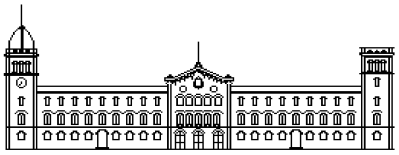

Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

ANÀLISI DE LA SUPERVIVÈNCIA PER DADES CENSURADES

Marta Bofill Roig

Director: Olga Julià de Ferran
Realitzat a: Departament de Probabilitat, Lògica i Estadística. UB

Barcelona, 27 de gener de 2014

Abstract

This work tries to introduce us in the survival analysis for censured information. Survival analysis born with the intention of analyzing information that represent the duration between two events. With him we will want to explain the evolution throughout the time of an event. We will define the time zero, from which the durations measure up, and the time of the event of interest, that is to say, the moment in the one that produces to himself the event that we are studying.

Examples where this type of study is applied are: demographic studies, industrial applications to prove the electronic reliability of components, epidemiology or clinical tests. In this work we will centre on the latter area, since there will be applied the concepts introduced to two real true databases of clinical tests.

We will begin the work defining the basic concepts that will be the functions that are in use for the survival analysis. In the chapter 2 we will introduce the concept of censoring and truncation, two characteristics of the information in this type of studies. In the chapter 3 we will present the non-parametric estimators for the survival analysis: the Kaplan-Meier's estimator for the survival function and the Nelson-Aalen's estimator for the hazard function, principally. Also, the properties of these. Finally, in the chapter 4, they will present the tests of hypothesis for the comparison of two or populations.

Índex

Introducció

L'anàlisi de supervivència és un conjunt de tècniques estadístiques amb el propòsit d'examinar dades que representen la duració entre dos events. Es pretén explicar l'evolució al llarg del temps d'un succés. L'origen es troba amb el mètode clàssic per estimar patrons de mortalitat de poblacions: les taules de vida. Aquest mètode es va usar per actuaris, demògrafs, metges i epidemiòlegs. Aquestes taules servien per resumir la mortalitat (o el succés d'interés per l'estudi) d'una població específica dins un període de temps determinat.

És a partir de la segona meitat del segle XX quan les aplicacions es centren en l'àmbit clínic i el nombre de beneficis i utilitats d'aquest camp a la salut són incomptables. En el camp de la medicina, l'anàlisi de la supervivència agafa cos al moment del diagnòstic d'una malaltia o al moment en que es realitza un tractament. Apareix la necessitat d'estudiar el comportament al llarg del temps d'una enfermetat o comparar l'eficiència entre tractaments. Aquests fets es transformen en els objectius principals de l'anàlisi de la supervivència.

La situació típica per l'ús d'aquests mètodes estadístics sorgeix quan es vol analitzar dades d'estudis de cohort o d'assajos clínics. Es selecciona una mostra de la població la qual es sol dividir en diferents grups, com per exemple: tractats i no tractats per un fàrmac (o diversos fàrmacs) o exposats i no exposats a un factor d'interès. Aquesta mostra de la població es segueix durant l'interval de temps que dura l'estudi i es registra, per a cada pacient (o participant), l'aparició de l'esdeveniment d'interès. És aquí on apareix un dels conceptes més importants i particulars de l'anàlisi de supervivència: la censura. La censura es dóna quan la informació de la que es disposa per alguns individus és incompleta. És a dir, dins la mostra dels pacients es poden diferenciar dos tipus de participants: els participants que presenten l'esdeveniment durant l'estudi i els que no el presenten (censurats). Aquest fet succeix per motius com: l'abandonament de l'estudi (no volen continuar amb el tractament, no volen seguir l'estudi o presenten efectes secundaris del tractament), la finalització de l'estudi sense haver set observat l'event o la mort del pacient per una causa no relacionada amb l'experiment.

La variable o mesura d'interés en aquest tipus d'estudi és el temps des de l'origen de l'experiment fins que succeix l'event d'interès. En conseqüència es tracta de temps en risc al temps des de l'origen fins a l'esdeveniment i temps de fallada al temps on succeix l'event.

Exposarem mètodes per explicar el comportament de temps de vida quan les dades recollides estan censurades per la dreta i la hipòtesi de normalitat no és adequada. A més, enfocarem l'estudi a les aplicacions mèdiques de l'anàlisi de supervivència, aplicant els coneixements que es van introduïnt a unes bases de dades reals d'assaigs clínics. Cal recalcar que també es podria aplicar a altres àmbits com: models economètrics, estudis psicològics, tecnologia d'aliments o aplicacions industrials.

L'objectiu del treball és introduir en els mètodes estadístics de l'anàlisi de la supervivència i en el tractament de dades censurades, establir l'ús no paramètric d'aquest tipus de dades tant en l'estimació de funció de supervivència com en els test de comparació de diferents corbes de supervivència. A més, aquestes tècniques s'aplicaran a dues bases de dades reals dins el món mèdic.

Capítol 0

Dades de supervivència

El propòsit principal d'aquest treball, juntament amb introduir-nos al món de l'anàlisi de supervivència, és implementar els mètodes estadístics que anirem coneixent a unes bases de dades reals. Comencem idò el primer capítol amb la presentació dels assaigs clínics dels quals usarem les dades.

0.1 Remissió d'un assaig clínic per a la leucèmia aguda

Freireich et al. (1963)

Dades d'un assaig clínic de la droga 6 − *mercaptopurina* vs *placebo*. Per aquest assaig es vàren seleccionar 42 nens amb leucèmia aguda de 11 hospitals diferents d'Estats Units. Els pacients seleccionats requerien tenir una completa o parcial remissió de la leucèmia induïda pel tractament amb prednisona. Per a fer l'estudi, es va agrupar fent parelles de pacients amb el mateix estat de remissió i tractats al mateix hospital, aleatòriament a un pacient de la parella se li va donar 6 − *mercaptopurina* i a l'altre *placebo*. D'aquesta manera, 21 van ser tractats amb 6 − *mercaptopurina* i 21 amb *placebo*. L'assaig va seguir l'evolució dels pacients fins la recaiguda o fins al final de l'estudi.

0.2 Càncer de mama, dades de Rotterdam

Conjunt de dades de Rotterdam sobre el càncer de mama. Per l'estudi es va comptar amb la informació de 2982 dones a les que se'ls hi va detectar el primer tumor mamari entre els anys 1978 i 1993. Es va fer un seguiment a les dones fins al final d'aquest any. La variable d'interés fou el temps des de la primera intervenció quirúrgica pel tumor fins a l'aparició del primer dels següents events:

- Recurrència locorregional. Es dóna quan el càncer torna a apareixer i s'ha diseminat a altres llocs de l'organisme.
- Tumor contralateral o secundari. És una metàstasis, té per origen un càncer primitiu. Aquest resulta de la difusió de les cèl·lules cancerígenes es diseminen i formen un nou tumor, per via limfàtica o sanguinea.
- Mort per càncer de mama.

El temps fins la mort per altres causes i els temps fins el final del seguiment si no apareixia cap dels events d'estudi varen considerarse com temps censurat. El rang de temps de seguiment

fou d'un mes fins 231 mesos, amb una mitjana de 5.3 anys. En total es van obervar 1518 events (50.9%).

L'interés de l'estudi va consistir en estudiar diferents factors associats a la supervivència de dones amb càncer de mama. Aquests factors, mesurats al moment del diagnòstic, van ser: edat, estat menopàusic (premenopausa o postmenopausa), grau del tumor, tamany del tumor, quantitat de nodes limfàtics, receptor de progesterona, tractament amb tamoxifen (hormonada o no-hormonada) i quimioteràpia (si havien estat tractades o no amb quimioteràpia).

Capítol 1

Conceptes bàsics

En aquest capítol volem presentar els conceptes que s'utilitzen per als models d'anàlisi de supervivència. Sigui X el temps fins l'event específic ϵ que volem estudiar. Aquest event potser, per exemple, l'aparició d'un tumor, el desenvolupament d'una enfermetat o la reaparició de símptomes. El temps X és una variable aleatòria no negativa sobre una població homogènia. La distribució de X ve caracteritzada per sis funcions:

- Funció de supervivència (survival function), $S(x)$.
- Funció de distribució (distribution function), $F(x)$.
- Funció de densitat de probabilitat (probability density function), $f(x)$.
- Funció de risc (hazard function), $\lambda(x)$.
- Funció de risc acumulat (cumulate hazard function), $\Lambda(x)$.
- Vida mitjana residual (mean residual life), $vmr(x)$.

1.1 Funcions de supervivència i de distribució

La *funció de supervivència* és la probabilitat de sobreviure més de x unitats de temps. O el que és el mateix, que el succés ϵ ocurreixi després de x. La denotem per S i és la funció $S : \mathbb{R}^+ \to [0, 1]$ que formalment queda definida com:

$$S(x) = Prob(X > x) \text{ on } x \geq 0.$$

En el contexte de l'industria, la funció de supervivència també es coneix com *funció de fiabilitat*.

La funció de supervivència pot tenir diverses formes però totes compleixen una sèrie de característiques. S és una funció monòtona decreixent i si X és una variable aleatòria continua, aleshores $S(x)$ és continua i estríctament decreixent. A més, es compleix

$$S(0) = 1 \text{ i } \lim_{x\to\infty} S(x) = 0.$$

Per tant, totes les funcions de supervivència comencen des del 1 quan $x = 0$ i decreixen monotònicament fins convergir al 0 quan $x \to \infty$. La velocitat de decreixement de S depèn del risc d'experimentar ϵ a l'instant x.

La *funció de distribució* associada a una variable aleatòria X és la funció $F : \mathbb{R} \to [0, 1]$ definida com

$$F(x) = Prob(X \leq x).$$

La funció F és complementaria de la funció de supervivència, és a dir, $F(x) = 1 - S(x)$. La funció de distribució F d'una variable aleatòria X, dins l'anàlisi de supervivència, representa la probabilitat de sobreviure menys de x unitats de temps. Com és sabut F compleix les propietats següents:

- F és monòtona creixent.
- F és contínua per la dreta.
- $\lim_{x\to\infty} F(x) = 1$ i $\lim_{x\to-\infty} F(x) = 0$.

Si X és una variable aleatòria continua, aleshores $F(x)$ és continua i estríctament creixent. Per tant, anàlogament a S, totes les funcions de distribució d'una variable positiva, comencen des del 0 quan $x = 0$ i creixen monotònicament fins convergir al 1 quan $x \to \infty$.

1.2 Funcions de densitat de probabilitat i de massa de probabilitat

Abans hem parlat de la velocitat de decreixement de S, aquesta velocitat depèn de l'instant x però no podem treure conclusions només amb el gràfic de S. Per calcular la intensitat de probabilitat, si les variables són absolutament continues, usem la *funció de densitat de probabilitat*. La denotem per f i formalment es defineix com:

$$f(x) = \lim_{\Delta x\to 0} \frac{1}{\Delta x} Prob[x \leq X < x + \Delta x]$$

Intuitivament $f(x)\Delta x$ s'interpreta com la probabilitat de que l'event ϵ passi a $(x, x + \Delta x)$. Es dedueixen les següents propietats:

- $f(x) \geq 0$, $\forall x$.
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$
- $S(x) = Pr(X > x) = \int_x^{+\infty} f(u)\, du$.
- $f(x) = \frac{\delta F(x)}{\delta x}$ i $f(x) = \frac{-\delta S(x)}{\delta x}$.

Anàlogament, es defineix la *funció de massa de probabilitat* per a variables discretes i es denota per p. Si X pren els valors $x_1, ..., x_n$, la funció de massa de probabilitat queda definida com:

$$p(x_j) = Prob(X = x_j), \forall j$$

Les propietats essencials de la funció de massa de probabilitat són

- $p(x_j) \geq 0$.

- $\sum_j p(x_j) = 1$.
- $S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j)$
  $F(x) = Pr(X \leq x) = \sum_{x_j \leq x} p(x_j)$

1.3 Funcions de risc i risc acumulat

La *funció de risc*, quan X és una variable aleatòria absolutament continua, es defineix formalment com:

$$\lambda(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} Prob[x \leq X < x + \Delta x | X \geq x].$$

Intuitivament, $\lambda(x)\Delta x$ s'interpreta com la probabilitat de que a un individu d'edat x li passi ϵ a $(x, x + \Delta x)$ sabent que en x encara no s'havia produït ϵ. La funció de risc expressa com el risc canvia amb el temps i conté la mateixa informació que la supervivencia però en termes de la velocitat de canvi. S'estima mitjançant la proporció de persones que fallen en el moment x entre les que no havien fallat abans.
Algunes de les propietats de la funció de risc són les següents:

- $\lambda(x)$ és una funció no negativa, ja que és límit de reals no negatius.
- $\lambda(x) = \frac{f(x)}{S(x)} = -\frac{\delta}{\delta x}(\ln S(x)) \Rightarrow S(x) = \exp\{-\int_0^s \lambda(u)du\}$.
- $\int_0^s \lambda(u)\, du < \infty \quad \forall s$ tal que $S(x) < 1$, per la propietat anterior.

Per les variables aleatòries absolutament continues, es defineix la *funció de risc acumulada* com:

$$\Lambda(x) = \int_0^x \lambda(u)du.$$

La funció de risc acumulada és una funció no negativa i monotona creixent. A més, satisfà: $\Lambda(x) = -\ln S(x)$.

Suposem que X és una variable aleatòria discreta que pren valors $\{x_1 < x_2 < ...\}$. Es defineix la *funció de risc* com:

$$\lambda(x_k) = Prob(X = x_k | X > x_{k-1})$$

Les propietats bàsiques de la funció de risc són:

- $\lambda(x_j)$ és una probabilitat.
- $\lambda(x_j) = \frac{p(x_j)}{S(x_{j-1})} = \frac{S(x_{j-1}) - S(x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})}$.
- Representació en forma de producte de la funció de supervivència
  $S(x) = \prod_{x_j \leq x} \frac{S(x_j)}{S(x_{j-1})} = \prod_{j:x_j \leq x}(1 - \lambda(x_j))$.
  En efecte, al punt x_k tenim:

$$\begin{aligned} S(x_k) &= P(X > x_k) = P(X > x_k, X > x_{k-1}) \\ &= P(X > x_k | X > x_{k-1}) P(X > x_k) \\ &= P(X > x_k | X > x_{k-1}) ... P(X > x_2 | X > x_1) P(X > x_1) \\ &= (1 - \lambda(x_k)) ... (1 - \lambda(x_2))(1 - \lambda(x_1)) \\ &= \prod_{i=1}^{k} (1 - \lambda(x_i)) \end{aligned}$$

Per tant, $S(x) = \prod_{j:x_j \leq x} (1 - \lambda(x_j))$.

Si X és una variable discreta que pren els valors $x_1 < x_2 < ...$, es defineix la funció de risc acumulada com:

$$\Lambda(x) = \sum_{j:x_j \leq x} \lambda(x_j)$$

en aquest cas, no es satisfà l'equació $\Lambda(x) = -\ln S(x)$ comentada anteriorment. Altres autors defineixen la funció de risc acumulada de la manera següent:

$$\Lambda(x) = -\sum_{j:x_j \leq x} \ln(1 - \lambda(x_j))$$

amb aquesta darrera equació si es satisfà $\Lambda(x) = -\ln S(x)$. Les dues definicions donen valors propers quan les taxes de fallada són petites.

La funció de risc juga un paper important a l'anàlisi de la supervivència. A continuació presentarem els patrons més usuals que pot pendre aquesta funció:

- Funció de risc creixent: Correpon a poblacions que envelleixen amb l'edat o per desgast.
- Funció de risc constant: Correspon a poblacions que no presenten envelliment. N'és un exemple la vida útil de les bombetes.
- Funció de risc decreixent: Correspon a poblacions que s'enforteixen amb el pas del temps. Per exemple el temps entre infarts, quan més temps fa des del darrer infart menys probable és que es repeteixi.
- Funció de risc amb forma de banyera: Correspon a aquelles poblacions que la seva funció de risc és decreixent a l'inici, constant durant un llarg període de temps i creixent al final de la vida. Aquest model s'ajusta a la vida dels éssers vius.
- Funció de risc amb forma de gepa: Correspon a aquelles poblacions que la seva funció de risc és creixent a l'inici i després d'un cert temps comença a decreixer. Aquest model es pot aplicar al temps de rebuig d'un òrgan en cas de transplantament.

