STATISTIQUE BIOMEDICALE

Support de cours accompagnant le projet

Cet enseignement intitulé **Statistique Biomédicale** a pour objectif d'aborder quelques méthodes d'analyse statistique spécifiques aux données biomédicales, qui feront l'objet de quelques séances de cours.

Il s'organise principalement autour d'un **bureau d'étude statistique** au cours duquel vous analyserez des données réelles sur l'infection à VIH. Ce travail vous permettra de revoir une bonne partie de vos connaissances statistiques en réalisant une analyse statistique complète, tout en appliquant de nouvelles méthodes.

C.Chouquet, LSP-UPS, cecile.chouquet@math.univ-toulouse.fr

Table des matières

Chapitre 1

Généralités sur les données biomédicales

1.1 Introduction : quelques définitions

Epidemiologie :
Discipline scientifique qui étudie la fréquence des maladies, leur répartition dans la société, les facteurs de risque et les décès liés à cette maladie. C'est l'étude de la distribution et des déterminants des états de santé ou des évènements de santé dans une population définie et l'application de cette étude au contrôle des pbs de santé.

Recherche Clinique :
Activité médicale visant à améliorer la connaissance soit d'une maladie soit d'une thérapeutique. Elle concerne l'être humain. En pharmacologie, la recherche clinique est dominée par les études du médicament administré à l'homme dans le cadre des essais cliniques.

Recherche biomédicale :
Tous essais ou expérimentations organisés et pratiqués sur l'être humain en vue du développement des connaissances biologiques ou médicales. Elles peuvent être réalisées sur des volontaires sains ou sur des malades. Afin de protéger les personnes qui se prêtent à des recherches biomédicales, les conditions de ces recherches sont réglementées par différents textes de lois.

Les **études épidémiologiques** peuvent être classées selon deux grandes catégories :
- **d'observation** : le chercheur analyse une réalité qu'il n'a pas choisie et sur laquelle il en peut pas intervenir (cas le plus fréquent en épidémiologie).
- **expérimentales** : déterminer le facteur à étudier pour ensuite observer son effet ; souvent limitées par des contraintes éthiques.

1.2 Principales sources de données biomédicales

1.2.1 Etude transversale

• Description de la fréquence d'une maladie, de ses facteurs de risque ou de ses autres caractéristiques dans une population donnée pendant un laps de temps donné.

• Utile pour déterminer la prévalence d'une maladie à un moment déterminé et pour évaluer un nouveau test diagnostique.

• Sélection des sujets pour leur appartenance à une population, indépendamment de leur statut vis-à-vis de l'exposition et de la maladie. Biais de sélection possible (par ex : sur le lieu de travail, malades plus absents que non malades, d'où sous-estimation de la fréquence de la maladie).

• Mesure de l'exposition et de la fréquence de la maladie : étude d'une population à un instant donné (image ponctuelle) avec mesure simultanée de l'exposition et de la maladie. Recherche d'une association (et non pas d'une relation causale) entre une intervention donnée et l'issue clinique. Discussion de son caractère causal.

1.2.2 Etude Cas-Témoins

• Sélection des sujets réalisée en fonction de l'issue, de la maladie. Choix des témoins appariés selon des facteurs de confusion déjà connus.

• Etude d'observation rétrospective dans laquelle les caractéristiques des malades (les cas) sont comparées à celles de sujets indemnes de la maladie (les témoins) : comparaison de la fréquence de l'exposition entre les cas et les témoins.

• Mesure de l'exposition, mais pas de mesure de l'incidence de la maladie.

• Particulièrement adaptée pour les maladies rares ou celles qui présentent un longue période entre l'exposition et l'issue.

• Exemple : sélection d'un groupe de malades atteints de cancers et d'un groupe témoins (sans cancer) => Etude des antécédents dans chaque groupe => Comparaison des antécédents et conclusion.

1.2.3 Etude de cohorte ou de suivi

• Sélection des sujets réalisée en fonction de l'exposition et non pas de l'issue.

• Etude d'observation dans laquelle un groupe de sujets exposés (à des facteurs de risque d'une maladie ou à un traitement particulier) est suivi pendant une période donnée et comparé à un groupe de sujets non exposés (groupe contrôle).

• Suivi des sujets jusqu'à l'apparition de l'issue recherchée, donc pouvant exiger un délai très long entre le début de l'étude et l'obtention des premiers résultats. Mesure directe du nombre de cas de maladie survenus spontanément avec le temps dans chacune des deux populations, c.a.d. de l'incidence de la maladie.

• Exemple : sélection d'un groupe de fumeurs (exposés) et d'un groupe de non-fumeurs (non-exposés) => Suivi des sujets et comparaison des issues.

1.2.4 Essai contrôlé randomisé

• Etude expérimentale, où les patients éligibles, sélectionnés pour une intervention thérapeutique, sont répartis de manière aléatoire en deux groupes : l'un recevant le nouveau traitement, le seond en général le placebo ou l'ancien traitement.

• Répartition des sujets au hasard dans chaque bras de traitement, ayant pour but d'assurer que les patients répartis dans les deux groupes de l'essai sont rigoureusement semblables en tous points, excepté en ce qui concerne l'intervention. Toute différence observée entre les issues des deux groupes sera probablement due à l'intervention et non à des facteurs potentiels d'interférence.

• Réalisation de l'étude en aveugle ou en double aveugle de manière à écarter tout biais éventuel.

• Coût et durée de ce type d'études contraignants ; problèmes d'éthique à prendre en considération.

