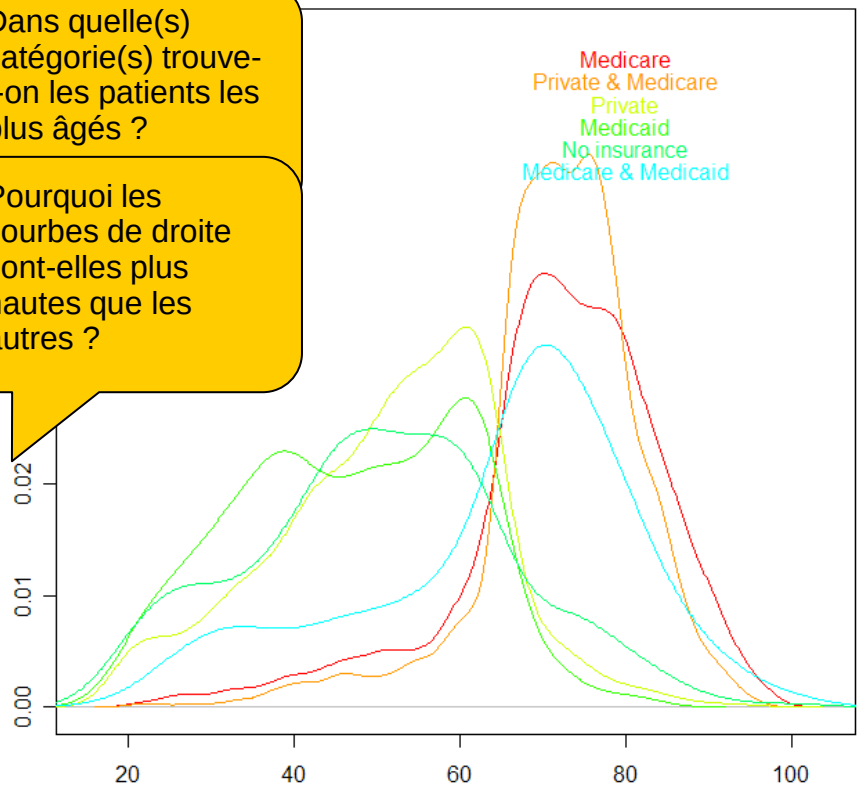
Superposition de courbes de densité de probabilité :

Y=f(âge) et X=âge

Courbes=types d'assurance

Dans quelle(s) catégorie(s) trouve-t-on les patients les plus âgés ?

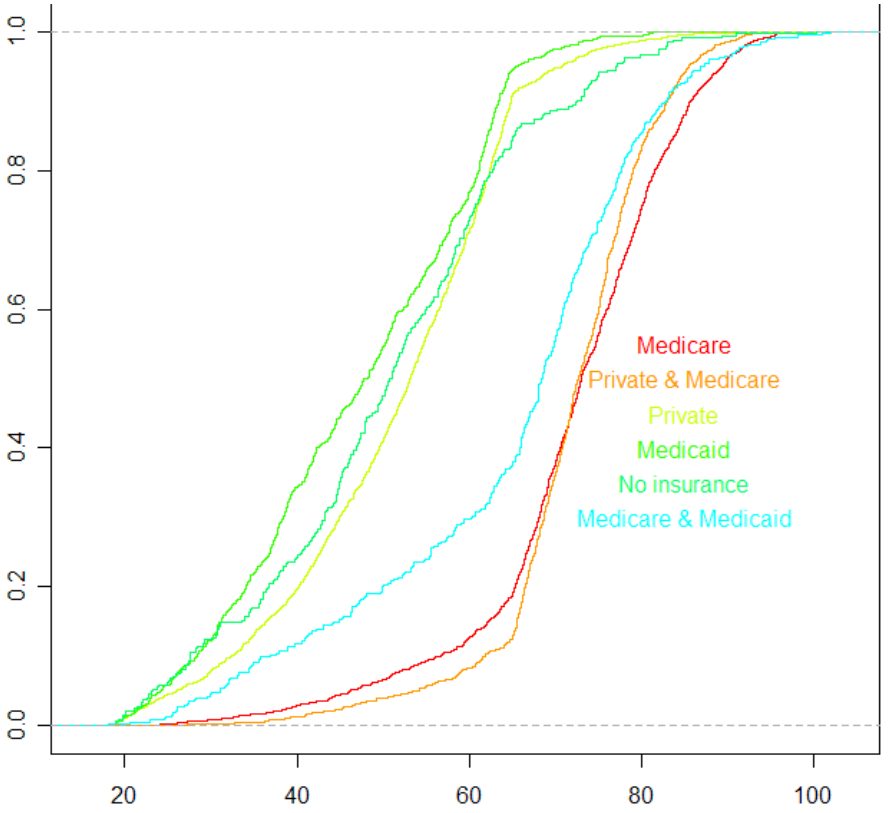
Pourquoi les courbes de droite sont-elles plus hautes que les autres ?