1.4 Vida residual

Per acabar aquest capítol de presentació dels conceptes bàsics de l'anàlisi de supervivència definirem la *vida residual mitjana*, denotada per $vrm(x)$, i la *vida residual mediana*, denotada per $vrmed(x)$. La vida residual mitjana mesura l'esperança de vida restant i es defineix formalment com:

$$vrm(x) = E(X - x | X > x).$$

Quan $x = 0$, la vida residual mitjana coincideix amb la mitjana.

La vida residual mediana a x és el temps en el que la supervivència restant descendeix a la meitat, és a dir, formalement:

$$u = vrmed(x), \text{ si } S(x+u) = \frac{S(x)}{2}.$$

Mesura als individus d'edat x la vida mediana restant.

Capítol 2

Censura i truncament

Quan s'estudia amb dades de supervivència ens trobem amb dificultats al mesurar els temps de vida. Un dels problemes més freqüents és la *censura*. La censura es dóna quan només observem els temps de vida si es troben dins un certs intervals de temps. És a dir, el temps quan es produeix ϵ no s'observa de forma exacta. Per exemple, perquè l'estudi ha acabat abans de que es produexi ϵ o perquè es produeix abans que la persona entri a l'estudi. En aquests casos, només sabem que ϵ es produeix dins un interval de temps. Per estudiar adequadament la censura caldrà considerar com es van obtenir les dades. Existeixen diversos tipos de censura: censura per la dreta, per l'esquerra o intervals de censura cada tipus dóna lloc a una funció de probabilitat diferent que serà la base per la inferència. En aquest treball ens centrarem en estudiar la censura per la dreta.

Per a l'estudi, considerarem un nombre n de pacients i notarem per X_1, ..., X_n les variables aleatòries, independents i idènticament distribuides que corresponen als temps fins ϵ de cada pacient, amb $f(x)$ funció de densitat i $S(x)$ funció de supervivència associades.

2.1 Censura

S'anomena *censura tipus I* quan l'event és observat només si succeeix abans d'un temps especificat a priori, C_R. Només sabrem la X d'un individu quan $X \leq C_R$. Si $X > C_R$, l'individu és un supervivent i la dada està censurada. Cada dada de l'experiment es representarà amb el parell de variables aleatòries (T, δ), on $T = min(X, C_R)$ i

$$\delta = \begin{cases} 1 & \text{si } X \leq C_R \\ 0 & \text{si } X > C_R \end{cases}$$

És a dir, $\delta = 1$ si la dada és no censurada i $\delta = 0$ si la dada és censurada. A la variable δ se l'anomena *indicador de censura*. Així, per una mostra de n pacients tindrem: $(T_1, \delta_1), ..., (T_n, \delta_n)$.

S'anomena *censura tipus I progressiva* quan el temps de censura C_R són diferents per a cada grup d'individus. Per aquesta situació s'estableix un nombre finit de temps de censura $C_1, ..., C_m$, on $m \leq n$, i es divideixen els individus en M grups de manera que cada grup té el seu temps de censura.

D'aquesta manera tindrem, $(1, ..., n_1)$ pacients que conformen el grup 1 amb temps de censura C_1, $(n_1 + 1, ..., n_2)$ els pacients que conformen el grup 2 amb temps de censura C_2, i així successivament fins $(n_{m-1}+1, ..., n_m)$, on $n_m = n$, grup m amb temps de censura C_m. Per tant, en aquest cas tindrem: $(T_1, \delta_1), ..., (T_n, \delta_n)$ on per a cada $j = 1, ..., m$ i per a cada $n_{j-1} + 1 \leq i \leq n_j$, $T_i = min(X_i, C_j)$ i

$$\delta_i = \begin{cases} 1 & \text{si } X_i \leq C_j \\ 0 & \text{si } X_i > C_j \end{cases}$$

S'anomena *censura tipus I generalitzada* quan els pacients entren a l'estudi en diferents moments però s'ha establert una data C_R on es tancarà l'estudi. Denotem per E_i el temps d'inici l'estudi del pacient i-èsim i F_i el temps potencial de fallada d'aquest mateix pacient. Si reescalem els temps de la manera següent:

$T_i = F_i - E_i$

$C_i = C_R - E_i$

obtenim censura tipus I, però amb cada individu i un temps de censura C_i que depèn de l'individu i. Aleshores, observarem $(T_1, \delta_1), ..., (T_n, \delta_n)$ on $T_i = min(X_i, C_i)$ i

$$\delta_i = \begin{cases} 1 & \text{si } X_i \leq C_i \\ 0 & \text{si } X_i > C_i \end{cases}$$

El segon tipus de censura per la dreta s'anomena *censura tipus II*, aquesta es dóna quan l'estudi continua fins que succeeix el temps de fallada r-èsima on r és un nombre natural predeterminat tal que $r < n$. Els experiments que segueixen aquest tipus de censura solen estar vinculats a l'estudi de temps de vida a l'enginyeria on es posen a prova equips molt costosos o molt fiables, amb aquest tipus de censura poden estalviar temps i diners. En aquest cas, tots els elements comencen el test a la vegada i l'experiment acaba quan fallen r elements. Per tant, no sabem quant de temps durarà l'experiment, és a dir, el temps en el qual s'acaba l'experiment, T_r, és aleatòri però el nombre d'errors, r, i el nombre de dades censurades, $n - r$, són fixades i per tant no aleatòri.

La *censura tipus II generalitzada*, on podem trobar semblances amb la censura tipus I progressiva i amb la censura tipus I generalitzada. Ens trobem amb la següent situació: una mostra de n elements per a l'experiment i un nombre r_1 fixat, $r_1 < n$. Comencem l'experiment i quan arribem al r_1-èsima fallada, prenem aleatòriament $n_1 - r_1$ elements del rest dels $n - r_1$ elements que no han fallat i els eliminem de l'experiment. Ara, la mostra per a l'experiment consta de $n - n_1$ elements. Es fixa un nombre r_2, natural tal que és menor que el nombre d'elements a la mostra i observem fins que succeeix el r_2-èsima fallada. Seguim iterant els procés fins un nombre fixat de fallades que volem observar o bé fins que ja no ens queden elements a la mostra.

El tercer tipus de censura per la dreta és la *censura per riscos competitius*. Un cas especial d'aquest és la *censura aleatòria*. Aquest tipus de censura sorgeix quan estem interessats en la estimació de la distribució marginal d'algun event però alguns dels individus han d'abandonar l'estudi per algún altre succés que anomenarem event competitiu. Per exemple, en el cas d'un assaig clínic, els pacients poden patir una mort accidental, migració, sortida de l'assaig per toxicitat, etc. L'event d'interés no es observat si els individus han experimentat l'event competitiu, aleshores aquests temps tenen censura per la dreta en aquell moment. Davant els riscos competitius és important determinar quina quantitat es desitja estimar.

2.2 Truncament

La segona característica que presenten les dades de l'anàlisis de supervivència és el *truncament*. El truncament succeeix quan només els pacients amb el temps de fallada dins un cert interval de temps són observats. Els pacients els qual el seu event s'ha produit fora d'aquest temps (Y_L, Y_R) no són observats i l'investigador no sap de la seva existència ni té dades sobre aquests.
Quan Y_R és infinit, parlem de *truncament per l'esquerra*, aquí els únics individus observats son aquells amb temps de l'event més gran que Y_L. És a dir, X és observat si, i només si, $Y_L < X$. Als estudis de supervivència, el truncament pot ser la exposició a algunes enfermetats, el diagnòstic d'alguna enfermetat, la entrada a un assaig clínic, etc. Notem que, a diferència de la

censura per l'esquerra on tenim informació parcial sobre els individus que experimenten l'event d'interés abans de l'edat d'ingrés, pel truncament per l'esquerra aquests individus mai van estar considerats per la seva inclusió dins l'estudi.
El *truncament per la dreta* ve donat quan Y_L és igual a zero. És a dir, observem X si, i només si, $X \leq Y_R$. Exemples per aquest tipus de truncament poden ser estudis d'estimació de la distribució de les estrelles des de la terra, on les estrelles molt llunyanes no són visibles i estan truncades, també estudis de mortalitat usant registres de defunció.

Capítol 3

Estimadors no paramètrics de l'anàlisi de supervivència

En aquest capítol introduïm i estudiem l'estimador per la funció de supervivència que van proposar Kaplan i Meier a l'any 1958 que es basa en la descomposició de la funció de supervivència en un producte de probabilitats condicionades. Distinguirem els casos on la mostra de la població té tots els temps de vida diferents (no hi ha empats) dels casos on la mostra té alguns temps repetits (hi ha empats). Veurem també possibles estratègies per estimar la funció de risc i l'estimador proposat per Nelson i Aalen per la funció de risc acumulada. Una vegada presentats tots els conceptes, veurem quines són les propietats més importants de l'estimador Kaplan i Meier i les propietats asimptòtiques dels estimadors Kaplan i Meier i Nelson i Aalen.

3.1 Estimador de Kaplan i Meier

Considerem $X_1, ..., X_n$ una mostra de variables aleatòries que medeixen el temps fins l'esdeveniment ϵ de la població que estudiem i $C_1, ..., C_n$ variables aleatòries independents i idènticament distribuides amb funció de distribució G que representen les censures. Suposarem que la censura és no informativa, és a dir, que les variables C_i són independents a les X_j. Usarem la notació que s'ha explicat al capítol anterior, és a dir, les dades observades consistiran en les parelles $(T_1, \delta_1), ..., (T_n, \delta_n)$ on

$T_i = min(X_i, C_i)$ i

$$\delta_i = \begin{cases} 1 & \text{si } X_i \leq C_i, \\ 0 & \text{si } X_i > C_i. \end{cases}$$

A més, denotarem per $x_{max} = X_{(n)}$ i $x_{ult} = max\{T_i;\ \delta_i = 1\}$.

3.1.1 Estimador Kaplan i Meier quan no hi ha empats

Suposem que totes les dades observades, T_i, són diferents. Considerem l'estadístic d'ordre $T_{(1)} < ... < T_{(n)}$ i denotem per $\delta_{(i)}$ la variable δ associada a $T_{(i)}$, per simplificar la notació posem $\tau_i = T_{(i)}$. Definim els següents conceptes:

- Els intervals aleatoris $I_i = (\tau_{i-1}, \tau_i] := (T_{(i-1)}, T_{(i)}]$, $i = 1, ..., n$. Observem que cadascun d'ells conté exactament un valor observat.
- Definim *conjunt de risc* al moment x, $R(x)$, al conjunt dels individus vius a l'instant x^-.
- $n_i = card(R(T_i))$

  $d_i = \sum_{j=1}^{n} \mathbf{1}_{\{\mathbf{T_j}=\mathbf{T_{(i)}},\delta_\mathbf{j}=\mathbf{1}\}}$, és a dir, d_i és el nombre de pacients que moren a $T_{(i)}$.

$$
\begin{aligned}
p_i &= Prob(\text{pacient viu al final de } I_i|\text{pacient viu al principi de } I_i) \\
&= Prob(X > \tau_i | X > \tau_{i-1}) \\
&= Prob(X > \tau_i)/Prob(X > \tau_{i-1})
\end{aligned}
$$

$$
\begin{aligned}
q_i = 1 - p_i &= Prob(\text{pacient mor a } I_i|\text{pacient viu al principi de } I_i) \\
&= Prob(X < \tau_i | X > \tau_{i-1}) \\
&= Prob(\tau_{i-1} < X \leq \tau_i)/Prob(X > \tau_{i-1}).
\end{aligned}
$$

Les probabilitats p_i, q_i s'estimen de forma natural com:
$\widehat{q_i} = \frac{d_i}{n_i} = \frac{d_i}{n-i+1} = \frac{\delta_i}{n-i+1}$ i $\widehat{p_i} = 1 - \frac{\delta_i}{n-i+1} = \frac{n-i+1-\delta_i}{n-i+1}$.

Ara, si descomponem la funció de supervivència com un producte de probabilitats condicionades, obtenim:

$$
\begin{aligned}
S(\tau_k) &= Prob(X > \tau_k) = Prob(X > \tau_k, X > \tau_{k-1}) \\
&= Prob(X > \tau_k | X > \tau_{k-1}) Prob(X > \tau_k) \\
&= Prob(X > \tau_k | X > \tau_{k-1}) ... Prob(X > \tau_2 | X > \tau_1) Prob(X > \tau_1) \\
&= p_1 p_2 ... p_k
\end{aligned}
$$

Per estimar la funció de supervivència sustituirem els valors p_i per les seves estimacions i utilitzem que $n_i = n - i + 1$, d'aquesta manera ens queda:

$$
\begin{aligned}
\widehat{S}(x) &= \prod_{i:T(i)<x} \widehat{p_i} = \prod_{i:T(i)<x} (1 - \frac{1}{n_i})^{\delta_i} \\
&= \prod_{i:T(i)<x} (1 - \frac{1}{n-i+1})^{\delta_i} = \prod_{i:T(i)<x} (\frac{n-i}{n-i+1})^{\delta_i}
\end{aligned}
$$

Observem que si $\delta_i = 1, \forall i$ es té:

$$
\widehat{S}(x) = \prod_{i:T(i)<x} (\frac{n-i}{n-i+1}) = \frac{n-k}{n} = 1 - \frac{k}{n} = 1 - \frac{\sum_{1 \leq i \leq n} \mathbf{1}_{\{\mathbf{T_{(i)}} \leq \mathbf{x}\}}}{n}.
$$

Per tant, l'estimador de Kaplan i Meier coincideix, en aquest cas, amb la supervivència empírica.

3.1.2 Estimador Kaplan i Meier quan hi ha empats

Suposem ara que hi poden haver empats. Considerem ara $Y_{(1)} < ... < Y_{(r)}$, l'estadístic ordenat que correspon als r temps de mort diferents. Definim $I_i = (Y_{(i-1)}, Y_{(i)}]$, suposem que en I_k tenim n_k individus vius abans de Y_k i que es produeixen $d_k \geq 1$ morts en $Y_{(k)}$. Amb aquesta nova situació, l'estimació de p_k varia respecte l'anterior ja que ara podem observar més d'un temps de mort. Podem raonar de la forma següent subdividim l'interval I_i en d_k intervals de manera que a cada subinterval tinguem una única mort. Aleshores, aplicant la idea de l'apartat anterior la probabilitat de sobreviure a $Y_{(k)}$ saben que estem vius a $Y_{(k-1)}$ s'estimaria de la forma:

$\widehat{p_k} = (1 - \frac{1}{n_k})(1 - \frac{1}{n_k - 1})...(1 - \frac{1}{n_k - d_k + 1}) = 1 - \frac{d_k}{n_k}$.

L'estimador de Kaplan i Meier queda definit per a tota x dins el rang de les dades, és a dir, $\forall x \leq x_{max}$ de la forma:

$$\widehat{S}(x) = \begin{cases} 1 & \text{si } x < T_{(1)} \\ \prod_{i:T_{(i)} \leq x} (1 - \frac{d_i}{n_i}) & \text{si } x \geq T_{(1)}. \end{cases}$$

3.1.3 Estimador Kaplan i Meier quan la darrera observació ordenada està censurada

Recordem que la funció de supervivència i la funció de distribució estan estretament vinculades per la fórmula: $F(x) = 1 - S(x)$. A més, recordem la funció de distribució satisfà: $\lim_{x\to\infty} F(x) = 1$ i $\lim_{x\to-\infty} F(x) = 0$. Quan la darrera observació ordenada està censurada, usant l'estimador $\widehat{S}$ que hem trobat a l'apartat anterior tenim $\lim_{x\to\infty} \widehat{S}(x) > 0$. Per tant, l'estimador no té una definició adient. Per solucionar aquest problema van sorgir dos possibles propostes per afrontar-ho.