1.3 Spécificités des données

1.3.1 Situation dans le temps : transversales ou longitudinales

Lorsque les informations concernant le facteur et la maladie sont recueillies à la même date, l'étude est dite **transversale**.

Lorsque les informations concernant le facteur et/ou la maldie s'étalent au cours du temps, l'étude est **longitudinale**. En médecine, la maladie peut être souvent considérée comme un processus évolutif, dont la connaissance nécessite un suivi longitudinal. Un des avantages des études longitudinales est de permettre la détermination de la séquence des évènements.

1.3.2 Mesures prospectives et rétrospectives

Cette notion de mesure prospective ou rétrospective ne fait pas référence au temps mais à la façon dont les sujets sont sélectionnés et dont les informations sont récoltées.

Dans une étude **prospective**, le chercheur sélectionne des sujets qui seront exposés à un risque ou à un facteur d'intervention, et chez qui on pourra observer un évènement. Une enquête prospective est une enquête longitudinale dans laquelle on connaît à un instant les expositions des sujets au facteur, puis la survenue de la maladie à laquelle on s'intéresse. C'est le cas des études de cohortes.

Dans une étude **rétrospective**, les sujets sont sélectionnés parce qu'ils ont déjà vécus l'évènement. C'est une étude longitudinale dans laquelle on interroge les patients, malades et non malades sur leur exposition passée au facteur. C'est le cas des études cas témoins.

1.3.3 Données longitudinales, données censurées

Les données longitudinales sont un cas particulier de **mesures répétées** dans le temps sur un ensemble d'unités. Ces mesures peuvent être faites à intervalles réguliers sur un même nombre d'individus ; on dispose alors du même nombre de mesures par individu.

Dans les études médicales traditionnelles dont l'objectif est d'étudier la survie des patients, on peut observer en général la date exacte d'un évènement ou la date de censure, c.a.d. la dernière date à laquelle l'individu n'avait pas encore eu l'évènement. Ce type de données est disponible quand il est facile de dater l'évènement tel que le décès, ou si les individus sont suivis très fréquemment de telle façon que l'on peut connaître la date exacte d'un évènement. On dispose alors de **données censurées à droite**.

Par ailleurs, dans les cohortes de sujets, les données longitudinales sur l'histoire d'une maladie sont souvent incomplètes au moins pour les 3 raisons suivantes :

- un sujet ne peut être observé sur toute la durée de sa maladie ;
- un sujet est vu par intermittence à des visites de suivi (à intervalles plus ou moins réguliers) au cours desquelles on observe si un évènement s'est produit ou non depuis la dernière visite, mais l'information sur les périodes entre deux visites est manquante ou incomplète ;
- le moment exact de survenue de la maladie ou d'autres évènements est inconnu.

Au lieu d'observer une date d'évènement, on observe un intervalle de temps pendant lequel l'évènement s'est réalisé. De telles données sont appelées **données censurées par intervalle**.

1.4 Indicateurs statistiques en recherche biomédicale

1.4.1 Fréquence d'une maladie : prévalence et incidence

• L'**incidence** mesure la fréquence de nouveaux cas apparaissant au cours d'une unité de temps. On peut également utiliser la notion d'incidence cumulée (par exemple, depuis le début de l'épidémie pour

les maladies infectieuses). Elle permet d'évaluer la *taille* de l'épidémie.

• La **prévalence** mesure la fréquence des cas présents (aussi bien les nouveaux cas que les anciens) à un moment donné dans la population à laquelle on s'intéresse.

• Incidence et prévalence s'expriment en nombre de cas (absolues) ou en nombre de cas rapportés à une population (relatives).

• **Relation entre incidence et prévalence** : Dans le cas de maladies dont la fréquence ne change pas au cours du temps, si P et I désignent la prévalence et l'incidence, et si D désigne la durée de la maladie, on a approximativement $P = I \times D$.

1.4.2 Mesures d'association : risque relatif et odds-ratio

L'objectif des études transversales, des cohortes ou des études cas-témoins est d'évaluer les conséquences de la présence ou de l'absence d'un facteur de risque : effet positif/protecteur ou effet négatif/délétère. Un même facteur peut avoir un effet positif sur certaines maladies et un effet délétère sur d'autres.

Pour analyser la relation facteur/maladie, les données sont résumées sous la forme d'un tableau de contingence (en colonnes : malades/Non malades ; en lignes : Exposés/Non exposés). On désigne par :
- M la présence de la maladie chez un patient, et NM son absence ;
- F la présence de l'exposition ou du facteur de risque chez un patient, et NF son absence.

Pour quantifier cette relation entre la survenue de la maladie et l'exposition, on dispose des indicateurs suivants :

• **Risques absolus** : c'est la probabilité pour qu' individu ait la maladie ; on peut la calculer :
- chez les exposés $Pr(M/F) = R1$,
- chez les non-exposés $Pr(M/NF) = R0$,
- dans la population totale $Pr(M)$.

• **Risque Relatif** : c'est le rapport de la probabilité de maladie chez les sujets exposés à celle chez les sujets non-exposés :

$$RR = Pr(M/F)/Pr(M/NF) = R1/R0$$

On l'interprète comme le facteur par lequel le risque de maladie est multiplié en présence de l'exposition. Le calcul du risque relatif est possible uniquement dans les études prospectives qui permettent de connaître les risques absolus de maladie dans les différents catégories de sujets exposés à la maladie.