Superposition de courbes de fonction de répartition :

Y=F(âge) et X=âge

Courbes=types d'assurance



5.B.4) Graphique quantitatif * quantitatif

Nuage de points :

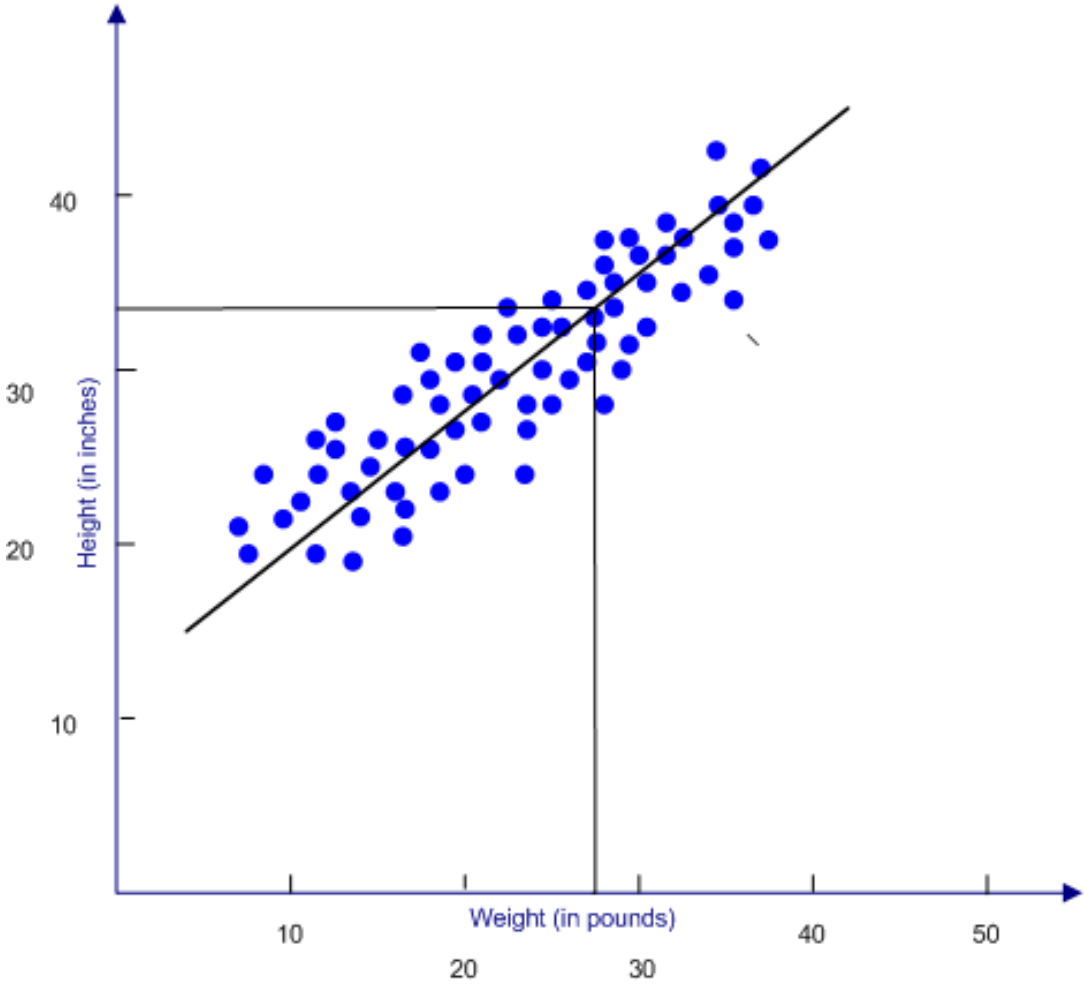
1 point = 1 individu

(précédemment, 1 courbe ou 1 rectangle par groupe)