Efron va proposar redefinir $\widehat{S}(x) = 0, \forall x \geq T_{(n)}$, mentres que Gill va proposar considerar que $\widehat{S}(x) = \widehat{S}(T_{(n)})$, quan $\delta_{(n)} = 0, \forall x > T_{(n)}$. Els dos suggeriments asimptòticament es comporten igual i les dues convergeixen a la funció de supervivència real per a mostres de tamany gran, però un estudi de les propietats per mostres petites revela que l'estratègia seguida per Gill (amb biaix positiu) té un millor comportament.

3.1.4 Variança de l'estimador de Kaplan i Meier

En primer lloc, introduirem la fórmula de Greenwood una de les fórmules d'aproximació possibles de la variància de l'estimador de la supervivència.

Proposició 3.1.1. *La variança de l'estimador de Kaplan i Meier es pot aproximar per*

$$Var(\widehat{S}(\tau_k)) \approx S(\tau_k)^2 \sum_{i=1}^{k} \frac{q_i}{p_i n_i}$$

Utilitzant aquesta expressió es defineix l'estimador de la variança de $\widehat{S}(\tau_k)$ *com*

$$\widehat{Var}(\widehat{S}(x)) = \widehat{S}(x)^2 \sum_{i:Y_{(i)} \leq x} \frac{d_i}{(n_i - d_i)n_i}$$

i es coneix com fórmula de Greenwood.

En general, el problema amb el que ens trobem és que els estimadors proposats presenten biaix per mostres petites i moderades.

Aalen i Johansen proposaren la següent modificació per l'estimació de la variança de $\widehat{S}(x)$:

$$\widehat{Var}_{AJ}(\widehat{S}(x)) = \widehat{S}(x)^2 \sum_{i:T_{(i)} \leq x} \frac{d_i}{n_i^2}$$

aquesta fórmula es coneix com *variança de Tsiatis en R* i es basa en la relació:

$\widehat{S}(x) = exp\{-\widehat{\Lambda}(x)\}$, de la que parlarem més al próxim apartat.

3.2 Estimació funció risc i l'estimador Nelson i Aalen

En aquesta secció introduirem l'estimador de Nelson-Aalen, estimarem les funcions de risc i risc acumulada i veurem que mitjançant la relació $S(x) = exp\{-\Lambda(x)\}$ podrem fer comparacions entre les estimacions de Nelson-Aalen i Kaplan-Meier.

Recordem que si $\lambda(x)$ és la funció de risc, $\lambda(x)\Delta x$ s'interpreta com la probabilitat de que a un individu d'edat x li passi ϵ a $(x, x+\Delta x)$. La funció de risc al moment $T_{(i)}$ s'estima mitjançant la proporció de pacients que han fallat a $T_{(i)}$ entre la proporció de pacients en risc a $T_{(i)}$, és a dir, $\widehat{\lambda}(x) = \frac{d_i}{n_i}$.

Es defineix l'estimador de Nelson-Aalen per la funció de risc com:

$$\widehat{\lambda}_{NA}(x) = \begin{cases} 0 & \text{si } x \neq T_{(i)} \\ \frac{d_i}{n_i} & \text{si } x = T_{(i)} \end{cases}$$

Per altra banda, recordem que es defineix la funció de risc acumulada com: $\Lambda(x) = \int_0^x \lambda(u)du$. L'estimador que van proposar Nelson-Aalen per aquesta funció va estar:

$$\widehat{\Lambda}_{NA}(x) = \begin{cases} 0 & \text{si } x \leq T_{(1)} \\ \sum_{i:T_{(i)} \leq x} \frac{d_i}{n_i} & \text{si } x \geq T_{(1)} \end{cases}$$

A partir d'aquests estimadors i tinguent en compte la relació $S(x) = exp\{-\Lambda(x)\}$ podem aproximar la funció de supervivència amb un nou estimador de la forma:

$$\widehat{S}_{NA}(x) = exp\{-\widehat{\Lambda}_{NA}(x)\}$$

De la mateixa manera, considerant la funció inversa d'aquesta podem estimar la funció de risc acumulada mitjançant l'estimador de Kaplan i Meier:

$$\widehat{\Lambda}_{KM}(x) = -\ln(\widehat{S}_{KM}(x))$$

L'estimador de Nelson-Aalen té principalment dos usos, per un costat seleccionar entre models paramètrics i per l'altre, serveix per estimar la funció de risc. Es comporta millor que el basat en Kaplan-Meier quan les mostres són petites. Quan no es satisfà la hipòtesis de que la censura és no informativa, les conclusions que extraiem dels dos estimadors poden ser totalment errònies, de fet, estarien estimant unes altres funcions.

Proposició 3.2.1. *La variança de l'estimador Nelson-Aalen, anomenada variança de Tsiatis en R ve donada per*

$$\sigma^2_{NA}(x) = \sum_{i:T_{(i)} \leq x} \frac{d_i}{n_i^2}$$

i la variança de la supervivència per

$$\widehat{Var}_{NA}(\widehat{S}(x)) = \widehat{S}_{NA}(x)^2 \cdot \sigma^2_{NA}(x).$$

3.3 Aplicació a dades reals: Estimadors no paramètrics

Recordem que al capítol 0 vam fer la presentació dels assaigs clínics reals dels quals usariem les dades per aplicar els conceptes que anem introduïnt.

3.3.1 Lèucemia

Les dades de l'assaig clínic van ser

Parella	Temps	Censura	Tractament	Estadi	Temps	Censura	Tractament	Estadi
1	1	1	Control	1	10	1	6-MP	1
2	22	1	Control	2	7	1	6-MP	2
3	3	1	Control	2	32	0	6-MP	2
4	12	1	Control	2	23	1	6-MP	2
5	8	1	Control	2	22	1	6-MP	2
6	17	1	Control	1	6	1	6-MP	1
7	2	1	Control	2	16	1	6-MP	2
8	11	1	Control	2	34	0	6-MP	2
9	8	1	Control	2	32	0	6-MP	2
10	12	1	Control	2	25	0	6-MP	2
11	2	1	Control	2	11	0	6-MP	2
12	5	1	Control	1	20	0	6-MP	1
13	4	1	Control	2	19	0	6-MP	2
14	15	1	Control	2	6	1	6-MP	2
15	8	1	Control	2	17	0	6-MP	2
16	23	1	Control	1	35	0	6-MP	1
17	5	1	Control	1	6	1	6-MP	1
18	11	1	Control	2	13	1	6-MP	2
19	4	1	Control	2	9	0	6-MP	2
20	1	1	Control	2	6	0	6-MP	2
21	8	1	Control	2	10	0	6-MP	2

Taula 3.1: Dades assaig leucèmia aguda.

En les dades tenim 42 pacients de leucèmia agrupats en 21 parelles. De cada parella un va ser tractat amb $6-MP$ i l'altre va formar part del grup control (placebo). De les dades 12 són censurades, mentres que la resta, 42, són dades completes. Per altra banda, 10 dels pacients es trobaven a l'estadi 1 (remissió parcial de la leucèmia) i 32 a l'estadi 2 (remissió completa). Es medeix el temps en setmanes fins la recaiguda o sortida de l'assaig.

Ara, farem un primer anàlisi de supervivència d'aquestes dades. El que ens proposem serà estimar la funció de supervivència i comparar gràficament la supervivència dels pacients tractats amb 6-PM i els pacients tractats amb placebo. Per això, en primer lloc, fem una taula amb els resultat obtinguts pels pacients tractats amb 6-MP utilitzant el programa estadístic R. A la primera columna trobarem el temps, que són els diferents x_i on hem observat mort. A les dues columnes següents trobem el nombre de persones en risc i el nombre d'events produïts a x_i. A la columna quarta hi trobem l'estimador de Kaplan-Meier avaluat a l'instant x_i, és a dir, $\widehat{S}_{KM}(x_i)$ i a la darrera columna l'estimació de la seva desviació estàndar, calculada a partir de la fórmula de Greenwod i de l'estimador de Kaplan-Meier, $\sqrt{\widehat{Var}_G(\widehat{S}_{KM}(x_i))}$. Després, de la mateixa manera, fem la taula amb els resultats obtinguts dels pacients tractats amb placebo.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	21	3	0.857	0.0764	0.720	1.000
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

Taula 3.2: Resultats pels pacients tractats amb 6-MP.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	21	2	0.9048	0.0641	0.78754	1.000
2	19	2	0.8095	0.0857	0.65785	0.996
3	17	1	0.7619	0.0929	0.59988	0.968
4	16	2	0.6667	0.1029	0.49268	0.902
5	14	2	0.5714	0.1080	0.39455	0.828
8	12	4	0.3810	0.1060	0.22085	0.657
11	8	2	0.2857	0.0986	0.14529	0.562
12	6	2	0.1905	0.0857	0.07887	0.460
15	4	1	0.1429	0.0764	0.05011	0.407
17	3	1	0.0952	0.0641	0.02549	0.356
22	2	1	0.0476	0.0465	0.00703	0.322
23	1	1	0.0000	NaN	NA	NA

Taula 3.3: Resultats pels pacients tractats amb placebo.

A la gràfica següent podrem observar les dues funcions de supervivència estimades per les dues poblacions, una estimant la mostra de pacients tractats amb 6-MP i l'altre estimant la mostra de pacients que se'ls hi va proporcionà placebo. L'estimació ha estat calculada mitjançant l'estimador de Kaplan-Meier.

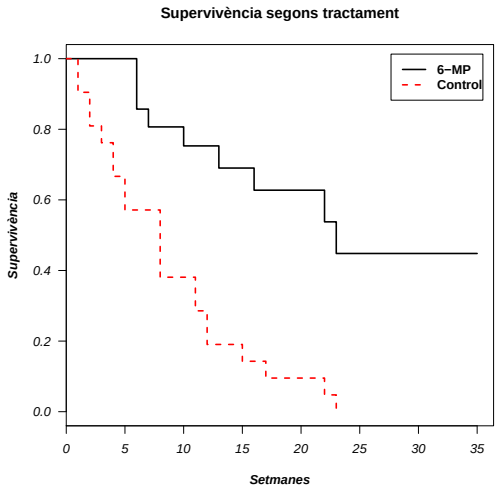

Figura 3.1: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Per altra banda, podem utilitzar l'estimador Nelson-Aalen per estimar les funcions de supervivència i risc. Mitjançant aquest estimador i calculant "a mà", hem realitzat els càlculs pel

grup tractat amb 6-MP que es presenten a la taula següent:

Time x	$\widehat{\Lambda}(x) = \sum_{x_i \leq x} \frac{d_i}{n_i}$	$\sigma^2_{NA} = \sum_{x_i \leq x} \frac{d_i}{n_i^2}$	std.err
$[0, 6)$	0	0	0
$[6, 7)$	$\frac{3}{21} = 0.1428$	$\frac{3}{21^2} = 0.0068$	0.0825
$[7, 10)$	$0.1428 + \frac{1}{17} = 0.2017$	$0.0068 + \frac{1}{17^2} = 0.0103$	0.1015
$[10, 13)$	$0.2017 + \frac{1}{15} = 0.2683$	$0.0103 + \frac{1}{15^2} = 0.0147$	0.1212
$[13, 16)$	$0.2683 + \frac{1}{12} = 0.3517$	$0.0147 + \frac{1}{12^2} = 0.0217$	0.1473
$[16, 22)$	$0.3517 + \frac{1}{11} = 0.4426$	$0.0217 + \frac{1}{11^2} = 0.0299$	0.1729
$[22, 23)$	$0.4426 + \frac{1}{7} = 0.5854$	$0.0299 + \frac{1}{7^2} = 0.0503$	0.2243
$[23, 35)$	$0.5854 + \frac{1}{6} = 0.7521$	$0.0503 + \frac{1}{6^2} = 0.0781$	0.2795

Usant el programa *R* i les mateixes dades que la taula anterior (nens tractats amb 6-MP), calculem la supervivència amb l'estimador Nelson-Aalen i dibuixem les funcions de supervivència estimades amb aquest mètode pels dos tractaments (6-MP i placebo). Notem que la supervivència de 23 setmanes és de 0.448 amb l'estimador Kaplan-Meier i en canvi 0.468 amb Nelson-Aalen. Aquesta vegada, la quarta columna és la funció de supervivència estimada per Nelson-Aalen i la cinquena la variança mitjançant l'estimador de Tsiatis.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	21	3	0.860	0.0767	0.723	1.000
7	17	1	0.811	0.0874	0.657	1.000
10	15	1	0.759	0.0971	0.591	0.975
13	12	1	0.698	0.1081	0.516	0.946
16	11	1	0.638	0.1159	0.447	0.911
22	7	1	0.553	0.1318	0.346	0.882
23	6	1	0.468	0.1405	0.260	0.843

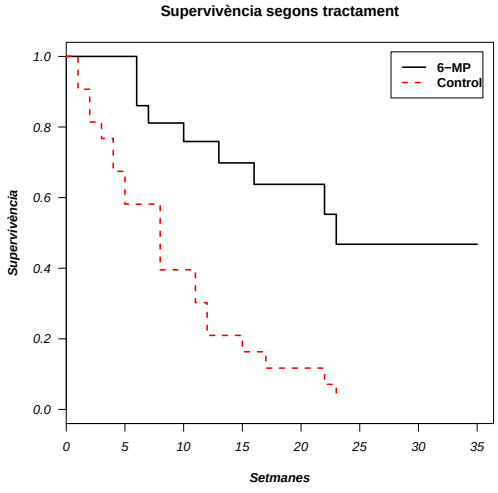

Figura 3.2: Corbes de supervivència mitjançant l'estimador Nelson-Aalen.

3.3.2 Dades de Rotterdam

Recordem que les dades de Rotterdam contenen la informació de 2982 dones a les quals es va detectar un primer tumor mamari. Es va fer un seguiment fins l'aparició de la recurrència locorregional, tumor contralateral o mort per càncer de mama. En aquest assaig clínic, l'interés consistia en estudiar els diferents factors associats a la supervivència de dones amb càncer de mama. Els factors que es van considerar per l'avaluació i estudi van ser: edat de la pacient, nombre de nodes limfàtics, receptor progesterona, menopausa (premenopausa o postmenopausa), grau del tumor, tamany del tumor, tractament amb tamoxifen (si va ser tractada o no amb aquest fàrmac hormonal) i quimioteràpia (si va ser tractada o no amb quimioterapia).

En primer lloc, per a estudiar l'anàlisi de supervivència d'aquest assaig clínic és observar amb una petita taula quants pacients van formar part de l'estudi i quants events van ser observats. També podem veure el comportament de la censura en aquestes dades i el nombre de dones premenopàusiques i postmenopàusiques.

Pacients	Events observats	No-Censurada 1	Censurada	Premenopausa	Postmenopausa
2982.0	1518.0	1518	1464	1312	1670

Per altra banda, fem un anàlisi previ d'alguns dels factors. A la primera taula es pot veure el nombre d'observacions classificades segons el tamany del tumor o si estàven hormonades o no. I, per últim, es mostra la taula descriptiva de l'edat de les pacients observades.

$T1(<= 20mm)$	$T2(> 20, <= 50mm)$	$T3(> 50mm)$	No-Hormonada	Hormonada
1387	1291	304	2643	339

Mínim	1r Quadrant	Mediana	Mitjana	3r Quadrant	Màxim
24.00	45.00	54.00	55.06	65.00	90.00

Fem un estudi de les corbes de supervivència segons el factor tamany del tumor, classificat segons la seva mesura en mil·límetres als grups $T1$, $T2$ o $T3$. Per als següents gràfics s'ha fet una estimació de la funció de supervivència mitjançant Kaplan-Meier.