• **Odds Ratio** : il peut être estimé dans tous les types d'études, et est équivalent au risque relatif dans grand nombre de situations (en particulier si la fréquence de la maladie est assez faible) :

$$OR = \frac{R1/(1-R1)}{R0/(1-R0)} = \frac{Pr(M/F)}{Pr(M/NF)} \times \frac{Pr(NM/NF)}{Pr(NM/F)}$$

• **Risque attribuable** : on peut calculer le risque attribuable à un facteur chez les exposés comme la différence entre la fréquence de la maladie chez les exposés et la fréquence de la maladie chez les non-exposés :

$$RA_e = Pr(M/F) - Pr(M/NF)$$

La fraction du risque attribuable ches les exposés est le rapport du risque attribuable chez les exposés sur le risque de maladie chez les exposés :

$$FRA_e = \frac{Pr(M/F) - Pr(M/NF)}{Pr(M/F)} = \frac{RR-1}{RR}$$

On peut également calculer ces deux quantités dans la population. Le risque attribuable à un facteur dans la population est la différence entre le risque de maladie dans la population et le risque de maladie qu'il y aurait en l'absence d'exposition au facteur :

$$RA_p = Pr(M) - PR(M/NF)$$

La fraction de risque attribuable à un facteur dans la population est la proportion maximale de cas qui pourraient être attribués au facteur :

$$FRA_p = \frac{RA_p}{Pr(M)} = \frac{f(RR-1)}{f(RR-1)+1} = g\frac{RR-1}{RR}$$

où f est la proportion de sujets exposés dans la population, et g est la proportion de sujets exposés ches les malades. On peut aussi interpréter cette fraction comme la proportion maximale de cas que l'on pourrait éviter si on supprimait l'exposition au facteur supposé causal.

1.4.3 Evaluation diagnostique : sensibilité, spécificité, valeurs prédictives

Dans le cadre du diagnostic (face à un patient) ou du dépistage (face à une population), la valeur d'un test clinique ou biologique en tant qu'outil diagnostique est quantifiée par sa sensibilité, sa spécificité, et ses valeurs prédictives positives et négatives. On suppose que tout sujet peut être classé soit comme atteint de la maladie, soit comme non atteint. On s'intéresse au cas où le test diagnostique utilisé fournit une valeur quantitative qui est classée en positif ou négatif selon qu'elle dépasse ou non un seuil fixé.

On note $p = Pr(M)$ la prévalence de la maladie. $D+$ désigne le fait qu'un patient soit trouvé positif par le test diagnostique, et $D-$ qu'il soit trouvé négatif. On dispose des indicateurs suivants pour juger de la qualité de ce test diagnostique :

• La **sensibilité** (Se) d'un test diagnostique est la proportion de malades classés positifs par le test :

$$Se = Pr(D+/M)$$

• La **Spécificité** (Sp) est la proportion de non-malades classés négatifs par le test :

$$Sp = Pr(D-/NM)$$

• La **valeur prédictive positive** d'un test est le pourcentage de malades parmi les patients trouvés positifs par ce test :

$$VPP = Pr(M/D+) = \frac{pSe}{pSe + (1-p)(1-Sp)}$$

• La **valeur prédictive négative** d'un test est le pourcentage de non malades par les patients qui sont classés négatifs par ce test :

$$VPN = Pr(NM/D-) = \frac{(1-p)Sp}{(1-p)Sp + p(1-Se)}$$

1.5 Applications

5.1 Une étude portant sur l'efficacité du vaccin contre la grippe a donné les résultats suivants :

	Malades	Non-Malades
Vaccinés	20	220
Non-vaccinés	80	140

Evaluer l'efficacité du vaccin par le calcul du risque relatif.

5.2 Dans une population, un sujet sur 100 a été exposé à un facteur F. Un sujet par millier développe la maladie M. Chez ceux qui développent la maladie, 80% ont été exposés à F. Caractériser les risques absolus, relatifs, et les différents risques et fractions attribuables.

5.3 La spondylarthrite ankylosante est une maladie rhumatismale inflammatoire. L'antigène HLA B27 a été identifié comme marqueur de cette maladie. Au vu des résultats donnés dans le tableau ci-dessous, évaluer les qualités diagnostiques de cet antigène (Se, Sp, VPP et VPN).

Antigène HLA B27	Malades	Non-Malades
Positif	90	510
Négatif	10	9390

5.4 Un test diagnostique T de l'infection urinaire a été évalué dans un service d'urologie où la prévalence de cette pathologie est de 20%. La sensibilité de ce test est de 95%, sa spécificité de 90%. On utilise maintenant ce test diagnostique dans un groupe de la population où la prévalence est de 1%.

1. Les valeurs de sensibilité et de spécificité seront-elles modifiées ?
2. Calculer la valeur prédictive positive de ce test.

5.5 En prenant comme critère diagnostique de l'infarctus du myocarde l'élévation des enzymes cardiaques (CPK) au dessus de 80 UI, quelle est la prévalence de l'infarctus du myocarde dans une population où 20% des individus remplissent ce critère, qui a par ailleurs une sensibilité de 93% et une spécificité de 88% ?

Chapitre 2

Analyse de données de survie

2.1 Introduction

• Objectif : s'exprimer sur la "chance" de survie de patients présentant une pathologie particulière.
Moyen : quantifier la probabilité qu'ont ces patients de survivre au moins un certain temps à compter d'un instant de référence.

• L'analyse des durées de survie englobe également l'**analyse des délais de survenue d'un évènement**, autre que le décés, tel que la survenue de complications, de rechutes, de disparitions de symptômes, ... ou des évènements plus complexes.
Les méthodes expliquées dans ce chapitre s'appliquent à tout type d'évèvement. On parlera en général de décès, de survie pour simplifier le discours.