Exemple : Y=taille, X=poids

Ajout fréquent de la droite de régression : modèle linéaire Y=a.X+b

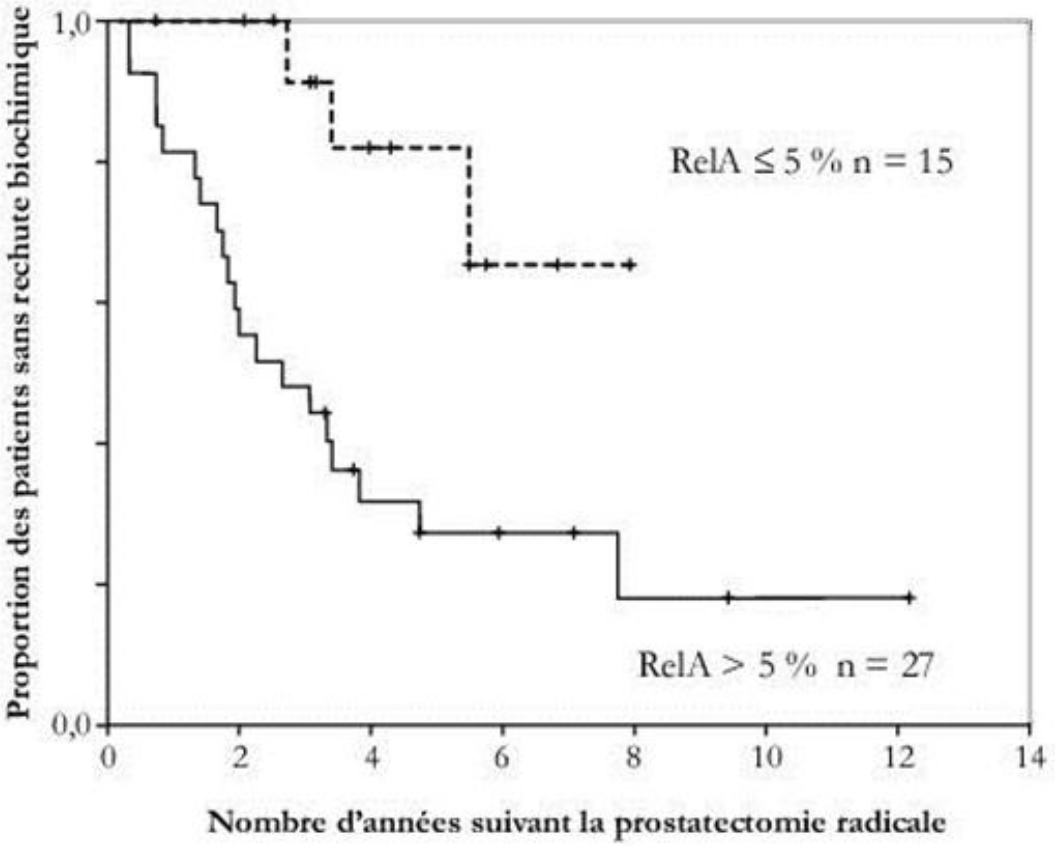
NB : il est possible de transformer l'une ou l'autre des variables en qualitative => graphiques précédents



5.B.5) Graphique qualitatif * survie

Une courbe de survie par catégorie de la variable qualitative

Exemple :
Y=probabilité cumulée de survie sans rechûte,
X=temps,
courbes=groupes



Dans quels groupe y a-t-il le plus d'événements ?

5.C) Paramètres bivariés

- Qualitatif * qualitatif :
  - Proportions conditionnelles
  - Odds ratio
  - Risque relatif (interdit dans les cas-témoins, que cohortes)
- Qualitatif * quantitatif :
  - Moyenne par sous-groupe
- Quantitatif * quantitatif :
  - Coefficient de corrélation empirique r de Pearson
    Varie entre -1 et 1
    Forte corrélation lorsque r^2 proche de 1
  - Coefficient de corrélation de Spearman, interprétation identique
  - Coefficients a et b de la régression Y=a.X+b
- Qualitatif/quantitatif * survie
  - Hazard ratio (« risque relatif dans le temps »)

5.D) Rappel sur les tests statistiques

1. Exemple introductif
2. Exemple d’un test portant sur une valeur (et non un paramètre)
3. Exemple du Test Exact de Fisher
4. Généralisation : la démarche des Tests Statistiques

5.D.1) Exemple introductif

- L’histoire du Yéti et du téléphone portable :
  - A sujet d’une observation ponctuelle

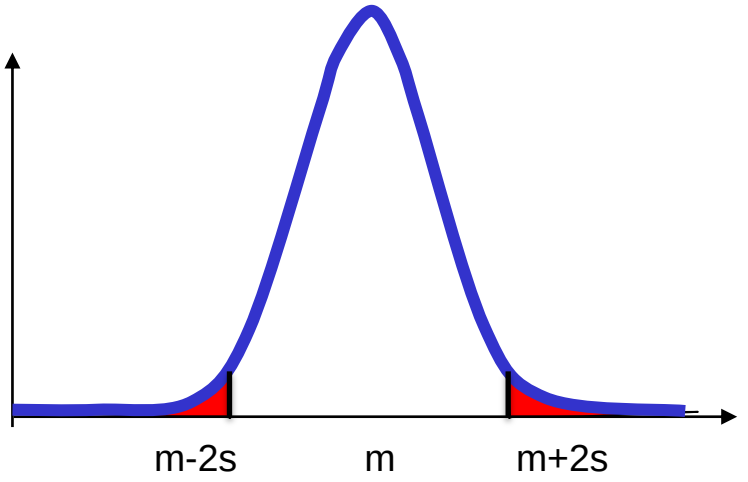
    (…)

  - Au sujet d’un groupe d’observations

    (…)

5.D.2) Exemple d'un test portant sur une valeur individuelle (et non un paramètre)

Distribution du taux de cholestérol chez les adultes sains (suit une loi normale). On définit l'intervalle qui contient 95% des valeurs des sujets sains : [1,2 ; 1,7].
Les patients en-dehors de ces bornes seront jugés malades.