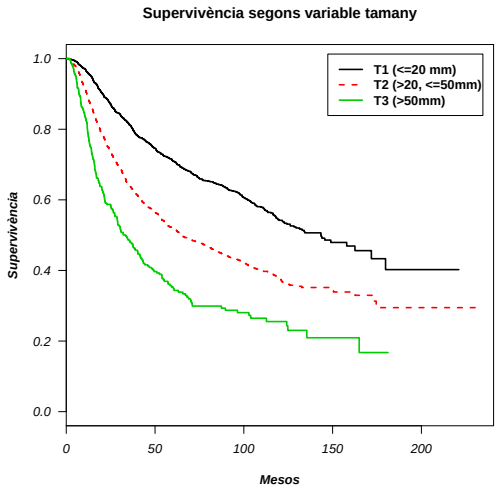

Figura 3.3: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

De la mateixa manera, fem un anàlisis de les corbes usant el factor menopausa.

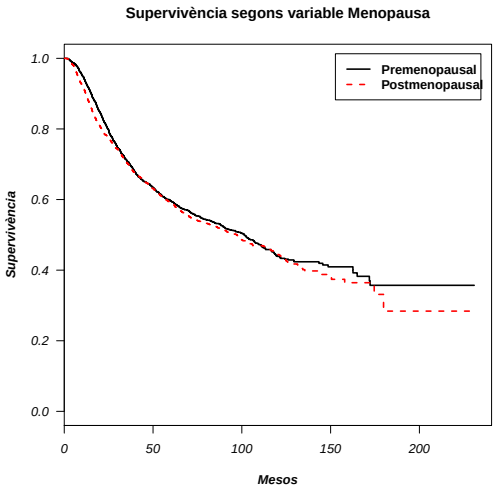

Figura 3.4: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Ara, fem l'estimació de la supervivència considerant els factors anteriors però mitjançant l'estimador Nelson-Aalen. Les gràfiques que s'obtenen són les següents:

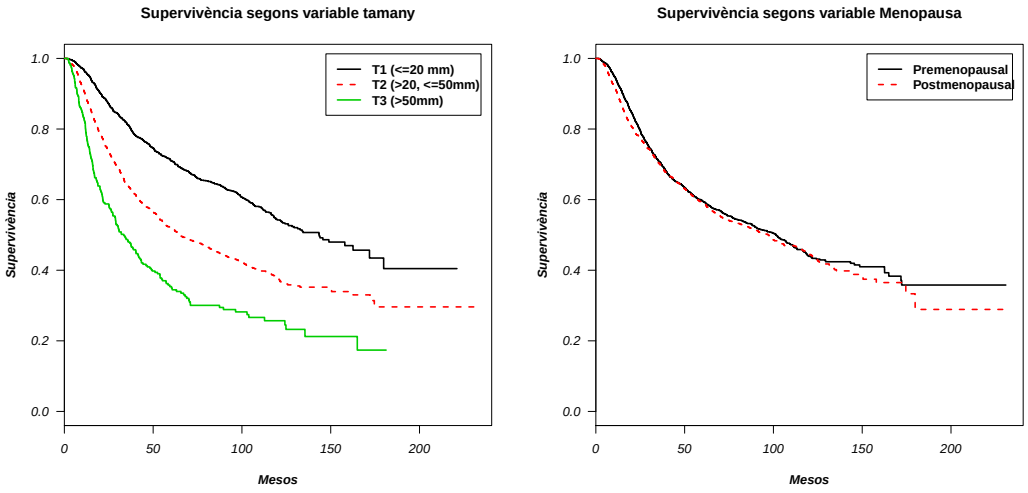

Figura 3.5: Corbes de supervivència mitjançant l'estimador Nelson-Aalen.

Es pot observar que les estimacions de Nelson-Aalen i Kaplan-Meier per la funció de supervivència són molt similars.

3.4 Propietats de l'estimador de Kaplan i Meier

Dediquem aquesta secció a fer un estudi de les propietats de l'estimador per la funció de supervivència de Kaplan i Meier. Les propietats que veurem amb més detalls són

- L'estimador de Kaplan i Meier és l'estimador que s'obté mitjançant l'algoritme de redistribució a la dreta.
- Sota certes condicions de regularitat, l'estimador de Kaplan i Meier és l'estimador de màxima versemblança no-paramètric generalitzat.

- L'estimador de Kaplan i Meier és autoconsistent.

3.4.1 Algoritme de redistribució a la dreta

Aquest és un algoritme proposat per Efron a l'any 1967 per estimar la funció de supervivència. Veurem que és equivalent a l'estimador Kaplan i Meier i ens proporciona un mètode recursiu pel seu càlcul. Es basa en la idea següent: si tinguéssim n dades al nostre conjunt i no hi hagués censura, l'estimador de $S(x)$ seria $1-F_n(x)$ ($F_n(x)$ la funció de distribució empírica). És a dir, donariem un pes igual a $1/n$ a cada temps observat. L'algoritme que va proposar Efron es basa en aquesta idea, redistribuint el pes corresponent de les observacions censurades entre els temps posteriors.

Considerem l'estadístic d'ordre $T_{(1)} < ... < T_{(n)}$. A continuació, donem els pasos per a implementar l'algoritme:

(i) Assignar a cada $T_{(i)}$ un pes $1/n$.

(ii) Començant pel valor més petit, deixem el pes anterior a totes les observacions no censurades fins arribar al primer valor censurat, el primer $i \in \{1, ..., n\}$, tal que $\delta_{(i)} = 0$, aleshores redistribuim el pes $1/n$ entre les següents observacions més grans. Per tant, si el primer valor censurat ha estat el $T_{(n-m)}$, a cadascun dels valors següents $T_{(n-m+1)}, ..., T_{(n)}$ li correspon un pes $\frac{1}{n}\left(1+\frac{1}{m}\right)$.

(iii) Deixem el pes anterior a totes les observacions no censurades fins arribar al segon valor censurat. Quan trobem el següent $j \in \{n-m+1, ..., n\}$, tal que $\delta_{(j)} = 0$, redistribuim el pes $\frac{1}{n}\left(1+\frac{1}{m}\right)$ entre les següents observacions més grans.

(iv) Iterem aquest procés fins que hem redistribuït el pes del darrer valor censurat o bé hem arribat al final.

Proposició 3.4.1. *Si l'última observació és no censurada, l'estimador de la funció de supervivència que s'obté d'aplicar l'algoritme de redistribució a la dreta és*

$$\widehat{S}_{RD}(x) = 1 - \sum_{k:Y_{(k)}\leq x} \frac{\delta_{(k)}}{n-k+1} \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$$

Aquest estimador és constant entre observacions i els salts en $Y_{(k)}$ són:

$$\widehat{S}_{RD}(Y_{(k)}^-) - \widehat{S}_{RD}(Y_{(k)}) = \frac{\delta_{(k)}}{n-k+1} \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}.$$

Demostració. Per simplificar la notació, considerem $u_k = Y_{(k)}$. L'algoritme dóna una estimació de la probabilitat de que l'event succeeixi a l'interval $(u_{k-1}, u_k]$. Per tant,

$$\begin{aligned}\widehat{Prob}(u_{k-1} \leq X \leq u_k) &= \widehat{F}_{RD}(u_k) - \widehat{F}_{RD}(u_k^-) = (1 - \widehat{S}_{RD}(u_k)) - (1 - \widehat{S}_{RD}(u_k^-)) \\ &= \widehat{S}_{RD}(u_k^-) - \widehat{S}_{RD}(u_k)\end{aligned}$$

Segons l'algoritme hem de redistribuir els pesos de les observacions censurades abans de $T_{(k)}$, de manera que $T_{(k)}$, si no està censurada, tindrà pes $1/n$ més els pesos derivats de les

observacions censurades anteriors. Per tant, tal com hem vist a la descripció de l'algoritme hem de multiplicar per $(1+\frac{1}{R(u_j)-1})$ sempre hi quan u_j $(j<k)$ sigui una observació censurada. Observem que quan multipliquem el nombre anterior al pes que ja tenia l'observació u_k, li estarem sumant aquest mateix valor dividit entre el nombre d'observacions que li queden a la dreta a la variable censurada u_j. Per tal cal multiplicar per $(1+\frac{1}{R(u_j)-1})^{1-\delta_{(j)}}$ $\forall j<k$

$$\begin{aligned}\widehat{F}_{RD}(u_k)-\widehat{F}_{RD}(u_{k-1}) &= \frac{\delta_{(k)}}{n}\prod_{j=1}^{k-1}\left(1+\frac{1}{R(u_j)-1}\right)^{1-\delta_{(j)}}\\ &= \frac{\delta_{(k)}}{n}\prod_{j=1}^{k-1}\left(1+\frac{1}{n-j+1-1}\right)^{1-\delta_{(j)}}.\end{aligned}$$

Recordem que per estar suposant que no hi ha empats es té $R(u_j)=n-j+1$. Ara, separant el productori i aplicant $\prod_{j=1}^{k-1}(1+\frac{1}{n-j})=\frac{n}{n-(k-1)}$ obtenim

$$\begin{aligned}\widehat{F}_{RD}(u_k)-\widehat{F}_{RD}(u_{k-1}) &= \frac{\delta_{(k)}}{n}\prod_{j=1}^{k-1}\left(1+\frac{1}{n-j}\right)^{-\delta_{(j)}}\prod_{j=1}^{k-1}\left(1+\frac{1}{n-j}\right)\\ &= \frac{\delta_{(k)}}{n-k+1}\prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}.\end{aligned}$$

Aquest algoritme únicament assigna pes positiu als temps no censurats, als temps censurats els hi assigna un pes igual a 0.
Per tant, l'estimador de la supervivència, tinguent en compte $\widehat{S}_{RD}(x)=1-\widehat{F}_{RD}(x)$ i que $\widehat{F}_{RD}(u_0)=0$, queda

$$\begin{aligned}\widehat{S}_{RD}(x) &= 1-\sum_{k:u_k\le x}(\widehat{F}_{RD}(u_k)-\widehat{F}_{RD}(u_{k-1}))\\ &= 1-\sum_{k:Y_{(k)}\le x}\frac{\delta_{(k)}}{n-k+1}\prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}\end{aligned}$$

□

Teorema 3.4.2. *Si l'última observació no és censurada, l'estimador de Kaplan i Meier és l'estimador que s'obté mitjançant l'algoritme de redistribució a la dreta,* $\widehat{S}_{RD}(x)=\widehat{S}_{KM}(x)$.

Demostració. Igual que abans utilitzem $u_k=Y_{(k)}$. En primer lloc, notem que $\widehat{S}_{KM}(x)=\widehat{S}_{RD}(x)=1, \forall x<T_{(1)}$ i també adonar-nos que els estimadors són constants als intervals $[u_k,u_{k+1})$. Així només ens cal veure que els salts a les dues funcions coincideixen, és a dir, $\widehat{S}_{RD}(u_k^-)-\widehat{S}_{RD}(u_k)=\widehat{S}_{KM}(u_k^-)-\widehat{S}_{KM}(u_k)$:

$$\begin{aligned}\widehat{S}_{KM}(u_k^-)-\widehat{S}_{KM}(u_k) &= \prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}-\prod_{j=1}^{k}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}\\ &= \prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}\left(1-\left(\frac{n-k}{n-k+1}\right)^{\delta_{(k)}}\right)\end{aligned}$$

Per acabar la demostració, separem per casos. Si $\delta_{(k)} = 0$, aleshores $\left(1 - \left(\frac{n-k}{n-k+1}\right)^{\delta_{(k)}}\right) = 0$ mentres que si $\delta_{(k)} = 1$, aleshores $(1 - \frac{n-k}{n-k+1}) = \frac{1}{n-k+1}$. Per tant, en els dos casos tenim:

$$\left(1 - \left(\frac{n-k}{n-k+1}\right)^{\delta_{(k)}}\right) = \frac{\delta_{(k)}}{n-k+1}$$

En conseqüència,

$$\widehat{S}_{KM}(u_k^-) - \widehat{S}_{KM}(u_k) = \prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} \frac{\delta_{(k)}}{n-k+1} \quad = \widehat{S}_{RD}(u_k^-) - \widehat{S}_{RD}(u_k).$$

□

3.4.2 Màxim de versemblança

Aquesta secció demostrem com l'estimador Kaplan-Meier és de màxima versemblança. En primer lloc, veiem aquest resultat en el cas discret.

Teorema 3.4.3. *Si X és una variable discreta, aleshores l'estimador de Kaplan-Meier és un estimador de màxima versemblança.*

Demostració. Sigui X una variable discreta que pren valors a $\{0 \leq a_1 < a_2 < ...\}$. La funció de risc llavors és $\lambda(a_j) = P(X = a_j | X > a_{j-1})$, $\forall j > 1$, i $\lambda(a_1) = P(X = a_1)$.

Recordem que la funció de supervivència en el cas discret queda definida amb la funció de risc de la manera següent

$$S(a_k) = \prod_{i=1}^{k}(1 - \lambda(a_i))$$

Aleshores, la funció probabilitat al valor $a_j, \forall j > 1$ es pot expressar

$$\begin{aligned} P(X = a_j) = S(a_{j-1}) - S(a_j) &= \prod_{i=1}^{j-1}(1 - \lambda(a_i)) - \prod_{i=1}^{j}(1 - \lambda(a_i)) \\ &= [1 - 1 + \lambda(a_j)]\prod_{i=1}^{j-1}(1 - \lambda(a_i)) = \lambda(a_j)\prod_{i=1}^{j-1}(1 - \lambda(a_i)). \end{aligned}$$

Suposem ara, de la mateixa manera que abans, que tenim una mostra $(T_1, \delta_1), ..., (T_n, \delta_n)$ i recordem que $T_i = min(X_i, C_i)$, $\delta_i = \mathbf{1}_{\{\mathbf{X_i} \geq \mathbf{C_i}\}}$ i la indendència entre les variables X_i i C_i. La funció de versemblança és:

$$\begin{aligned}
L((T_1,\delta_1),...,(T_n,\delta_n)) &= \prod_{i=1}^{n} P(X_i = T_i, \delta_i = 1)^{\delta_i} P(C_i = T_i, \delta_i = 0)^{1-\delta_i} \\
&= \prod_{i=1}^{n} P(X_i = T_i, X_i \leq C_i)^{\delta_i} P(C_i = T_i, X_i > C_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} (P(X_i = T_i)P(X_i \leq C_i))^{\delta_i} (P(C_i = T_i)P(X_i > C_i))^{1-\delta_i} \\
&\approx \prod_{i=1}^{n} P(X_i = T_i)^{\delta_i} P(X_i > C_i)^{1-\delta_i}
\end{aligned}$$

Suposem $T_i = a_k$ i per facilitar la notació $\lambda_j = \lambda(a_j)$

$$\begin{aligned}
P(X = a_k)^{\delta_i} P(X > a_k)^{1-\delta_i} &= \left(\lambda_k \prod_{j=1}^{k-1}(1-\lambda_j)\right)^{\delta_i} \left(\prod_{j=1}^{k}(1-\lambda_j)\right)^{1-\delta_i} \\
&= \lambda_k^{\delta_i} \left(\prod_{j=1}^{k-1}(1-\lambda_j)^{\delta_i}\right) \left(\prod_{j=1}^{k-1}(1-\lambda_j)^{1-\delta_i}\right) (1-\lambda_k)^{1-\delta_i}
\end{aligned}$$

Observem

$$\left(\prod_{i=1}^{k-1}(1-\lambda_i)^{\delta_k}\right) \left(\prod_{i=1}^{k-1}(1-\lambda_i)^{1-\delta_k}\right) = \prod_{i=1}^{k-1}(1-\lambda_i)^{\delta_k+1-\delta_k}$$

Llavors

$$\begin{aligned}
P(X = a_k)^{\delta_i} P(X > a_k)^{1-\delta_i} &= \lambda_k^{\delta_i}(1-\lambda_k)^{1-\delta_i} \prod_{j=1}^{k-1}(1-\lambda_j) \\
&= \lambda_k^{\delta_i}(1-\lambda_k)^{-\delta_i} \prod_{j=1}^{k}(1-\lambda_j)
\end{aligned}$$

Calculem ara $\prod_{i=1}^{n} P(X_i = T_i)^{\delta_i} P(X_i > C_i)^{1-\delta_i}$. Per cada temps a_k tindrem el producte de tants λ_k com individus morin en a_k, i tindrem tants $(1-\lambda_k)$ com individus en risc en a_k:

$$L \approx \prod_k \lambda_k^{d_k}(1-\lambda_k)^{n_k - d_k}$$

on $d_k = \sum_{i=1}^{n} \delta_i \mathbf{1}_{(\mathbf{T_i}=\mathbf{a_k})}$ i $n_k = \sum_{i=1}^{n} \mathbf{1}_{(\mathbf{T_i}\geq\mathbf{a_k})}$, observem que d_k és el nombre d'individus que fallen a a_k i n_k és el nombre d'individus en risc a a_k.