• L'élément majeur de l'étude des phénomènes de survenue d'évènements est la **fonction de survie**. Cette fonction, notée $S(t)$, dépendant de t est définie comme :

$$S(t) = Pr(\text{délai de survenue de l'évènement d'intérêt à compter de l'instant de référence} > \text{t})$$

Sa représentation graphique s'appelle **courbe de survie**.

•Définitions et notations :

- $S(t)$ représente la **probabilité pour qu'un patient soit encore vivant après un délai** t, ou encore la proportion de survivants après un délai t.
- La durée de vie d'un patient est une caractéristique variable d'un patient à l'autre ; c'est une v.a. que l'on notera T. $S(t)$ est donc la **probabilité pour que** T **soit supérieure à** t :

  $$S(t) = Pr(T > t) = 1 - F(t)$$

  où F est la fonction de répartition de la durée T.
- La fonction de survie permet de calculer la **probabilité que le décès survienne après un délai** t_1 **et avant un délai** t_2 $(t_2 > t_1)$:

  $$Pr(T \in]t_1, t_2]) = F(t_2) - F(t_1) = S(t_1) - S(t_2)$$

- La fonction de survie donne une nouvelle information : la probabilité de survivre après un délai t sachant que l'on est survivant après un délai τ $(\tau < t)$, que l'on notera $S(t|\tau)$:

  $$S(t|\tau) = Pr(T > t|T > \tau) = \frac{Pr(T > t)}{Pr(T > \tau)} = \frac{S(t)}{S(\tau)}$$

- Le **risque instantané de décès** au délai t est la probabilité de décéder juste après t, au cours d'un intervalle de durée Δt (tendant vers 0) :

  $$\lambda(t) = -\frac{dS}{dt}(t)/S(t)$$

— La **date d'origine**, variable selon les sujets, est la date à laquelle le sujet peut être considéré comme entrant dans l'étude et à partir de laquelle on comptera les délais le concernant. La **date de point** est la date à laquelle l'étude se termine (fin du recueil des informations), elle est la même pour tous les patients. Toutefois, le suivi d'un patient aura pu se terminer avant cette date si le patient est **perdu de vue** ou s'il a déclaré l'évènement d'intérêt à une date ultérieure.
— Dans le cas où le patient n'a pas déclaré l'évènement à sa dernière visite dans l'étude, on parle d'**information censurée**.
— Pour un patient donné, le délai entre la date d'origine et la date de ses dernières nouvelles s'appelle **temps de participation** ou **durée de suivi** du patient.
— Le **recul** pour un patient est le délai entre la date d'origine et la date de point.

Lorsque l'on s'intéresse à la survenue d'évènements, on rencontre l'un ou l'autre des problèmes suivants : estimer une fonction de survie et/ou évaluer l'impact d'une action de survie en comparant deux (ou des) fonctions de survie.

2.2 Méthode actuarielle

2.2.1 Informations disponibles

L'analyse de données de survie est basée sur les données d'études cliniques consistant à inclure n patients au cours du temps et à les suivre pour observer un évènement particulier. Pour un patient i ($i = 1, ..., n$), on prend en considération :
— sa durée de suivi dans l'étude (sa date d'origine et sa date des dernières nouvelles),
— son statut vis-à-vis de l'évènement (présent ou non) à la date des dernières nouvelles.

2.2.2 Principe de la méthode

Cette méthode est utilisée pour des études de grande taille. Le principe de la méthode est d'estimer la fonction de survie en des temps définis à l'avance. On note 0, b_1, b_2, ..., b_n les différents temps retenus. On utilise la formule :

$$S(b_j) = S(b_{j-1}) \times S(b_j|b_{j-1})$$

En d'autres termes, la probabilité d'être survivant au temps b_j, c'est la probabilité pour un patient d'être survivant au temps b_j sachant qu'il était survivant au temps précédent b_{j-1}, par la probabilité d'être survivant au temps b_{j-1}. On sait que tous les patients sont vivants à la date d'origine : $S(0) = 1$.

L'estimation de $S(b_j|b_{j-1})$ est donnée par :

$$\widehat{S}(b_j|b_{j-1}) = 1 - \frac{D_j}{N_j - C_j/2}$$

où :
— N_j est le nombre de patients que l'on sait vivants au temps b_{j-1},
— D_j est le nombre de patients décédés dans l'intervalle $]b_{j-1}, b_j]$,
— C_j est le nombre de sujets censurés dans l'intervalle $]b_{j-1}, b_j]$.
On a la relation : $N_{j+1} = N_j - D_j - C_j$. Le dénominateur $(N_j - C_j/2)$ représente le nombre moyen de patients à risque de décéder sur l'intervalle $]b_{j-1}, b_j]$.

La fonction de survie ainsi estimée par la méthode actuarielle peut être représentée graphiquement : en abscisse le temps et en ordonnée l'estimation de la survie. Pour chaque temps retenu b_j, on trace l'estimation de la survie correspondante. Les valeurs de la fonction de survie à d'autres temps s'obtiennent par interpolation linéaire (en d'autres termes, on trace une droite entre chacun des points représentés).

2.3 Méthode de Kaplan-Meïer

2.3.1 Le contexte

Cette méthode peut être utilisée dans toutes les circonstances, mais est plus souvent employée dans le cadre d'études de faibles effectifs. Le principe de base de cette méthode est le même que celui de la méthode actuarielle, à deux différences près :

— la fonction de survie est supposée constante entre deux instants de décés observés,
— la fonction de survie est estimée à chaque instant de décès observé.