Un individu présente un taux de 1,9 : est-il sain ?

NON

Un individu présente un taux de 1,5 : est-il sain ?

On ne sait pas (la question n'est pas « a-t-il un taux supérieur à 1,7 »)

Le risque de première espèce est la probabilité qu'un individu sain soit déclaré malade. Quel est-il ?

α=5% car c'est la définition ici de la zone de rejet

Le risque de deuxième espèce est la probabilité qu'un individu malade soit déclaré sain. Quel est-il ?

β est ici inconnu : il faudrait connaître la distribution chez les malades

Exemple d'un test portant sur une valeur (et non un paramètre)

- Cet exemple introductif :
  - Test de la valeur (unique) d'un attribut chez un individu
- Etudes épidémiologiques :
  - But général = mettre en évidence la non-indépendance entre deux variables (identifier un FDR d'une maladie, etc.)
  - On testera donc des paramètres statistiques calculés sur un groupe d'individus (ex : moyenne calculée sur un échantillon)

5.D.3) Exemple du test exact de Fisher

Taux de prévalence du diabète dans la population P : p=0,05

Sur mes 4 grands-parents, 2 sont diabétiques.

Est-ce théoriquement possible ? Oui

Si mes grands-parents sont issus de cette population P, quelle est la probabilité d'observer 2 diabétiques sur 4 (ou d'observer moins probable encore) ?

Loi binomiale

$P(x) = p^x \cdot (1-p)^{(4-x)} \cdot C^x_4$

x	P(x)		
0	0,81451		
1	0,17148		
2	0,01354	\	
3	0,00048	|	0,01402
4	0,00001	/	
total	*1*		

donc p=0,014

C'est la vraisemblance de l'observation sous l'hypothèse

Conclusion, au risque de première espèce α=5%, mes grands-parents sont-ils issus de la population P ?

NON, avec p=0,014 (significativité)

Mon observation est trop peu vraisemblable sous H_0, donc H_0 est considérée comme fausse.

Exemple du test exact de Fisher

Suite de l'exemple précédent :

Un seul des 4 grands-parents de mon voisin est diabétique. La vraisemblance de cette observation sous l'hypothèse est p=18,5%.

Ses grands-parents sont-ils issus de la population P ?

On ne sait pas car p>5%

Quel est le risque de première espèce β (dire qu'ils sont issus de cette population alors que ce n'est pas vrai) ?

On ne sait pas : il faudrait connaître le taux de prévalence du diabète en-dehors de cette population P.

5.E) Démarche générale des tests statistiques

- Poser l'hypothèse nulle H_0 :
  - Hypothèse la plus simple, en général « groupe1=groupe2 »
  - Cette hypothèse doit permettre de calculer une vraisemblance si elle est vraie
- Calculer la vraisemblance de l'observation sous H_0 :
  - C'est le « petit p » associé à l'observation ponctuelle
  - Vraisemblance calculable uniquement sous H_0 car il faut savoir *comment* la calculer ! Deux possibilités :
    - Soit calcul direct de probabilité (test exact de Fisher)
    - Soit calcul d'une quantité de décision, et une table de loi fournit le p (les autres tests)
- Conclure :
  - Si $p<\alpha$: observation trop peu vraisemblable sous H_0, donc H_0 est présumée fausse. Rejet de H_0 au risque α.
  - Si $p>\alpha$: l'observation est vraisemblable… donc impossibilité de conclure !! H_0 peut être vraie ou fausse.