Per calcular l'estimador de màxim de versemblança haurem de veure per a quin λ_k s'assoleix el màxim de $L(X, \lambda_j)$. Prenem logaritmes

$$\ln L = \sum_k d_k \ln(\lambda_k) + \sum_k (n_k - d_k)\ln(1-\lambda_k) + \text{const}$$

Derivem i igualem a zero per trobar l'estimador

$$\begin{aligned}\frac{\partial \ln L}{\partial \lambda_k} &= d_k \frac{1}{\lambda_k} - (n_k - d_k)\frac{1}{1-\lambda_k} = 0\\ \frac{d_k}{\lambda_k} &= \frac{n_k - d_k}{1-\lambda_k}\\ \lambda_k(n_k - d_k) &= d_k(1-\lambda_k)\\ \lambda_k &= \frac{d_k}{n_k}\end{aligned}$$

Les segones derivades són negatives, aleshores el màxim de versemblança es troba a

$$\widehat{\lambda_k} = \frac{d_k}{n_k}$$

i per tant, l'estimador del màxim de versemblança per la supervivència és

$$\widehat{S}(x) = \begin{cases} 1 & \text{si } x < a_1 \\ \prod_{i=1}^{j}(1-\widehat{\lambda_i}) = \prod_{i=1}^{j}(1-\frac{d_i}{n_i}) & \text{si } a_j \leq x < a_{j+1} \end{cases}$$

que coincideix amb l'estimador Kaplan-Meier,

$$\widehat{S}_{KM}(x) = \begin{cases} 1 & \text{si } x < T_{(1)} \\ \prod_{i:T_{(i)}\leq x}(1-\frac{d_i}{n_i}) & \text{si } t \geq T_{(1)} \end{cases}$$

En conseqüència,

$$\widehat{S}(x) = \widehat{S}_{KM}(x)$$

□

Teorema 3.4.4. *Si X és una variable absolutament continua, aleshores l'estimador de Kaplan-Meier és un estimador de màxima versemblança.*

3.4.3 Autoconsistència

Introduirem el concepte d'autoconsistència que, de la mateixa manera que l'algoritme de redistribució a la dreta, fou definit per Efron.

Diem que un estimador $\widehat{S}_{AC}$ de la funció de supervivència S és *autoconsistent* si satisfà l'equació següent:

$$\widehat{S}_{AC}(x) = E_{\widehat{S}_{AC}}\left[\frac{1}{n}\sum_{i=1}^{n} 1\{X_i > x\} | (T_1,\delta_1),...,(T_n,\delta_n)\right]$$

on $E_{\widehat{S}_{AC}}$ indica que la esperança condicionada per les dades observades, $(T_1,\delta_1),...,(T_n,\delta_n)$, es pren respectela llei determinada per la funció de distribució $F = 1 - \widehat{S}_{AC}$.

Per a dades censurades per la dreta la solució de l'equació d'autoconsistència ve donada per:

$$\widehat{S}_{AC}(x) = \frac{1}{n}R(x) + \frac{1}{n}\sum_{i:T_i\leq x}(1-\delta_i)\frac{\widehat{S}_{AC}(x)}{\widehat{S}_{AC}(T_i)}$$

on $R(x) = card\{i : T_i > x\}$, és a dir, són les persones en risc a l'instant x.

Teorema 3.4.5. *Sota el supòsit de que no tenim empats, l'estimador de Kaplan-Meier $\widehat{S}$ és l'únic estimador de S autoconsistent per tot $x < T_{(n)}$. L'estimador de Kaplan-Meier no coincideix amb l'estimador autoconsistent per aquells $x \geq T_{(n)}$ quan $\delta_{(n)} = 0$, perquè en aquest cas $\widehat{S}_{AC} = 0$ i $\widehat{S}$ no està definit.*

3.5 Propietats asimptòtiques

Comencem aquesta secció la dediquem al comportament asimptòtic dels estimadors Kaplan-Meier i Nelson-Aalen. Les propietats fonamentals són les següents:

- L'estimador de Kaplan-Meier és consistent
- Sota certes condicions de regularitat l'estimador Nelson-Aalen de la funció de risc acumulada convergeix dèbilment a un procés Gaussià. En particular, per a x fixe, els estimadors convergeixen a una distribució normal.
- Sota certes condicions de regularitat l'estimador de Kaplan-Meier convergeix dèbilment a un procés Gaussià. En particular, per x fixe, els estimadors convergeixen a una distribució normal.

Es pot demostrar que l'autoconsistència estudiada a la secció anterior implica aquestes propietats. A la bibliografia es troba un article on es demostra aquesta implicació. Per altra banda, la normalitat dels estimadors Kaplan Meier i Nelson-Aalen permetrà construir els intervals de confiança per la funció de supervivència i funció de risc.
Tot seguit, vegem amb més deteniment aquestes propietats anomenades.

3.5.1 Consistència de l'estimador Kaplan i Meier

La primera de les propietats asimptòtiques que estudiarem és la consistència de l'estimador Kaplan-Meier. Per això, començarem aquesta subsecció amb un recordatori de la definició de consistència i, seguidament, ens endinsarem amb la introducció a les notacions i proposicions i teoremes que ens portaran a veure perquè l'estimador Kaplan-Meier és consistent.

Definició 3.5.1. Direm que una successió d'estimadors de $g(X)$, $\{T_n, n \geq 1\}$, és *fortament consistent* per a tot $\omega \notin N$, $P(N) = 0$ tenim que

$$lim_{n\to\infty} T_n(\omega) = g(X)$$

La successió d'estimadors que pendrem seràn estimadors de la funció de supervivència amb mides n cada cop més grans. El que voldrem aconseguir amb la consistència és que quan $n \to \infty$, la successió d'estimadors de la supervivència convergeixi quasi-segurament a la funció de supervivència real.

Siguin F, G les funcions de distribució de les variables X (temps de mort) i C (temps de censura) respectivament. Sigui H la funció de distribució de la variable aleatòria observada Y i considerem $S^* = 1 - H$ la funció de supervivència de Y. Aleshores, es compleix:

$$S^*(x) = 1 - H(x) = Prob(T > x) = P(min(X, C) > x)$$
$$= P(X > x, C > x) = P(X > x)P(C > x) = (1 - F(x))(1 - G(x)).$$

Definim la funcions de *sub-supervivència* S_u^* com

$$S_u^*(x) = Prob(T > x, \delta = 1) = P(X > x, X \leq C) = \int_x^\infty ((1 - G(u))dF(u)$$

Proposició 3.5.2. *Si les distribucions F i G no tenen salts en comú, aleshores la funció de supervivència $S(x)$ es pot expressar en funció de $S_u^*(x)$ i $S_c^*(x)$ de la manera següent*

$$S(x) = exp\left\{c\int_0^x \frac{dS_u^*(u)}{S^*(u)}\right\} exp\left\{d\sum_{u\leq x} log\frac{S^*(x^+)}{S^*(x^-)}\right\}$$

Aquesta expressió on S queda definida en funció de les sub-supervivències s'anomena representació de Peterson.
La notació que s'utilitza és la següent: $c\int$ denota l'integració sobre els intervals de continuitat de $S_u^*(x)$ mentres que $d\sum$ denota la suma sobre els salts discrets de $S_c^*(x)$.

Es pot demostrar també que l'estimador de Kaplan i Meier compleix la representació de Peterson. Aquest fet juntament amb la darrera proposició ens permet anunciar el següent teorema:

Teorema 3.5.3. *L'estimador de Kaplan-Meier és fortament consistent.*

3.5.2 Normalitat asimptòtica Nelson-Aalen

Seguint la notació de la subsecció anterior, definim:

$$\widehat{H}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[0,t]}(\mathbf{T_i}) \text{ i } \widehat{\mathbf{F}}_\mathbf{u} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}} \delta_\mathbf{i}\mathbf{1}_{[0,t]}(\mathbf{T_i}).$$

Teorema 3.5.4. *L'estimador de Nelson-Aalen normalitzat, és a dir, $\sqrt{n}(\widehat{\Lambda}_{NA}(x) - \Lambda(x))$, convergeix dèbilment, quan $n \to \infty$, a un procés Gaussià Z_Λ amb esperança 0 i covariància*

$$Cov(Z_\Lambda(x_1), Z_\Lambda(x_2)) = \int_0^{t_1\wedge t_2} \frac{d\widehat{F}_u(u)}{(1 - \widehat{H}(u))^2}.$$

3.5.3 Normalitat asimptòtica Kaplan-Meier

Com abans, denotarem per F, G les distribucion dels temps de mort, X, i de la censura, C.

Teorema 3.5.5. *Si les dues distribucions són continues a l'interval $[0, \tau]$ i $F(\tau) < 1$, aleshores*

- *L'estimador de Kaplan-Meier normalitzat, és a dir* $\sqrt{n}(\widehat{S}(x)-S(x))$*, convergeix dèbilment, quan* $n \to \infty$*, a un procés Gaussià* Z_S *amb mitjana* 0 *i la matriu de covariancies donada per*

$$Cov(Z_S(x_1), Z_S(x_2)) = S(x_1)S(x_2) \int_0^{t_1 \wedge t_2} \frac{d\widehat{F}_u(u)}{(1-\widehat{H}(u))^2}.$$

- *L'estimador Kaplan-Meier al punt* x*,* $\widehat{S}(x)$*, és aproximadament normal de mitjana* $S(x)$ *i variança*

$$\frac{S^2(x)}{n} \int_0^x \frac{d\widehat{F}_u(u)}{(1-\widehat{H}(u))^2}.$$

- *La variança asimptòtica s'estima mitjançant*

$$\widehat{Var}(\widehat{S}(x)) = \widehat{S}^2(x) \sum_i \frac{\delta_{(i)}}{(n-i)(n-i+1)}.$$

Capítol 4

Comparació de poblacions

En aquest capítol tractarem el problema de comparar la supervivència de dues o més poblacions. Quan la mostra no té censura els mètodes més usuals són la prova de Mann-Whitney-Wilcoxon per comparar dues poblacions, la prova de Savage (coneguda per logrank quan es té censura o també anomenada prova de scores exponencials) i la prova de Kruskal-Wallis per comparar més de dues poblacions. Ens centrarem en el cas de quan la mostra té censura, que és el més habitual a l'anàlisi de supervivència.

Els estimadors de Nelson-Aalen per la funció de risc acumulada i de Kaplan-Meier per la funció de supervivència són bàsics per descriure la supervivència d'una població. Si volem comparar dues poblacions podem fer un anàlisis de les corbes estimades per cada població amb aquestes dues funcions. Amb el propósit de quantificar de millor manera les comparacions i ens proposem desenvolupar proves de hipòtesis que avaluïn les diferències ponderades entre els riscos observats i els esperats. Definirem pesos que ens permetran plantejar proves que siguin més sensibles a les diferències de la hipòtesi nul·la quan els temps de supervivència són petits o quan els temps de supervivència són més grans. Com veurem, els estadístics es basaran en el nombre de events a cada temps de fallada i el nombre de pacients en risc a cadascún d'aquests temps.

Anem a explicar una mica el plantejament que seguirem per a fer la comparació de poblacions, hem de tenir en compte que estem suposant que la censura per la dreta és no-informativa. En primer lloc, definirem dos (o més, segons l'experiment que estem analitzant) grups d'individus, el *Grup 1* i el *Grup 2*, habitualment es sol referir a grup 1 pel grup control (placebo) i el grup 2 pel grup alternatiu (experimental). Per cada *Grup* j ($j = 1, 2$) tindrem n_j individus amb els seus temps de supervivència que suposem independents idènticament distribuïts amb funció de supervivència comú $S_j(\cdot)$. L'objectiu és construir una prova no-paramètrica per

$$H_0 : S_1(x) = S_2(x)\ \forall x \quad \text{contra} \quad H_1 : S_1(x) \neq S_2(x) \text{ per algun } x$$

o equivalentment

$$H_0 : \lambda_1(x) = \lambda_2(x)\ \forall x \quad \text{contra} \quad H_1 : \lambda_1(x) \neq \lambda_2(x) \text{ per algun } x$$

Com no poden existir proves uniformement de màxima potència per aquests contrastos d'hipòtesi, ens plantejarem també proves *direccionals* o *omnibus*. Les proves direccionals estan plantejades per detectar una diferència específica entre les funcions (una ser proporcional a altre, exponencials, ...), aquestes proves són poc potents respecte a altres opcions. Les proves omnibus són tests simples que ens permetran detectar alguna de les possibles diferències, són proves relativament potents per molts tipus d'alternatives.

Començarem aquest capítol amb el cas on únicament tenim una mostra de dades i volem construir un test per comprovar si la població de la mostra esmentada segueix una funció de

risc prefixada. En la secció segona considerarem K tractaments diferents i construirem un test per estudiar si existeixen diferències en la supervivència sota els K tractaments. La hipòtesi alternativa en aquest cas mantindrà que existeix un tractament amb una supervivència diferent. A la secció tercera es presentaran test per K mostres amb potència per detectar hipòtesis alternatives ordenades. A la secció quarta, presentarem com construir un test tinguent en compte covariats que poden afectar l'anàlisi i que no ens interessa estudiar. Una opció per abastar aquest problema és fer regressió, una altra són els tests estratificats que explicarem a la secció. Per últim, a la secció cinquena, presentarem una classe de tests amb alta potència per detectar funcions de risc creuades. Ens centrarem en el cas de dos mostres.

Tots els mètodes que estudiarem en aquest capítol, es podran aplicar quan les mostres presenten censura per la dreta i quan les mostres que presenten censura per la dreta i truncament per l'esquerra.

4.1 Tests per una mostra

Suposem una mostra amb dades censurades de tamany n de certa població. Volem veure si la funció risc de la població, λ, és una funció de risc prefixada, λ_0. La alternativa que mantindrem serà $\lambda \neq \lambda_0$. És a dir, tindrem el següent contrast d'hipòtesi

$$H_0 : \lambda(x) = \lambda_0(x), \forall x \leq \tau$$
$$H_1 : \lambda(x) \neq \lambda_0(x), \text{ per algun } x \leq \tau$$

on τ és el temps màxim observat a l'estudi.

Notarem per $x_1, ..., x_D$ els temps no censurats que hem observat a la mostra. A cada moment x_i, denotarem per d_i el nombre de temps de mort i per R_i el nombre d'individus en risc moment x_i.

Sigui $H(x)$ l'estimador de Nelson-Aalen per la funció de risc acumulada:

$$H(x) = \sum_{x_i \leq x} \frac{d_i}{R_i}, \;\; x_1 \leq x$$

Observem que $\frac{d_i}{R_i}$ és l'estimador de $\lambda(x_i)$. Sota la hipòtesi nul·la, el valor que esperem és $\lambda(x_i) = \lambda_0(x_i)$. L'estratègia que seguirem serà comparar la suma ponderada entre el que s'observa i el que s'espera en la hipòtesi nul·la. Això ho farem considerant diferents funcions pes positives, W. Depenent de la funció de pes que considerem potenciarem més una secció de la funció de risc o una altra. L'estadístic amb el que treballarem serà el següent:

$$Z(\tau) = \sum_{i=1}^{D} W(x_i)\frac{d_i}{R_i} - \int_0^\tau W(s)\lambda_0(s)ds$$

Si és certa la hipòtesi nul·la, la variança de l'estadístic serà:

$$Var(Z(\tau)) = \int_0^\tau W^2(s)\frac{\lambda_0(s)}{R(s)}ds.$$

A més a més, $Z(\tau)^2/Var(Z(\tau)) \to \chi^2$ quan la mida de la mostra tendeix a infinit. La funció pes més usada és: $W(x) = R(x)$, on recordem que $R(x)$ és el nombre de persones en risc a l'instant x.