Comme précédemment, pour un patient i, on dispose des données suivantes :

— t_i, son temps de participation dans l'étude,
— son statut vis-à-vis de l'évènement (déclaré ou non) à la date des dernières nouvelles.

On note habituellement t_i la durée de suivi d'un patient décédé au temps t_i, et t_i^* la durée de suivi d'un patient censuré au temps t_i (perdu de vue ou connu vivant à la date de point). Pour l'application de cette méthode, les patients sont classés par ordre croissant de t_i et/ou t_i^*, et numérotés de 1 à n.

2.3.2 Le principe

Estimateur de Kaplan-Meier :
On cherche à estimer la fonction de survie S aux seuls temps de décès observés, par la formule :

$$S(t_i) = S(t_{i-1})S(t_i|t_{i-1})$$

L'estimation proposée pour $S(t_i|t_{i-1})$ est donnée par :

$$\widehat{S}(t_i|t_{i-1}) = 1 - \frac{D_i}{N_i - C_i} = 1 - \frac{\text{nombre de décés à } t_i}{\text{nombre de patients à risque à } t_i}$$

où

— D_i est le nombre de décés observé au temps t_i,
— C_i est le nombre de patients censurés entre t_{i-1} et t_i,
— N_i est le nombre de patients connus vivants juste après t_{i-1}.

Le dénominateur $N_i - C_i$ représente le nombre de patients susceptibles ou "à risque" de décéder au temps t_i. On notera qu'un patient censuré au temps t_i est à risque au temps t_i.

Représentation graphique :
On peut représenter l'évolution de la survie en fonction du temps avec, en ordonnée, l'estimation de la fonction de survie entre 1 et 0, et en abscisse, la durée de suivi. C'est une fonction décroissante au cours du temps, représentée en marches d'escalier : on suppose que la fonction de survie est constante entre les temps t_i et t_{i+1} et qu'elle vaut $\widehat{S}(t_i)$ sur cet intervalle.

Médiane de survie :
La médiane de survie correspond au temps pour lequel la fonction de survie estimée est égale à 0.5.

Ecart-type de l'estimateur de Kaplan-Meier :
L'écart-type de la probabilité estimée de survie $\widehat{S}$ au temps t_i est donné par :

$$ETS(t_i) = \widehat{S}(t_i)\sqrt{\sum_{t_i<t} \frac{D_i}{(N_i - C_i)(N_i - C_i - D_i)}}$$

Les bornes de l'intervalle de confiance à 95% de la probabilité de survie au temps t s'écrivent (sous l'hypothèse que S(t) suit asymptotiquement une distribution gaussienne) :

$$S(t) - 1.96 \times ETS(t)\ ;\ S(t) + 1.96 \times ETS(t)$$

On peut faire figurer les bornes de ces intervalles à différents temps sur la courbe de survie. Il est également souhaitable de compléter la représentation graphique par l'effectif des sujets encore à risque de décès pour quelques instants, afin de pouvoir juger rapidement de la précision de la courbe.

Remarque : A noter qu'il existe d'autres formulations plus appropriées pour le calcul de l'écart-type, proposées par les logiciels de statistique.

2.3.3 Exemple

On dispose des valeurs de t_i et t_i^* suivantes :

$$6; 6; 6; 6.1*; 7; 9*; 10; 10.1*; 11*; 13; 16; 17*; 19*; 20*; 22; 23; 25*; 32*; 32*; 34*; 35*$$

2.4 Modèles paramétriques de survie

Un modèle paramétrique de survie est un modèle dans lequel la fonction de risque instantané $\lambda(t)$ est une fonction mathématique dépendant d'un ou de plusieurs paramètres. Dans un premier temps, nous décrivons le **modèle exponentiel** qui suppose que la fonction de risque $\lambda(t)$ est constante au cours du temps :

$$\lambda(t) = \lambda_0$$

On l'appelle modèle exponentiel parce que la fonction de survie est exponentielle :

$$S(t) = exp(-\lambda_0 t)$$

La distribution de la v.a. T, la durée de survie, est définie par :
- sa fonction de répartition : $F(t) = 1 - S(t) = 1 - exp(-\lambda_0 t)$
- sa densité de probabilité : $f(t) = \lambda_0\ exp(-\lambda_0 t)$

L'estimateur du paramètre du modèle λ_0 s"obtient par la méthode de maximum de vraisemblance. Dans un échantillon de taille n, on a observé m décès et $n - m$ censures. La contribution à la vraisemblance d'un décès observé en t_i est $f(t_i)$, et celle d'un sujet censuré en t_i est $S(t_i)$. La vraisemblance s'écrit donc :

$$L(\lambda_0; t_1, ...t_n) = \prod_{deces} f(t) \prod_{censures} S(t) = \lambda_0^m exp(-\lambda_0 \sum_{i=1}^{n} t_i)$$

En annulant la dérivée de LnL par rapport à λ_0, on obtient l'estimateur du maximum de vraisemblance :

$$\widehat{\lambda}_0 = m/\sum_{i=1}^{n} t_i$$

Le risque instantané de décès est estimé par le nombre de décès divisé par la somme des temps de participation.

On vérifie que la dérivée seconde de LnL par rapport à λ_0 est négative, et on en déduit une estimation de la variance de $\widehat{\lambda}_0$:

$$\widehat{Var}(\widehat{\lambda}_0) = \widehat{\lambda}_0^2/m = m/(\sum_{i=1} t_i)^2$$

Un avantage de ce modèle est l'exitence de solutions explicites par le maximum de vraisemblance. On dispose donc d'estimateurs et de tests que l'on peut pratiquement calculer à la main, ce qui permet une approche exploratoire rapide et économique de données complexes.