Démarche générale des tests statistiques

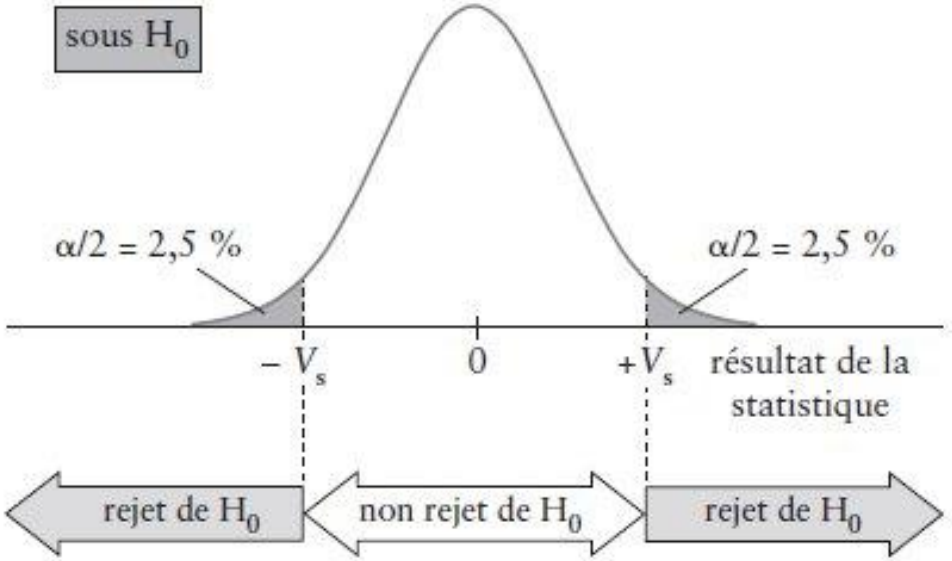
Le rejet de H0 peut se décider sur deux formulations équivalentes, même si en pratique les usages sont souvent consacrés :

- Situation 1 :
  - Sous H_0, la quantité Q vaut **0**
  - Dire « Q est différent de **0**, avec p<5% » est équivalent à dire « IC95% de Q = $[Q_1;Q_2]$ avec **0** non compris dans $[Q_1;Q_2]$ »
  - Exemples usuels : tester que t ≠ 0 (student), que $X^2 \neq 0$ (Khi²), que coefficient de corrélation ≠ 0, etc.
- Situation 2 :
  - Sous H_0, la quantité Q vaut **1**
  - Dire « Q est différent de **1**, avec p<5% » est équivalent à dire « IC95% de Q = $[Q_1;Q_2]$ avec **1** non compris dans $[Q_1;Q_2]$ »
  - Exemples usuels : tester que risque relatif≠1, que odds ratio ≠1
- => Bien comprendre la signification de H_0 pour le paramètre étudié

Démarche générale des tests statistiques

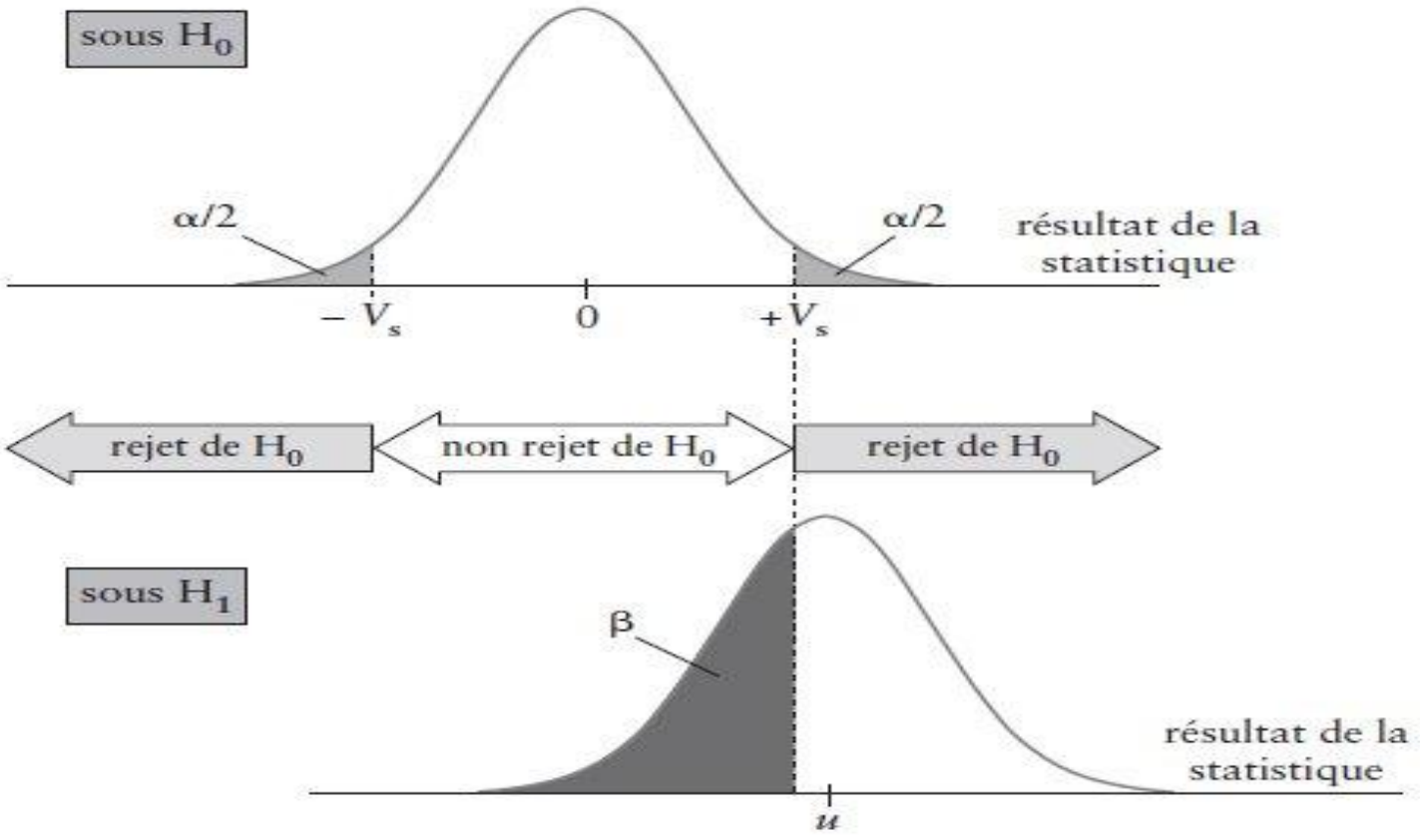
Risque de première espèce α :
C'est le risque de rejeter H_0 si H_0 est vraie.
Généralement on prend un risque α = 5% (fixé *a priori*)
Si on rejette H_0, car le test est plus grand que la valeur seuil Vs, on aura 5% de chances de se tromper.