4.2 Tests per dos o més mostres

En aquesta secció, extendrem els mètodes que hem començat anteriorment al problema de comparar funcions de risc de K ($K \geq 2$) poblacions. Suposem K mostres de dades de supervivència d'un assaig. Resumirem les dades observades amb la terna (T_i, δ_i, Z_i), $i = 1, ..., n$, on

$$T_i = min(X_i, C_i)$$

$$\delta_i = \begin{cases} 1 & \text{si } X_i \leq C_i \\ 0 & \text{si } X_i > C_i \end{cases}$$

$$Z_i = \begin{cases} 1 & \text{si la dada } i \text{ és de la mostra 1} \\ 2 & \text{si la dada } i \text{ és de la mostra 2} \\ ... & ... \\ K & \text{si la dada } i \text{ és de la mostra } K \end{cases}$$

La prova d'hipòtesi que voldrem plantejar serà la següent

$$H_0 : \lambda_1(x) = \lambda_2(x) = ... = \lambda_K(x) \ , \forall x \leq \tau$$
$$H_1 : \lambda_i(x) \neq \lambda_l(x) \text{ per algun } x \text{ i per algún parell } (i, l)$$

o equivalentment

$$H_0 : S_1(x) = S_2(x) = ... = S_K(x) \ , \forall x \leq \tau$$
$$H_1 : S_i(x) \neq S_l(x) \text{ per algun } x \text{ i per algún parell } (i, l)$$

on $S_1, ..., S_K$ són les funcions de supervivència i $\lambda_1, ..., \lambda_K$ les funcions de risc de les K poblacions. Notarem per $x_1, ..., x_D$ els temps no censurats ordenats que hem observat entre totes les mostres. A cada moment x_i, observem $d_{i1}, d_{i2}, ..., d_{iK}$ events dels $R_{i1}, R_{i2}, ..., R_{iK}$ individus en risc de cadascuna de les K mostres. Definim, $d_i = \sum_{j=1}^{K} d_{ij}$ el nombre total de morts a l'instant x_i.

El propòsit, igual que a la secció anterior, és fer inferència damunt d'un conjunt de dades per tal d'establir la funció de risc a tots els punts menors que τ. En aquesta nova situació amb K tractaments, considerem τ com el menor del conjunt de temps finals dels estudis de les K mostres.

El test està basat en la comparació de funcions de risc de cada població j ponderada per una funció pes, W. El que es tractarà és d'estimar les funcions de risc mitjançant Nelson-Aalen i després fer una comparació sota les hipòtesis nul·la i alternativa. Sota la hipòtesi nul·la, l'estimació de la funció risc és $\frac{d_i}{R_i}$. Per la mostra j-èsima, l'estimador de la funció de risc és $\frac{d_{ij}}{R_{ij}}$. Sigui $W(x)$ una funció pes, $W(x) \geq 0$ i que pren valor 0 quan $R_{ij} = 0$. Es defineix:

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(x_i) \left(\frac{d_{ij}}{R_{ij}} - \frac{d_i}{R_i} \right), j = 1, .., K$$

Observem que si la hipòtesi nul·la és certa, les funcions de risc són iguals i per tant les diferències $\left(\frac{d_{ij}}{R_{ij}} - \frac{d_i}{R_i} \right)$ tendiran a zero i l'estadístic Z_j pendrà valors petits. Per això, si els estadístics $Z_j(\tau)$ estan propers al zero, llavors no tindrem evidencies per creure que la hipòtesi nul·la és falsa. Altrament, si els valors són llunyans al zero tindrem una evidència de que les poblacions no segueixen una distribució de la funció risc plantejada a H_0. Per saber si estem

lluny o no del zero cal coneixer la distribució assimptòtica de les Z_j que evidentment depèn de la tria de les W_j.

Es sol definir la funció pes com $W_j(x_i) = R_{ij}W(x_i)$ on W és un pes comú compartit per les dos mostres. Aleshores, l'estadístic $Z_j(\tau)$ quedarà expressat com:

$$\begin{aligned} Z_j(\tau) &= \sum_{i=1}^{D} W_j(x_i)\left(\frac{d_{ij}}{R_{ij}} - \frac{d_i}{R_i}\right) \\ &= \sum_{i=1}^{D} W(x_i)\left(d_{ij} - R_{ij}\frac{d_i}{R_i}\right) \ , j = 1, ..., k \end{aligned} \tag{4.1}$$

$Z_j(\tau)$ estableix la suma ponderada de les diferencies entre el nombre observat de morts a t_i i el nombre esperat de morts sota H_0 en x_i a la mostra j. Sota la hipòtesi nul·la, l'esperança de $Z_j(\tau)$ és zero.

A continuació, presentem estimadors consistents de la variança de l'estadístic $Z_j(\tau)$ i de la covariança entre $Z_j(\tau)$ i $Z_g(\tau)$,

$$\sigma_{jj} = \sum_{i=1}^{D} W^2(x_i)\frac{R_{ij}}{R_i}\left(1 - \frac{R_{ij}}{R_i}\right)\frac{R_i - d_i}{R_i - 1}d_i, \text{ on } j = 1, ..., K.$$

$$\sigma_{jg} = -\sum_{i=1}^{D} W^2(x_i)\frac{R_{ij}}{R_i}\frac{R_{ig}}{R_i}\frac{R_i - d_i}{R_i - 1}d_i, \text{ on } g \neq j.$$

Els components del vector $(Z_1(\tau), Z_2(\tau), ..., Z_K(\tau))$ són linealment dependents degut a que $\sum_{j=1}^{D} Z_j(\tau) = 0$. Per tant, construirem l'estadístic amb $K-1$ estadístics $Z_j(\tau)$ i amb la matriu de dimensió $(K-1) \times (K-1)$ formada pels estadístics σ_{jg} dels corresponents $K-1$ components escollits. La prova, idò, es basa amb la forma quadràtica

$$(Z_1(\tau), Z_2(\tau), ..., Z_{K-1}(\tau))\Sigma^{-1}(Z_1(\tau), Z_2(\tau), ..., Z_{K-1}(\tau))^t \tag{4.2}$$

Quan la hipòtesi nul·la és certa aquest estadístic per n suficientment gran, segueix una llei χ^2 amb $K-1$ graus de llibertat.
Segons la funció pes W que escollim, tindrem diferents proves estadístiques. El pes que utilitzem dependrà de la alternativa. Per això, la idea és pendre un pes que emfatitzi aquesta diferencia.

A continuació, explicarem el cas per dues mostres més detalladament i després exposarem les proves més rellevants per als estudis de comparació de dues poblacions. Totes elles estan formulades en funció de $Z_j(\tau)$, com la diferència entre l'esperat i l'observat no està al quadrat, únicament seran potents quan el risc d'una població és més gran que la funció risc de l'altra. Quan les funcions de risc es creuen potser que no detectin bé les diferències.

4.2.1 Proves d'hipòtesi per dos mostres

Hem vist al començament de la secció com plantejar una prova d'hipòtesi per comparar K poblacions. En aquesta subsecció mirarem amb més deteniment el cas particular de la comparació entre dues poblacions.

De la mateixa manera que anteriorment, suposem dues mostres de dades de supervivència d'un assaig amb dos tractaments. Resumirem les dades observades amb la terna (T_i, δ_i, Z_i), $i = 1, ..., n$, on

$T_i = min(X_i, C_i)$

$\delta_i = \begin{cases} 1 & \text{si } X_i \leq C_i \\ 0 & \text{si } X_i > C_i \end{cases}$

$Z_i = \begin{cases} 1 & \text{si la dada } i \text{ és de la mostra 1} \\ 2 & \text{si la dada } i \text{ és de la mostra 2} \end{cases}$

Davant d'aquestes dades, ens plantegem formalment la següent prova d'hipòtesi:

$$H_0 : S_1(x) = S_2(x) \ \forall x \leq \tau \ \text{ contra } \ H_1 : S_1(x) \neq S_2(x) \text{ per algun } x \leq \tau$$

o equivalentment

$$H_0 : \lambda_1(\cdot) = \lambda_2(\cdot)$$

$$H_1 : \lambda_1(\cdot) \neq \lambda_2(\cdot)$$

Observem D temps de mort diferents $x_1, ..., x_D$ entre les dos mostres.

Definirem les variables aleatories d_{ij} com el nombre de morts a x_i sota el tractament T_j on $j = 1, 2$ i $d_i = d_{i1} + d_{i2}$, nombre total de morts a x_i entre les dues mostres. Les dades de l'experiment les podem resumir en la següent taula:

Mostra	Events	Supervivents	Totals en risc
1	d_{i1}	$R_{i1} - d_{i1}$	R_{i1}
2	d_{i2}	$R_{i2} - d_{i2}$	R_{i2}
Total	d_i	$R_i - d_i$	R_i

Prenem per estimador de la mostra j al temps x_i de la funció de risc: $\widehat{\lambda}_{ij} = \frac{d_{ij}}{R_{ij}}$. Si comparem les diferències entre les taxes de mort observades $\left(\frac{d_{ij}}{R_{ij}},\ j = 1, 2\right)$ i les esperades $\left(\frac{d_i}{R_i}\right)$, contruirem una classe de proves ponderant aquestes diferències mitjançant els pesos de $W(x_i)$ escollits de manera convenient.

Els pesos $W_j(x)$, $j = 1, 2$ són funcions de pes positives tals que $W_j(x) = 0$ si $R_{ij} = 0$, $j = 1, 2$. Definim els estadistics

$Z_j(\tau) = \sum_{i=1}^{D} W_j(x_i) \left(\frac{d_{ij}}{R_{ij}} - \frac{d_i}{R_i}\right)$, $j = 1, 2$

que tendiran a pendre valors petits quan H_0 és certa. Per tant, si $Z_1(\tau)$, $Z_2(\tau)$ són estadístics propers al zero, acceptarem la hipòtesi nul·la, mentres que si $Z_1(\tau)$ o $Z_2(\tau)$ és llunyà al zero, rebutjarem la hipòtesi nul·la.

Com al cas de K poblacions, es defineix la funció pes com $W_j(x_i) = R_{ij} W(x_i)$ on W és un pes comú compartit per les dos mostres. Aleshores,

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(x_i) \left(\frac{d_{ij}}{R_{ij}} - \frac{d_i}{R_i}\right) = \sum_{i=1}^{D} W(x_i) \left(d_{ij} - R_{ij} \frac{d_i}{R_i}\right) \ , \ j = 1, 2 \tag{4.3}$$

$Z_j(\tau)$ estableix la suma ponderada de les diferencies entre el nombre observat de morts a t_i i el nombre esperat de morts sota H_0 en x_i a la mostra j. Sota la hipòtesi nul·la, l'esperança de $Z_j(\tau)$ és zero.

A continuació, presentem estimadors consistents de la variança de l'estadístic $Z_j(\tau)$ i de la covariança entre $Z_1(\tau)$ i $Z_2(\tau)$

$$\widehat{Var}(Z_j(\tau)) = \sum_{i=1}^{D} W^2(x_i)\frac{R_{ij}}{R_i}\left(1-\frac{R_{ij}}{R_i}\right)\frac{R_i-d_i}{R_i-1}d_i$$

$$\widehat{Cov}(Z_1(\tau), Z_2(\tau)) = -\sum_{i=1}^{D} W^2(x_i)\frac{R_{i1}}{R_i}\frac{R_{i2}}{R_i}\frac{R_i-d_i}{R_i-1}d_i$$

Com $Z_1(\tau)+Z_2(\tau)=0$, les components del vector $(Z_1(\tau), Z_2(\tau))$ són linealment dependents i, per tant, podem simplificar la prova a un únic estadístic Z_W que es defineix com

$$Z_W = \frac{\sum_{i=1}^{D} W(x_i)\left(d_{i1}-R_{i1}\frac{d_i}{R_i}\right)}{\sqrt{\sum_{i=1}^{D} W^2(x_i)\frac{R_{i1}}{R_i}\left(1-\frac{R_{i1}}{R_i}\right)\frac{R_i-d_i}{R_i-1}d_i}} \tag{4.4}$$

Per n suficientment gran, Z_W es distribueix seguint una normal estàndard.

4.2.2 Prova del log-rank

En aquesta secció presentarem la prova del log-rank, veurem les hipòtesis que contrasta i els estimadors que caldrà usar. També veurem un petit exemple motivador que ens ajudarà a entendre el comportament de les variables aleatòries que després usarem per definir els estimadors per contrastar el test.
La prova del log-rank es pot utilitzar com un test direccional i s'obté considerant $W(x)=1$, $\forall x$. Posa el mateix pes en cada observació, per això és més sensible a exposicions amb un risc constant. Té òptima potència per detectar aquelles alternatives on les funcions de risc són proporcionals. Les funcions de supervivència que satisfan aquesta propietat s'anomenen *alternatives de Lehmann*, són les que tenen la forma: $S_j(x)=S(x)^{\theta_j}$. Així doncs, per la prova del log-rank la prova d'hipòtesi que voldrem contrastar és

$$H_0 : \lambda_1(x) = \lambda_2(x), \forall x \leq \tau$$
$$H_1 : \lambda_2(x) = e^{\beta}\lambda_1(x).$$

Exemple motivador 4.2.1. Suposem dos tractaments diferents T_1, T_2 i suposem una mostra de cadascuna d'aquestes poblacions. Observem D temps de mort diferents $x_1, ..., x_D$ entre les dos mostres. Definim

$$\begin{aligned} p_{i1} &= \text{probabilitat de morir en } x_i \text{ si es pertany a la mostra } 1, i \in \{1, ..., D\} \\ p_{i2} &= \text{probabilitat de morir en } x_i \text{ si es pertany a la mostra } 2, i \in \{1, ..., D\} \end{aligned} \tag{4.5}$$

Definim per aquest problema la prova d'hipòtesi següent:

$$H_0 : p_{11} = p_{12}, ..., p_{D1} = p_{D2}$$
$$H_1 : p_{i1} \neq p_{i2},\ i \in \{1, ..., D\} \tag{4.6}$$

Suposem que volem estudiar la diferència de mortalitat al moment x_i. Definirem les variables aleatories d_{ij} com el nombre de morts a x_i sota el tractament T_j on $j = 1, 2$ i $d_i = d_{i1} + d_{i2}$, és a dir, d_i correspon al nombre total de morts a x_i entre les dos mostres. Les dades de l'experiment les podem resumir en la següent taula:

Mostra	Events	Supervivents	Totals en risc
1	d_{i1}	$R_{i1} - d_{i1}$	R_{i1}
2	d_{i2}	$R_{i2} - d_{i2}$	R_{i2}
Total	d_i	$R_i - d_i$	R_i

Ens interessa estimar les probilitats p_{ij} per tal de construir un estimador per contrastar el test d'hipòtesi. Un possible estimador per p_{ij} és $\widehat{p}_{ij} = \frac{R_{ij}}{R_i} d_i$ i per la funció de risc: $\widehat{\lambda}_{ij} = \frac{d_{ij}}{R_{ij}}$. Si comparem les diferències entre les taxes de mort observades $\left(\frac{d_{ij}}{R_{ij}},\ j = 1, 2\right)$ i les esperades $\left(\frac{d_i}{R_i}\right)$, contruirem una classe de proves ponderant aquestes diferències mitjançant els pesos de $W(x_i)$.