On peut également donner l'exemple du **modèle de Weibull** qui suppose que la fonction de risque instantané varie avec le temps :

$$\lambda(t) = \gamma\ \lambda_0^\gamma\ t^{\gamma-1}$$

Le modèle de Weibull est un modèle à deux paramètres λ_0 et γ. Si γ est égal à 1, on retrouve le modèle exponentiel qui en est un cas particulier. Si γ est supérieur à 1, la fonction de risque est croissante. Si γ est inférieur à 1, la fonction de risque est décroissante. La fonction de survie du modèle de Weibull s'exprime comme :

$$S(t) = exp[-(\lambda_0\, t)^{\gamma}]$$

Nous venons de décrire plusieurs estimateurs des fonctions de survie : des estimateurs non-paramétriques (actuariel et Kaplan-Meier) et des estimateurs paramétriques (exponentiel et Weibull). En pratique, les modèles paramétriques, en dehors des modèles exponentiels et de Weibull, sont rarement utilisés. Les données disponibles ne permettent généralement pas de justifier un choix quelconque entre plusieurs fonctions de risque. Il est clair que si le modèle sous-jacent est un modèle paramétrique, exponentiel par exemple, l'estimateur correspondant est le meilleur. Dans le cas d'une survie exponentielle, l'estimateur paramétrique constant du risque instantané est opimal (sans biais et de variance minimale). Si en revanche, le modèle paramétrique utilisé ne correspond pas à la vraie fonction de risque, alors l'estimation de $S(t)$ est biaisée. Pour choisir entre plusieurs estimateurs paramétriques, il faut avoir assez d'informations à l'endroit où ces modèles diffèrent, c'est-à-dire en queue de distriution.

2.5 Test du Log-rank : Comparaison de fonctions de survie par l'approche non-paramétrique

2.5.1 Le contexte

Si on veut montrer qu'une caractéristique ou une action (traitement ou intervention) a un lien avec la survie, la démarche est de même nature que dans le cas de variables non temporelles, à savoir un test de comparaison.

On suppose que chaque patient appartient à un des deux groupes de traitement (A ou B). On note :

- t_{Ai} (respectivement t_{Bi}), la durée de suivi si le patient i du groupe A (respectivement du groupe B) est décédé,
- t^*_{Ai} (respectivement t^*_{Bi}), la durée de suivi si le patient i du groupe A (respectivement du groupe B) est censuré, encore vivant ou perdu de vue.

La comparaison des survies des deux groupes s'effectue grâce au test du log-rank.

2.5.2 Le principe du test du Log-rank

L'**hypothèse nulle** est l'égalité des probabilités de survie dans les deux groupes :

$$H_0 : S_A(t) = S_B(t)$$

IL s'agit donc de tester si la différence observée entre les données de survie dans les groupes A et B permet de rejeter l'hypothèse nulle. Cette méthode consiste donc à comparer le nombre d'évènements observés au nombre d'évènements attendus sous l'hypothèse nulle d'égalité de la survie dans les deux groupes.

Le **nombre de décès attendus sous l'hypothèse nulle** sont calculés de la façon suivante :

- au temps t_i, il y a $N_i^A - C_i^A$ patients à risque dans le groupe A et $N_i^B - C_i^B$ patients à risque dans le groupe B ;
- on a observé au total $D_i = D_i^A + D_i^B$ décès au temps t_i.

Sous l'hypothèse nulle, on "s'attend" à ce que les effectifs de décès dans l'intervalle $]t_{i-1}; t_i]$ dans les deux groupes se distribuent proportionnellement aux effectifs à risque dans les deux groupes :

$$E_i^A = D_i \frac{N_i^A - C_i^A}{N_i - C_i} = (1 - \widehat{S}(t_i|t_{i-1}))(N_i^A - C_i^A)$$

$$\text{et } E_i^B = D_i \frac{N_i^B - C_i^B}{N_i - C_i} = (1 - \widehat{S}(t_i|t_{i-1})) \times (N_i^B - C_i^B)$$

On calcule E_A et E_B le nombre total de décès attendus dans chaque groupe, et D_A et D_B le nombre total de décès observés dans chaque groupe. On peut montrer que, sous l'hypothèse nulle, la quantité

$$K = \frac{(D_A - E_A)^2}{E_A} + \frac{(D_B - E_B)^2}{E_B}$$

est distribuée selon une loi du Chi-deux à un ddl. Ce test, dit du Log Rank approché, permet de rejeter l'hypothèse d'égalité des courbes de survie quand $K > 3.84$.

On peut également généraliser la méthode pour comparer L courbes de survie ($L > 2$) avec le même principe de calcul. Sous l'hypothèse nulle, on s'attend à ce que les D_i décès au temps t_i se répartissent proportionnellement aux effectifs à risque dans les L groupes. On calcule les effectifs totaux attendus (E_l) et observés (D_l) pour chaque groupe l ($l = 1, ...L$). On en déduit la quantité K :

$$K = \sum_{l=1}^{L} \frac{(D_l - E_l)^2}{E_l}$$

qui sous l'hypothèse nulle, est distribuée selon une loi du Chi-deux à $L - 1$ ddl. On rejette l'hypothèse d'égalité des courbes de survie si K est supérieure à la valeur limite du χ^2 à $p - 1$ ddl au risque $\alpha = 5\%$.