Ci-dessous : exemple d'un test bilatéral (les tests unilatéraux sont généralement proscrits notamment car plus avantageux pour le chercheur).



Démarche générale des tests statistiques

Hypothèse H_1 = hypothèse alternative à H_0.
Parfois, la distribution de la statistique de test est connue sous H_1.
Risque de 2° espèce β = probabilité de ne pas rejeter H_0 alors que H_1 est vraie.
1 - β = puissance du test



Démarche générale des tests statistiques

	réalité inconnue	
	Ho Vraie	Ho fausse
Ho vraisemblable → non rejet de Ho	Pas d'erreur	Risque β
Ho non vraisemblable → rejet de Ho	Risque α	Pas d'erreur

Démarche générale des tests statistiques

- Presque tous les tests : conditions de validité à vérifier
- Tests paramétriques :
  - Le paramètre testé doit suivre une loi précise. Cette condition peut être atteinte si la variable initiale suit une loi précise ou si l'effectif est suffisant.
  - Ex : test de Student pour comparer les moyennes de X
    nécessite que $\overline{X} \sim T$
    atteint si $X \sim N$ ou $n>30$.
- Tests non paramétriques :
  - Ne font pas d'hypothèse de distribution des paramètres
  - Généralement plus souples, notamment utilisés pour effectifs faibles
  - Ex : test U de Mann Whitney pour comparer les moyennes