Teorema 4.2.2. *La llei de la variable aleatòria d_{ij} condicionada a R_i,R_{ij} i d_i i sota la hipòtesi nul·la definida en 4.6, és una hipergeomètrica de paràmetres R_i,R_{ij} i d_i,*

$$Prob\{d_{ij} = d\} = \frac{\binom{d_j}{d}\binom{R_j - d_j}{R_{ij} - d}}{\binom{R_j}{R_{ij}}}$$

on $d \in \{max(0, d_j - R_{1j}), ..., min(d_j, R_{2j})\}$ i la seva esperança i variança venen donades per

$$E(d_{ij}|R_i, R_{ij}, d_i) = \frac{R_{ij} d_i}{R_i}$$
$$Var(d_{i1}|R_i, R_{ij}, d_i) = \frac{R_{i1} R_{i2} (R_i - d_i) d_i}{(R_i)^2 (R_i - 1)}.$$

L'estadístic definit com

$$\frac{d_{i1} - \frac{R_{ij} d_i}{R_i}}{\sqrt{\frac{R_{i1} R_{i2} (R_i - d_i) d_i}{(R_i)^2 (R_i - 1)}}}$$

s'aproxima a una Normal $(0, 1)$ i rebutja H_0 per valors extrems de d_{i1}. Equivalentment, l'estadístic

$$LR_i = \frac{(d_{i1} - \frac{R_{ij} d_i}{R_i})^2}{\frac{R_{i1} R_{i2} (R_i - d_i) d_i}{(R_i)^2 (R_i - 1)}}$$

s'aproxima un χ^2_1.

En general, ens interessa plantejar-nos la prova de hipòtesi composta i per això definim l'estadístic log-rank

$$LR = \frac{(\sum_{i=1}^{D}(d_{i1} - \frac{R_{ij}d_i}{R_i}))^2}{Var(\sum_{i=1}^{D}(d_{i1} - \frac{R_{ij}d_i}{R_i}))}$$

la variança del qual ve donada per

$$Var(\sum_{i=1}^{D}(d_{i1} - \frac{R_{ij}d_i}{R_i})) = \sum_{i=1}^{D}\frac{R_{i1}R_{i2}(R_i - d_i)d_i}{(R_i)^2(R_i - 1)}.$$

Després d'aquest exemple estem ja en disposició de definir formalment la prova log-rank; la prova d'hipòtesi, l'estadistic i el comportament quan el nombre d'individus és suficientment gran.

Definició 4.2.3. Plantegem la prova d'hipòtesi composta

$$H_0 : p_{11} = p_{12}, ..., p_{D1} = p_{D2}$$

$$H_1 : p_{i1} \neq p_{i2},\ i \in \{1, ..., D\}$$

i definim l'*estadistic logrank*

$$LR = \frac{\left(\sum_{i=1}^{D}(d_{i1} - \frac{R_{i1}d_i}{R_i})\right)^2}{Var\left(\sum_{i=1}^{D}(d_i - \frac{R_{i1}d_i}{R_i})\right)}$$

on

$$Var\left(\sum_{i=1}^{D}\left(d_i - \frac{R_{i1}d_i}{R_i}\right)\right) = \sum_{i=1}^{D}\frac{R_{i1}R_{i2}(R_i - d_i)d_i}{R_i^2(R_i - 1)}$$

A més a més, encara que les D taules no són independents es pot demostrar que $\sqrt{LR}$ es comporta asimptòticament com una $N(0, 1)$. El que ens permet enunciar el següent teorema:

Teorema 4.2.4. *La variable aleatòria LR s'aproxima a una χ_1^2 quan el nombre total d'individus tendeix a infinit.*

4.2.3 Prova Mantel-Haenszel

La prova de Mantel-Haenszel té similituts amb la prova de log-rank. Aquesta prova es construeix per provar una hipòtesi com la platejada 4.6, ara les probabilitats corresponen a diferents unitats. Posem per exemple que estem interessats en les diferències entre dos tractaments administrats en K hospitals diferents. Plantegem les hipòtesis:

$$\begin{aligned} H_0 &: p_{11} = p_{12}, ..., p_{K1} = p_{K2} \\ H_1 &: p_{i1} \neq p_{i2},\ i \in \{1, ..., K\} \end{aligned} \tag{4.7}$$

on p_{ij} = Prob(Mort | Tractament j, Hospital i).
Pot existir variabilitat entre els hospitals però a nosaltres únicament ens interessa la variabilitat entre tractaments. Per aquest motiu, no combinarem les K taules en una sola, treballarem amb K taules independents.

L'estadísticde Mantel-Haenszel es defineix com

$$MH = \frac{\sum_{i=1}^{K} \left(d_{i1} - \frac{R_{i1} d_i}{R_i} \right)}{\sqrt{\sum_{i=1}^{K} Var \left(d_{i1} - \frac{R_{i1} d_i}{R_i} \right)}}$$

Quan n és suficientment gran, l'estadístic MH és asimptòticament normal amb mitjana 0 i variança 1. El que ens permet construir la regió d'acceptació del test.

4.2.4 Prova Gehan

Aquesta prova fou introduida per Gehan l'any 1965 com una generalització de l'estadístic de Wilcoxon. L'objectiu és comparar cada temps de fallada de la mostra 1 amb cada temps de fallada de la mostra 2 mitjançant una funció Φ que anomenem *score*. La idea que es vol trasmetre amb la funció score és una comparació ràpida del temps de l'individu i de la mostra 1 amb l'individu l de la mostra 2. Assignarem el valor $+1$ al score si el temps de i és inferior al temps de l i, si pel contrari, i té un temps superior a l, li assignarem el valor -1 al score i 0 si no els podem comparar. Formalment, el score es definirà de la següent manera

$$\Phi((Y_{i1}, \delta_{i1}), (Y_{l2}, \delta_{l2})) = \begin{cases} +1 & \text{si } Y_{i1} \leq Y_{l2} \text{ i } \delta_{i1} = 1, \delta_{l2} = 0 \\ +1 & \text{si } Y_{i1} < Y_{l2} \text{ i } \delta_{i1} = 1, \delta_{l2} = 1 \\ -1 & \text{si } Y_{i1} \geq Y_{l2} \text{ i } \delta_{i1} = 0, \delta_{l2} = 1 \\ -1 & \text{si } Y_{i1} > Y_{l2} \text{ i } \delta_{i1} = 1, \delta_{l2} = 1 \\ 0 & \text{altrament.} \end{cases}$$

L'estadístic de Gehan es defineix com

$$GH = \sum_{i=1}^{n} \sum_{l=1}^{n} \Phi((Y_{i1}, \delta_{i1}), (Y_{l2}, \delta_{l2}))$$

Teorema 4.2.5. *L'estadístic de Gehan és un cas particular de la familia d'estadistics $Z_j(\tau)$ definits a 4.1.*

La prova s'obté considerant $W(x_i) = R_i$, per tant els pesos depenen fortament del moment on succeix l'event i de la distribució de la censura. Podem arribar a falses conclusions si els patrons de censures de les poblacions són molt diferents.

4.2.5 Prova Tarone i Ware

La prova de Tarone-Ware s'obté considerant $W(x_i) = \sqrt{R_i}$. De la mateixa manera que la prova Gehan, els pesos depenen del moment quan succeix l'event i de la distribució de la censura i en conseqüència podem arribar a falsos resultats si els patrons de censura de les poblacions són molt diferents.

4.2.6 Prova Peto-Peto

La prova de Peto-Peto s'obté considerant el pes $W(x) = \widehat{S}(x)$, on $\widehat{S}(x)$ és un estimador de la funció de supervivència per les dos mostres combinades. La corva d'aquest estimador $\widehat{S}(x)$ és molt semblant a la de l'estimador Kaplan-Meier. Es defineix com

$$\widehat{S}(x) = \prod_{x_i \leq x} \left(1 - \frac{d_i}{R_i + 1}\right).$$

Així doncs, l'estimador Z depèn de la funció de supervivència estimada i no està tan afectada pels patrons de censura de les diferents mostres.

4.2.7 Familia proves Fleming i Harrington

Per definir les proves Fleming i Harrington es considera la familia de pesos
$W_{p,q}(x_i) = \widehat{S}(x_{i-1})^p(1 - \widehat{S}(x_{i-1})^q)$, on $p, q \geq 0$.

Mitjançant l'elecció de p, q es permet ponderar la importancia donada a les diferències al principi o al final de la vida dels individus. Amb aquesta familia de proves es poden construir estadístics potents per a alternatives amb riscs diferens a una regió en concret. Aquesta família conté algunes proves d'hipòtesi vistes abans segons l'elecció de p, q

$$\begin{cases} p = q = 0, & \text{porta a la prova de log-rank} \\ p = 1, q = 0, & \text{porta a una extensió de la prova de Gehan} \\ p > 0, q = 0, & \text{porta a una prova per detectar diferències tempranes} \\ p = 0, q > 0, & \text{porta a una prova per detectar diferències tardies.} \end{cases}$$

4.2.8 Model de curació

S'anomena *model de curació* a les situacions on el model presuposa que una proporció desconeguda π de pacients es cura en un temps relativament curt i, en conseqüència, no està en risc a partir d'aquell moment. Per altre costat, la proporció $1 - \pi$ de pacients continuen en risc. Aquesta situació es dona quan, per exemple, l'efecte d'alguns tractaments curen a una proporció de pacients en un periode de temps relativament curt. Per aquesta situació no ens és útil les proves per compararque hem usat anteriorment. En aquest treball, no s'estudiarà aquest tipus de model, però si recalcarem que en aquest cas és necessàri un altre disseny de tests.

4.3 Tests de tendència

En aquesta secció, s'examinaran proves basades en la comparació ponderada entre la quantitat observada i la esperada de morts a cadascuna de les K mostres per comparar la hipòtesi

nul·la de que les K poblacions segueixen una mateixa funció de risc enfront a la hipòtesi alternatva de que cadascuna de les funcions de risc són diferents. En particular, presentarem nous tests per detectar hipòtesis alternatives ordenades. Voldrem contrastar les hipòtesis del tipus

$$H_0 : \lambda_1(x) = \lambda_2(x) = ... = \lambda_K(x) \ , \forall x \leq \tau$$
$$H_1 : \lambda_1(x) \leq \lambda_2(x) \leq ... \leq \lambda_K(x) \ , \forall x \leq \tau$$

o equivalentment

$$H_0 : S_1(x) = S_2(x) = ... = S_K(x) \ , \forall x \leq \tau$$
$$H_1 : S_1(x) \geq S_2(x) \geq ... \geq S_K(x) \ , \forall x \leq \tau$$

Els tests es basaran en el estadístics $Z_j(\tau)$ donats a 4.1. Els pesos exposats anteriorment serviran per ponderar els estadístics d'aquesta nova prova d'hipòtesis. Necessitem pendre un conjunt de constats creixent per construir el test. Siguin $a_1 < a_2 < ... < a_K$ un conjunt de constants. En molts casos es prenen, $a_j = j$ per $j = 1, ..., K$. El estadístic pel test és

$$Z = \frac{\sum_{j=1}^{K} a_j Z_j(\tau)}{\sqrt{\sum_{j=1}^{K} \sum_{g=1}^{K} a_j a_g \widehat{\sigma}_{jg}}}$$

Si la hipòtesi nul·la és certa, aleshores l'estadístic segueix asimptòticament una distribució normal. Si la hipòtesi alternativa és certa, aleshores els $Z_j(\tau)$ associats amb els valors a_j tendeixen a ser grans.

4.4 Tests estratificats

Quan volem aplicar els tests de comparació de K poblacions a vegades ens topem amb altres variables que afecten a l'estudi però no són les que estem estudiant. Aquestes variables sense interés a l'estudi però que poden influir en el resultat s'anomenen *factors soroll*. Anomenarem nivell a cada possible valor del factor. Per exemple, en l'estudi de la capacitat funcional de les persones majors, es seleccionen tots els individus majors de 65 anys amb l'objectiu de dispondre d'una mostra aleatòria de la ciutat de Barcelona. Potser el factor sexe influeixi a l'estudi, d'aquesta manera tindríem el factor sexe com a factor soroll i com a nivells d'aquest: home i dona.

Si el factor és desconegut i totalment incontrolable, ens cal aleatoritzar el màxim possible la mostra que escollim per l'estudi. Si és incontrolable però mesurable, haurem de fer l'estudi de la variable d'interés tinguent en compte els factors (covariants). Per últim, si el factor és controlable experimentalment, farem blocs i compararem la variable d'interés dins cada bloc fet segons els factors soroll.

Quan ens topem amb la situació d'ajustar les covariants, aleshores els mètodes que hem presentat no són vàlids. Una possibilitat és plantejar models de regressió i una altra alternativa és usar una versió estratificada dels mètodes anteriors.

Aquesta manera de plantejar el problema és possible sempre que el nombre de nivells de la covariant no sigui massa gran o quan, en el cas de ser continua, es discretitzai en un nombre

asequible de nivells. En aquesta secció es discutirà la manera de construir aquest tipus de tests i com s'utilitzen aquestos tests a l'anàlisi de supervivència.

Considerem el següent contrast d'hipòtesi estratificat on hi tenim una variable amb M nivells. Proposem la hipòtesi nul·la

$$H_0 : \lambda_{1s}(x) = \lambda_{2s}(x) = ... = \lambda_{Ks}(x) \text{ , on } s = 1, ..., M \text{ , } \forall x \leq \tau$$

contra

$$H_1 : \lambda_{is}(x) \neq \lambda_{js}(x) \text{ per algun parell } i, j$$

Sigui $Z_{js}(\tau)$ l'estadistic definit a 4.1 i sigui Σ_s la matriu formada pels estadístics σ_{jg}. Qualsevol pes dels exposats a les seccions anteriors es pot usar per Z_{js}. Aquests estadístics es poden utilitzar per provar la hipòtesi de la diferència en la funció de risc dins cada estrat, mitjançant la construcció del test de dos o més mostres (4.2). El test que crearem per discutir el contrast d'hipòtesi que acabem de definir és el que es presenta a continuació.

Siguin $Z_{j.}(\tau) = \sum_{s=1}^{M} Z_{js}$ i $\widehat{\sigma}_{jg.} = \sum_{s=1}^{M} \widehat{\sigma}_{jgs}$. El test que pendrem serà el basat en l'estadístic

$$(Z_{1.}(\tau), Z_{2.}(\tau), ..., Z_{(K-1).}(\tau))\Sigma_{.}^{-1}(Z_{1.}(\tau), Z_{2.}(\tau), ..., Z_{(K-1).}(\tau))^t \tag{4.8}$$

on $\Sigma_{.}$ és la matriu de dimensió $(K-1) \times (K-1)$ formada pels estadístics $\widehat{\sigma}_{jg.}$. Sota la hipòtesi nul·la, aquest estadístic segueix asimptòticament una distribució χ^2 amb $K-1$ graus de llibertat.
Quan $K = 2$, l'estadístic del test estratificat és

$$\frac{\sum_{s=1}^{M} Z_{1s}(\tau)}{\sqrt{\sum_{s=1}^{M} \widehat{\sigma}_{11s}}}$$

i es comporta asimptòticament com una distribució normal estàndar quan la hipòtesi nul·la és certa.

4.5 Test tipus Rényi

A la secció 4.2 es presenten un conjunt de test per comparar la supervivència de K poblacions. Totes aquestes proves es van basar en el sumatori ponderat de les diferències estimades entre les taxes de risc acumulats en les mostres. Quan apliquem aquests test a mostres on les funcions de risc es creuen, aquestes proves tindran poca potència i potser es aportin conclusions errònies. En aquesta secció, presentarem una classe de tests amb bona potència per detectar funcions de risc creuades. Enfocarem aquesta presentació al cas $K = 2$.

El test que presentarem s'anomena test Rényi i és anàleg al test Kolmogorov-Smirnov de comparació de dos mostres de dades no censurades. Per a construir el test, avaluarem cada instant observat a l'estimador calculat anteriorment 4.1. Tenint en compte que quan les funcions de risc es creuen, el valor absolut del conjunt d'avaluacions de Z pendrà un valor màxim en algun instant x abans de l'instant observat amb temps més gran. Quan aquest valor és gran, llavors, la hipòtesi nul·la d'interès $H_0 : \lambda_1(x) = \lambda_2(x), \forall t < \tau$ és rebutjada en favor de $H_A \lambda_1 \neq \lambda_2$, per algun t.