2.6 Le modèle de Cox

2.6.1 Principe et définitions

Dans le cadre de l'analyse de données de survie, l'effet éventuel de covariables sur la survie est étudiée par biais du modèle de Cox. Les variables explicatives, notées X^j ($j = 1, ...p$), peuvent être quantitatives ou qualitatives, et sont renseignées pour chaque patient lors de l'entrée dans l'étude. Ces variables explicatives peuvent représenter des facteurs de risque, des facteurs pronostiques, des traitements, ou des caractéristiques intrinsèques au patient, ... Le modèle de Cox permet d'**exprimer le risque instantané de survenue de l'évènement en fonction de l'instant t et des variables explicatives X^j**, pour expliquer la survie sans donner aux fonctions de survie des formes paramétriques précises.

Rappelons que le risque instantané de survenu de l'évènement $\lambda(t, X^1, X^2, ...X^p)$ représente la probabilité d'apparition de l'évènement dans un intervalle de temps $]t, t + \Delta t]$ sachant que l'évènement ne s'était pas produit avant l'instant t. Le modèle de Cox exprime $\lambda(t, X^1, X^2, ...X^p)$ sous la forme :

$$\lambda(t, X^1, X^2, ...X^p) = \lambda_0(t)\ exp\{\sum_{j=1}^{p} \beta_j X^j\} = \lambda_0(t)\ exp\{\beta' X\}$$

Plusieurs remarques peuvent être apportées sur cette formule :

- Le risque instantané se décompose en deux termes dont l'un dépend du temps t, et l'autre des variables X^j.
- Si par exemple, les variables X^j représentent des facteurs de risque et si elles sont toutes égales à 0, alors $\lambda_0(t)$ est le risque instantané pour les sujets ne présentant aucun facteur de risque.
- La forme de $\lambda_0(t)$ n'étant pas précisée, c'est plutôt l'association entre les variables X^j et la survenue de l'évènement considéré qui est l'intérêt central du modèle. Cela revient donc à déterminer les cœfficients β_j.
- Le raport des risques instantanés de deux individus donc les caractéristiques respectives sont $X^1, X^2, ...X^p$ et $X'^1, X'^2, ...X'^p$ est :

$$\frac{\lambda(t, X^1, X^2, ...X^p)}{\lambda(t, X'^1, X'^2, ...X'^p)} = \frac{exp\{\sum_{j=1}^{p} \beta_j X^j\}}{exp\{\sum_{j=1}^{p} \beta_j X'^j\}}$$

  Ce rapport ne dépend pas du temps. De tels modèles sont dits à risques proportionnels. C'est une hypothèse importante du modèle de Cox.

2.6.2 Interprétation des cœfficients

Dans le cas d'un modèle comprenant une variable explicative dichotomique prenant les valeurs 0 ou 1 selon l'absence ou la présence de la caractéristique considérée, le rapport des risques instantanés des patients de la modalité 1 par rapport à la modalité 0 est :

$$\frac{\lambda(t,1)}{\lambda(t,0)} = e^{\beta} \Rightarrow \beta = ln\frac{\lambda(t,1)}{\lambda(t,0)}$$

Le coefficient β est donc le logarithme du risque instantané relatif de la modalité 1 par rapport à la modalité 0.

De façon plus générale, les cœfficients β_j représentent l'effet de la caractéristique X^j sur la survenue de l'évènement. Son interprétation peut être résumée en trois cas :

- Si β_j est nul, la caractéristique X^j n'a pas d'effet sur l'évènement considéré ;
- Si β_j est positif, des valeurs élevées de la caractéristique X^j sont associées à un risque instantané plus élevé ;
- Si β_j est négatif, des valeurs élevées de la caractéristique X^j sont associées à un risque instantané plus faible.

De plus, $exp(\beta_j)$ représente le facteur par lequel le risque de décès est multiplié en cas de présence de la caractéristique X_j.

2.6.3 Estimation et tests

Le principe pour le modèle de Cox est de n'estimer que les cœfficients β_j. On ne cherche pas à estimer $\lambda_0(t)$. Les estimations des cœfficients β_j sont obtenus par la méthode du maximum de vraisemblance. Plus exactement, seule la partie de la vraisemblance comportant de l'information sur les cœfficients β_j est retenue pour les calculs ; on parle de "vraisemblance partielle" ou de "vraisemblance de Cox".

Pour estimer le vecteur des paramètres β, il faut calculer la vraisemblance :

$$L(\beta; t_1, ...t_n, X_1, ...X_n) = \prod_{deces} f(t_i, X_i) \prod_{censures} S(t_j, X_j)$$

où :

- $f(t_i, X_i)$ représente la probabilité pour un patient de caractéristique X_i de décéder au temps t_i ;
- $S(t_j, X_j)$ représente la probabilité pour un patient de caractéristique X_j de survivre jusqu'à l'instant t_j.