5.E) Tests statistiques bivariés

a) Comparer des effectifs/proportions

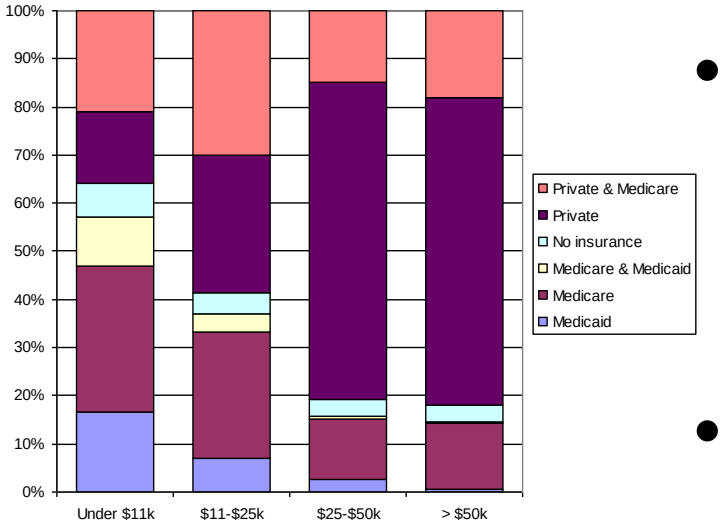
b) Comparer des moyennes non appariées

c) Comparer des moyennes appariées

d) Tester l'association entre deux variables quantitatives

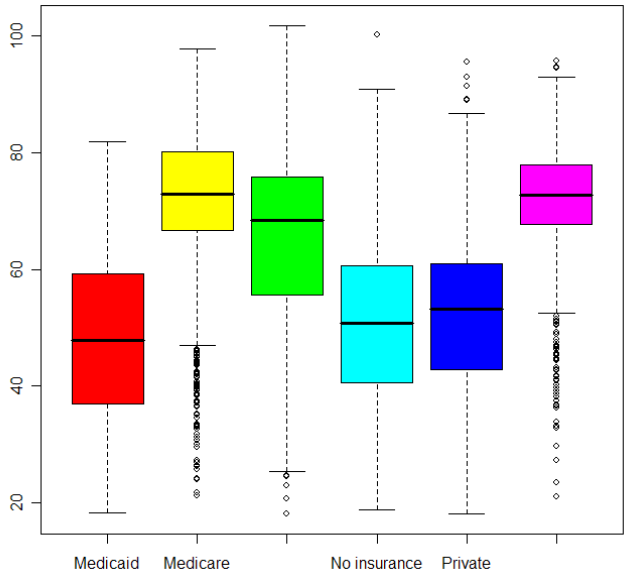
e) Association entre une variable quanti/quali et la survie

5.E.a) Comparer des proportions/effectifs



- H_0 : les variables A et B sont indépendantes
- Test du Khi² : non paramétrique, effectifs théoriques≥5
- Test exact de Fisher : non paramétrique, tous effectifs

5.E.b) Comparer des moyennes non appariées (échantillons indépendants)



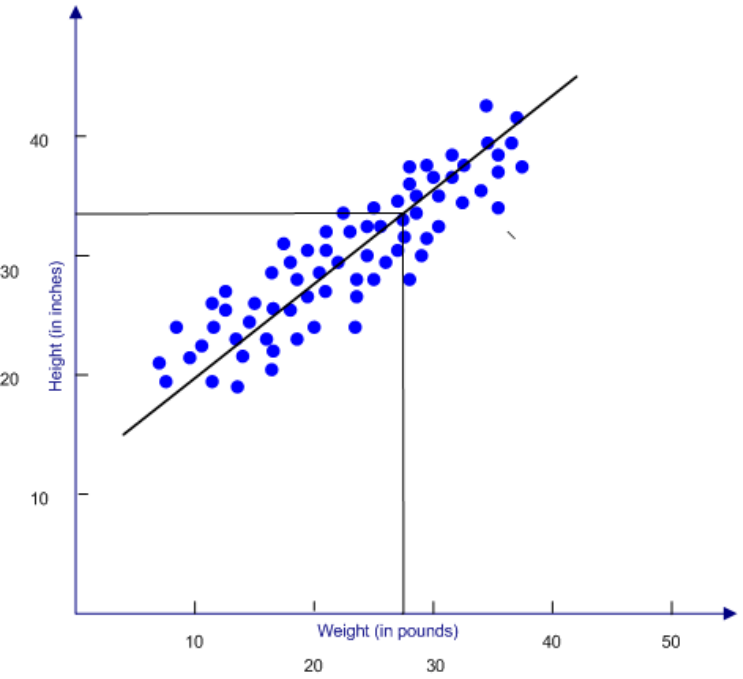
- H_0 : les variables A et X sont indépendantes (autrement dit, μ_x dans A_1 = μ_x dans A_2…)

- Paramétrique : Test t de Student (var quali à 2 catégories)
- Non paramétrique : Test U de Mann-Whitney (var quali à 2 catégories)
- Non paramétrique : ANOVA (2 catégories ou plus)

5.E.c) Comparer des moyennes appariées

- Deux variables sont appariées lorsque toute réalisation de l'une correspond à une réalisation de l'autre. Exemple typique : X_{avant} et $X_{après}$
- H_0 : « les moyennes avant et après sont identiques »… formulation courante mais inexacte. Plutôt : soit $\Delta = X_{après} - X_{avant}$, $\mu_\Delta = 0$
- Paramétrique : Test t de Student pour variables appariées
- Non paramétrique : Test t de Wilcoxon

5.E.d) Association entre deux variables quantitatives

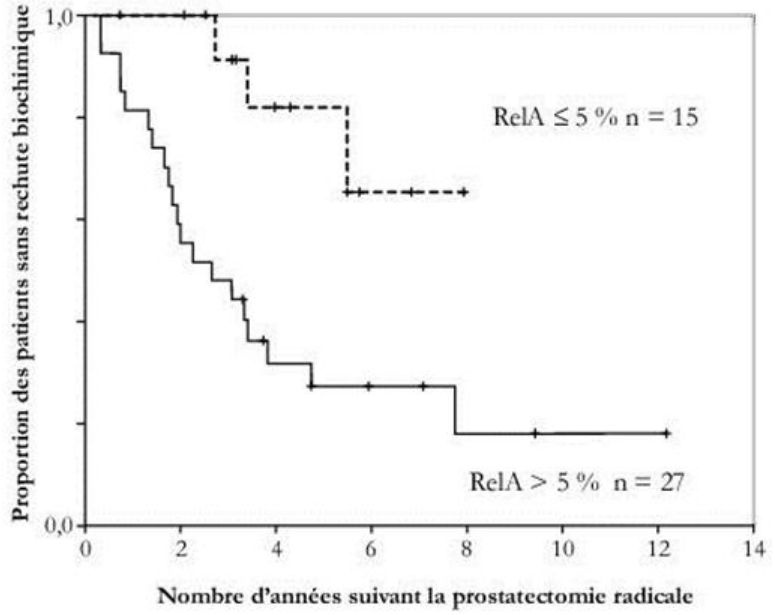


- H_0 : X et Y sont indépendantes. Le coefficient … est nul.