La vraisemblance L contient la fonction $\lambda_0(t)$ dont l'étude ne nous intéresse pas réellement et que l'on considère comme un paramètre nuisible. La solution proposée par Cox conduit à éliminer $\lambda_0(t)$, en supposant que les instants auxquels se produisent les censures n'apportent pas d'information sur β. On ne retient donc que les k temps de décès. La contribution à la vraisemblance d'un patient de caractéristique X_i décédé au temps t_i s'exprime comme la probabilité conditionnelle d'observer un décès à l'instant t_i, sachant que l'on avait R_i sujets à risque en t_i :

$$L_i^*(\beta; t_i, X_i) = \frac{\lambda(t_i, X_i)}{\sum_{j \in R_i} \lambda(t_j, X_j)} = \frac{exp(\beta' X_i)}{\sum_{j \in R_i} exp(\beta' X_j)}$$

On obtient alors la log-vraisemblance de Cox :

$$LnL^* = \sum_{i=1}^{k} [\beta'(\sum_{l=1}^{m_i} X_l) - m_i Ln(\sum_{j \in R_i} exp(\beta' X_j))]$$

où m_i est le nombre de décès au temps t_i.
L'estimateur du maximum de vraisemblance du vecteur β des pramètres du modèle vérifie :

$$\frac{dLnL^*(\widehat{\beta})}{d\beta} = 0$$

On peut définir :

— $U(\beta)$ le vecteur des dérivées de LnL^* par rapport à β ;

— $I(\widehat{\beta}) = d^2LnL^*(\widehat{\beta})/d^2\beta$ la matrice des dérivées secondes par rapport à β, appelée matrice d'information.

On teste l'hypothèse que le vecteur des effets $(\beta_1, \beta_2, ...\beta_p)$ est nul. Trois tests peuvent être utilisés :

— le test du score : $U'(0)I^{-1}(0)U(0)$

— le test de Wald (ou maximum de vraisemblance) : $\widehat{\beta}'I(\widehat{\beta})\widehat{\beta}$

— le test du rapport de vraisemblance : $-2[LnL^*(0) - LnL^*(\widehat{\beta})]$

Ces trois statistiques suivent approximativement, sous l'hypothèse nulle, des distributions de Chi-deux à p degrés de liberté (où p est le nombre de variables explicatives prises en compte dans le modèle). Si les valeurs de ces statistiques sont supérieures à la valeur limite du χ^2 à p degrés de liberté au risque $\alpha = 5\%$, alors on rejette H_0 et on conclut à l'effet significatif d'au moins une des covariables sur la survie.

On peut également utiliser ces tests pour tester l'absence d'effet d'une (ou de plusieurs) covariable(s) en comparant le modèle complet au modèle réduit sous H_0. Dans ce cas, on compare la statistique de test à la valeur d'un chi-deux dont le nombre de degrés de liberté est le nombre de contraintes posées sous H_0.

2.7 Applications

7.1 Cent patients atteints de cancers ont été suivis pendant un maximum de 72 mois (soit 6 ans), pour étudier la survenue de décès chez ces patients. Les données de cette étude sont présentées par année de suivi, dans le tableau ci-dessous. Pour chaque année de suivi, on a noté le nombre de décès survenus et le nombre de censures : au bout de 12 mois de suivi, on a observé 6 décès et 2 patients sont censurés.

Suivi (en mois)	Nb de décès	Nb de censurés
$[0-12[$	6	2
$[12-24[$	10	6
$[24-36[$	9	11
$[36-48[$	7	10
$[48-60[$	4	18
$[60-72[$	1	12

1. Quelles peuvent être les différentes causes de censures ?
2. Combien de décès a-t-on observé sur l'ensemble du suivi ?
   Quel est le pourcentage de données censurées ?
3. Evaluer la survie de ces patients par la méthode actuarielle en utilisant l'année comme intervalle de temps.
4. Donner, si possible, une estimation de la survie médiane.
   Estimer le temps au delà duquel 75% des patients ont survécu.
5. Représenter la courbe de survie actuarielle. Commenter les résultats.
6. Quelle autre méthode aurait-on pu appliquer pour estimer la survie de ces patients et quelles données auraient été nécessaires pour cela ?

7.2 On veut comparer la mortalité dans deux groupes A et B. Dans le tableau suivant, n désigne l'année d'étude. X_n^A est le nombre de malades du groupe A vivant au début de l'année n , D_n^A le nombre de décès observé dans le groupe A. E_n^A est le nombre de décès attendu dans le groupe A sous l'hypothèse nulle d'égalité des survies dans les deux groupes. Tous groupes confondus, $X_n = X_n^A + X_n^B$ et $D_n = D_n^A + D_n^B$.

Dans cet exemple, on suppose qu'il n'y a pas de censures.

n	X_n^A	D_n^A	$S_{n/n-1}^A$	S_n^A	X_n^B	D_n^B	$S_{n/n-1}^B$	S_n^B	X_n	D_n	E_n^A	E_n^B
1	100	10			200	30						
2		10				40						
3		20				70						
4		30				40						
5		20				10						
$Total$	/		/	/	/		/	/	/			

1. Calculer les survies dans les deux groupes, en complétant les colonnes $S_{n/n-1}^A$, S_n^A, $S_{n/n-1}^B$, S_n^B. Tracer les courbes de survie estimées par groupe.
2. Calculer le nombre de décès attendus sous l'hypothèse d'égalité des survies (en complétant les colonnes E_n^A et E_n^B) et construire le test du Log-Rank. Conclure sur l'égalité des survies dans les deux groupes.

7.3 Une expérience vise à expérimenter la toxicité de la toxine botulique. On considère une cohorte de 10 rats et on veut étudier la probabilité de survenue d'un décès après injection de la toxine. Les données de cette expérience sont présentées dans le tableau ci-dessous : pour chaque rat, on dispose de la durée du suivi et du statu à la fin du suivi (1=décédé ou 0=vivant).

N° du rat	Durée du suivi	Statut à la fin du suivi
1	3	0
2	2	0
3	15	1
4	1	1
5	5	0
6	13	1
7	6	0
8	7	1
9	10	0
10	9	1

1. Calculer la fonction de survie et représenter la courbe correspondante.
2. Quelle est la survie médiane ?