- Paramétrique : Test de nullité du coefficient de corrélation de Pearson
- Non paramétrique : Test de nullité du coefficient de corrélation de Spearman
- Test de nullité de la pente d'une régression linéaire Y=a.X+b

5.E.e) Association entre une variable quali/quanti et une survie

- H_0 : la survie est la même dans les deux groupes.



- Log-Rank : non paramétrique pour X qualitative
- Modèle de Cox : non paramétrique pour X qualitative ou quantitative

6) Analyses multivariées

- Modèle linéaire généralisé :
  Y est fonction de $a_0 + a_1.X_1 + \ldots + a_k.X_k$
- Différents méthodes :
  - Régression linéaire multiple pour Y quantitatif
  - Régression de Poisson pour Y « nombre de… »
  - Régression logistique pour Y binaire
  - Modèle de Cox pour Y de type Survie
- Intérêts :
  - Permet de découvrir puis tester la participation de chaque variable X dans Y… effet propre de cette variable, tenant compte en même temps de l'effet des autres variables : AJUSTEMENT
  - On peut ensuite « prédire » Y

Exemple de régression linéaire

On cherche à prédire le poids de patients en fonction de plusieurs variables. On réalise une régression linéaire multiple.

Variable	Coefficient	p
(Intercept)	60	0.00035 ***
Taille	0.49	0.0166 *
Végétarien	-3.4	0.0467 *
Sportif	-2.68	0.68

Combien a-t-il de tests ? Quelles sont les hypothèses H0 correspondantes ?

Peut-on écrire une équation qui prédit la taille ?

En moyenne, quelle est la différence de poids des végétariens ?

En moyenne, quelle est la différence de poids si la taille augmente de 10cm ?

Exemple de régression de Poisson

On cherche à prédire le nombre d'hospitalisations par an de certains malades. On réalise une régression de Poisson.

Variable	Coefficient	p
(Intercept)	3.045e+00	<2e-16 ***
Outcome2	-4.543e-01	0.0246 *
Outcome3	-2.930e-01	0.1285
Treatment2	8.717e-16	1.0000
Treatment3	4.557e-16	1.0000

Combien a-t-il de tests ? Quelles sont les hypothèses H0 correspondantes ?

Quels sont les termes significatifs à conserver ?

Aurait-on pu formuler la significativité des coefficients autrement ?

Analyses multivariées

- Modèle linéaire généralisé :
  $Y = a_0 + a_1.X_1 + \ldots + a_k.X_k$
- Interprétation générale : on teste la contribution de chaque coefficient a_k, 4 possibilités :
  - Test de l'hypothèse nulle $a_k = 0$
  - Calcul de l'intervalle de confiance de a_k, à comparer à 0
  - Test de l'hypothèse nulle $e^{-ak} = 1$
  - Calcul de l'intervalle de confiance de e^{-ak}, à comparer à 1
- Quelques exponentielles connues :
  - Régression logistique (Y binaire) : e^{-ak} =odds ratio ajusté
  - Modèle de Cox (Y survie) : e^{-ak} =hazard ratio ajusté

7) Conclure

Résumé (abstract)

Introduction

Matériel et méthodes

Résultats

Discussion

Résultats, biais

Conclusion

Conclure

Exemple : un test du Khi² réalisé entre une exposition et une maladie.

Comment interprétez-vous ce résultat ?

A cette étape, interpréter le test seulement, pas le problème médical

On observe une association statistiquement significative entre…

Vous n'avez pas le droit d'affirmer l'indépendance !!

On n'observe pas d'association statistiquement significative entre…

Que pouvez-vous conclure ?

Vous ne devez jamais vous prononcer sur une relation causale !! Evoquez toujours les biais

Pour conclure à une relation causale, il faudrait tout d'abord éliminer les biais…

Conclure

Trois biais à connaître :

- Biais de sélection
  patients non représentatifs
- Biais de classement
  patients dans le mauvais groupe : exp/non exp ou mal/non mal
  inclut biais de remémoration (exposition)
  inclut biais de mesure (exposition/maladie)
- Biais de confusion
  Toujours citer l'âge et le sexe
  Le seul biais qu'on puisse prendre en compte dans l'analyse
  Solutions : ajustement (multivarié) / appariement / stratification