Suposem dos mostres independents de tamanys n_1 i n_2. Definim $n = n_1 + n_2$, el tamany de les mostres conjuntes. Siguin $x_1 < x_2 < ... < x_D$ els temps de mort observats entre les dues mostres. Sigui d_{ij} el nombre de temps de morts i R_{ij} el nombre de pacients en risc a la mostra j $(j = 1, 2)$ a l'instant de temps x_i $(i = 1, ..., D)$. Considerem una funció de pes positiva $W(x)$ i l'estadístic Z definit a 4.1. Aquesta vegada però, avaluarem l'estadístic a cada instant x_i on s'ha observat un event, usant únicament els temps de mort anteriors a aquest. És a dir,

$$Z_j(x_i) = \sum_{x_k \leq x_i} W(x_k) \left(d_{kj} - R_{kj} \frac{d_k}{R_k} \right) , \, i = 1, ..., D, \, j = 1, 2 \tag{4.9}$$

Sigui $\sigma(\tau)$ la desviació estàndar de $Z(\tau)$ que ve donada per

$$\sigma^2(\tau) = \sum_{x_k \leq x_i} W^2(x_k) \frac{R_{k1}}{R_k} \frac{R_{k2}}{R_k} \frac{R_k - d_k}{R_k - 1} d_k$$

on τ és l'event de temps més gran x_k on $R_{k1}, R_{k2} > 0$.

Aleshores, l'estadístic per la prova ve donat per

$$Q = \frac{sup\{|Z(x)|, x \leq \tau\}}{\sigma(\tau)}.$$

Quan la hipòtesi nul·la és certa, aleshores la distribució de Q pot ser aproximada per la distribució de $sup(|B(x)|, 0 \leq x \leq 1)$ on B és el procés estàndar del moviment brownià.

4.6 Aplicació a bases de dades reals: Comparació de poblacions

4.6.1 Leucèmia

En aquesta secció, com al tema anterior, apliquem els nous coneixements adquirits per fer un estudi de comparació de poblacions a les nostres bases de dades reals. Primer començarem amb l'assaig clínic de la remissió per a la leucèmia aguda. Aplicarem el *test log-rank* per comparar la mostra tractada amb la droga 6-MP amb la mostra de pacients als que se'ls hi va administrar placebo. Voldrem estudiar mitjançant les nostres dues mostres si hi ha diferències significatives entre els dos tractaments i extrapolar aquests resultats a la població. Com hem comentat anteriorment, aquest test té una molt bona potència per detectar aquelles hipòtesi alternatives on les funcions de risc són proporcionals. Plantejarem el següent contrast:

$$H_0 : \lambda_1(x) = \lambda_2(x)$$
$$H_1 : \lambda_1(x) \neq \lambda_2(x)$$

Vam veure que quan el nombre d'individus era suficientment gran la distribució de referència pel contrast és χ^2_1.

Abans de fer el test de comparació entre la població tractada amb placebo i la tractada amb 6-MP, recordem la gràfica de corbes de supervivència calculada anteriorment:

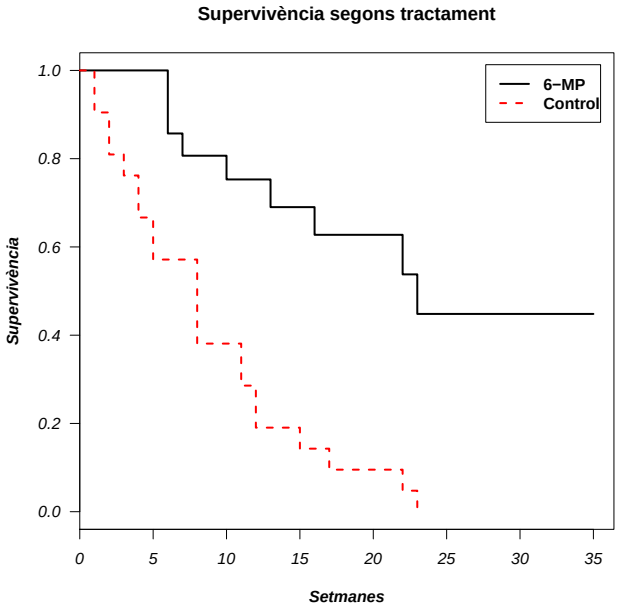

Figura 4.1: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Comparant les corbes de les dues poblacions, sembla que la supervivència de la població placebo és inferior a la de la població tractada amb 6-MP. Per veure si aquest fet és cert d'una manera rigurosa, calculem el test log-rank usant R.

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
6-MP	21	9	19.3	5.46	16.8
Control	21	21	10.7	9.77	16.8

I ens retorna també la següent informació:

$$\chi^2 = 16.8 \text{ on } 1 \text{ degrees of freedom, } p = 4.17e-05$$

Per tant, si considerem un nivell de significació $\alpha = 0.05$, rebutgem la hipòtesi nul·la i ens quedem amb l'alternativa. Aleshores, és certa la intuïció que teniem al comparar les corbes: la supervivència de la població tractada amb placebo és significativament inferior a la supervivència de la població tractada amb la droga 6-MP.

De la mateixa manera, per estudiar si existeixen diferències entre les funcions de risc segons el tractament, també podem fer el test Peto-Peto. En aquest cas, els resultats obtinguts són

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
6-MP	21	5.12	12.00	3.94	14.5
Control	21	14.55	7.68	6.16	14.5

$$\chi^2 = 14.5 \text{ on } 1 \text{ degrees of freedom, } p = 0.000143$$

Per tant, considerant el nivell de significació $\alpha = 0.05$, també rebutgem la hipòtesi nul·la i ens quedem amb l'alternativa: les funcions de risc són diferents per les poblacions control i

experimental. I per tant, les conclusions són les mateixes que les obtingudes amb el test log-rank: la supervivència de la població tractada amb placebo és significativament inferior a la supervivència de la població tractada amb la droga 6-MP.

Recordem que al conjunt de dades d'aquest assaig hi trobem parelles de pacients amb el mateix estat de remissió i tractats al mateix hospital. És natural plantejar-se l'estat de remissió com un possible efecte soroll. Per aquest motiu plantegem el test d'hipòtesi estratificant segons l'estat del pacient. Els resultats obtinguts fent aquest anàlisi són

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
6-MP	21	9	19.3	5.51	17.9
Control	21	21	10.7	9.96	17.9

$$\chi^2 = 17.9 \text{ on } 1 \text{ degrees of freedom, } p = 2.28e-05$$

Aleshores, considerant el nivell de significació $\alpha = 0.05$, rebutgem la hipòtesi nul·la i ens quedem amb l'alternativa: les funcions de risc són diferents per les poblacions control i experimental.

4.6.2 Dades de Rotterdam

Anàlogament a l'assaig clínic dels pacients amb leucèmia, volem plantejar un test per estudiar de manera rigurosa si existeixen diferències entre dues poblacions. Per això usarem els test *log-rank* i *Peto-Peto* que hem introduït a aquest capítol i que hem usat anteriorment. En primer lloc, estudiarem si el factor *tamany* del tumor influeix a la supervivència de la població i, després, estudiarem l'influència del factor *menopausa*.

Estudi del factor *tamany*

Abans de fer el test de comparació entre les poblacions classificades segons el tamany del tumor, recordem la gràfica de corbes de supervivència calculada anteriorment:

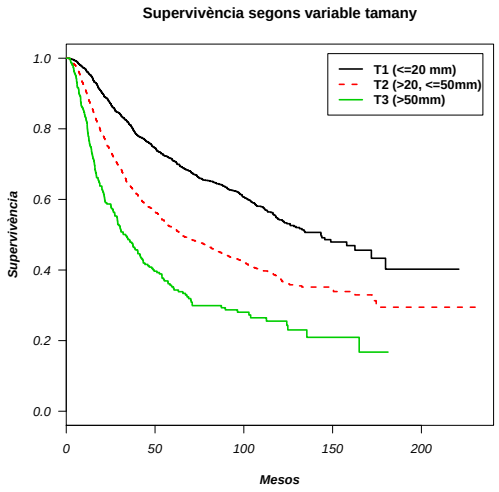

Figura 4.2: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Podem observar que les corbes de supervivència semblen diferents, on la corba de supervivència de la població amb tamany del tumor tipus T3 és la que pren valors més baixos i

la corba de supervivència de la població amb tamany tipus T1 la que pren valors més alts. Compararem d'una manera més rigurosa les tres poblacions amb un contrast d'hipòtesi.

La següent taula ens mostra els resultats obtinguts amb R quan estudiem si les diferents poblacions (és a dir, pacients amb tamany del tumor de tipus $T1$, $T2$ o $T3$) tenen o no la mateixa supervivència aplicant el test log-rank:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
size= $T1$ ($<= 20mm$)	1387	565	802	69.9	149.0
size= $T2$ ($> 20, <= 50mm$)	1291	745	611	29.3	49.1
size= $T3$ ($> 50mm$)	304	208	105	101.1	109.1

$$\chi^2 = 202 \text{ on 2 degrees of freedom, } p \cong 0$$

el p-valor obtingut és menor al nivell de significació α i, per tant, rebutgem la hipòtesi nul·la, les poblacions tenen funcions de supervivència diferents. Aleshores era certa la intuïció que havíem pres amb la gràfica, les corbes de supervivència són diferents. Si el pacient té un tumor de tamany T3 té menys probabilitat de sobreviure que un pacient detectat amb un tumor de tamany T1.

El següent test que apliquem és el de Peto-Peto, de la mateixa manera que abans, volem estudiar si el factor tamany del tumor és significatiu. Per aquest motiu comparem les tres poblacions per veure si les funcions de supervivència i risc són iguals. Els resultats obtinguts es mostren a la següent taula:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
size= $T1$ ($<= 20mm$)	1387	391	582.4	62.7	170.6
size= $T2$ ($> 20, <= 50mm$)	1291	559	454.1	24.2	52.8
size= $T3$ ($> 50mm$)	304	168	81.2	92.0	124.3

$$\chi^2 = 229 \text{ on 2 degrees of freedom, } p \cong 0$$

el p-valor obtingut és menor al nivell de significació α, en conseqüència rebutgem la hipòtesi nul·la, les poblacions tenen funcions de supervivència diferents. Les conclusions que obtenim són les mateixes que amb el test log-rank: les poblacions tenen funcions de supervivència diferents. I, segons la gràfica, la població T1 té una supervivència més alta i la T3 una supervivència més baixa.

Estudi del factor *menopausa*

Per estudiar el factor menopausa farem el mateix procés que el seguit a l'estudi del factor tamany. Volem veure si aquest factor és significatiu, per tant, estudiarem si les poblacions de dones premenopàusiques i dones postmenopàusiques tenen la mateixa funció de supervivència i risc. Com al cas anterior, aplicarem el test log-rank i després el test Peto-Peto.

Els resultats que hem obtingut amb el test log-rank són:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	1312	682	699	0.425	0.79
meno=Postmenopausal	1670	836	819	0.363	0.79

$$\chi^2 = 0.8 \text{ on } 1 \text{ degrees of freedom, } p = 0.374$$

Aleshores, com el p-valor és més gran que el nivell de significació, acceptem la hipòtesi nul·la. Les poblacions premenopausa i postmenopausa no tenen diferències significatives en la funció de supervivència: el factor no és significatiu a l'anàlisi de supervivència.

Recordem les corbes de supervivència calculades anteriorment:

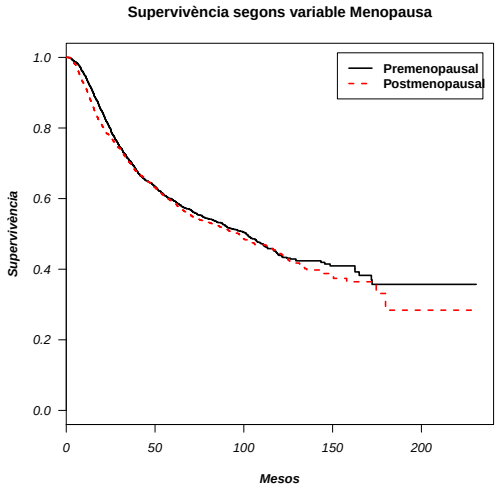

Figura 4.3: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Podem observar que les dues corbes s'assemblent molt, únicament als darrers temps és diferencien. Així doncs, és coherent que el test hagi donat per resultat que les funcions de supervivència no són significativament diferents.

Amb el test Peto-Peto els resultats són:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	1312	494	511	0.557	1.33
meno=Postmenopausal	1670	623	606	0.470	1.33

$$\chi^2 = 1.3 \text{ on } 1 \text{ degrees of freedom, } p = 0.248$$

Aleshores, com el p-valor és més gran que el nivell de significació, acceptem la hipòtesi nul·la. Les dues poblacions no presenten diferències significatives en la funció de supervivència.

Estudiem més a fons el factor menopausa. Per això, aplicarem el test estratificant el factor tamany del tumor per veure si aquest influeix i el factor menopausa és significatiu amb aquesta districció. La taula que es mostra a continuació són els resultats obtinguts al fer aquest test mitjaçant R

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	1312	682	685	0.00950	0.0174
meno=Postmenopausal	1670	836	833	0.00781	0.0174

$$\chi^2 = 0 \text{ on } 1 \text{ degrees of freedom, } p = 0.895$$

Podem veure amb més claretat que el p-valor que ens retorna el test és molt superior al nivell de significació habitual $\alpha = 0.05$, i, per tant el factor no és significatiu. No influeix a la supervivència de la població, és a dir, les dues poblacions (premenopausa, postmenopausa) tenen funcions de supervivència i risc similars.

Estudiem ara l'influència del factor menopausa quan estratifiquem pel factor quimioterapia. El que volem veure és si el factor menopausa és significatiu o no segons si tractem els pacients amb quimioterapia. Els resultats obtinguts fent el test són:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	1312	682	720	1.99	4.52
meno=Postmenopausal	1670	836	798	1.79	4.52

$$\chi^2 = 4.5 \text{ on } 1 \text{ degrees of freedom, } p = 0.0336$$

Aleshores, com el p-valor és inferior al nivell de significació, rebutgem la hipòtesi nul·la. Les poblacions de premenopausa i postmenopausa tenen diferent funció de supervivència estratificant pel factor quimioterapia.

Quan calculem la gràfica de la influència del factor menopausa damunt el subconjunt de dones tractades amb quimioteràpia obtenim:

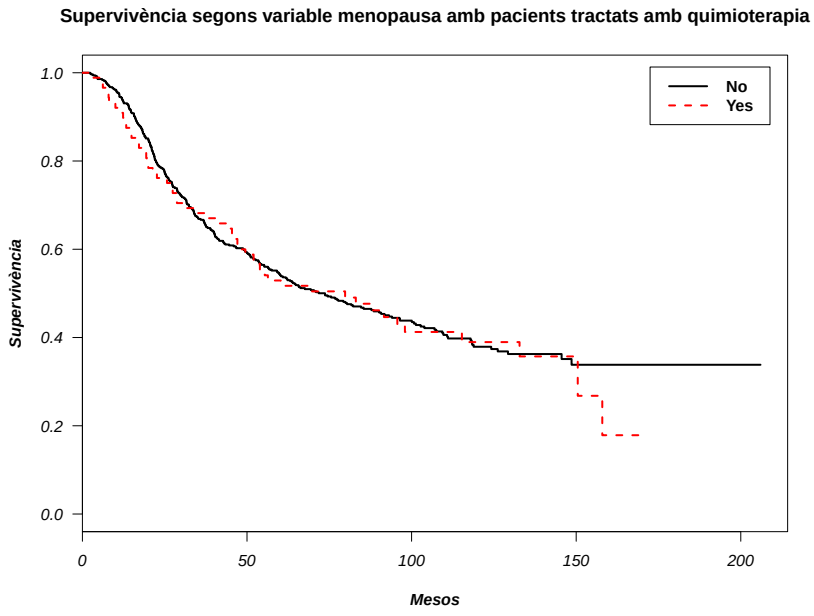

Figura 4.4: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Mentres que la gràfica de la influència del factor menopausa calculada damunt el subconjunt de dones no tractades amb quimioteràpia és:

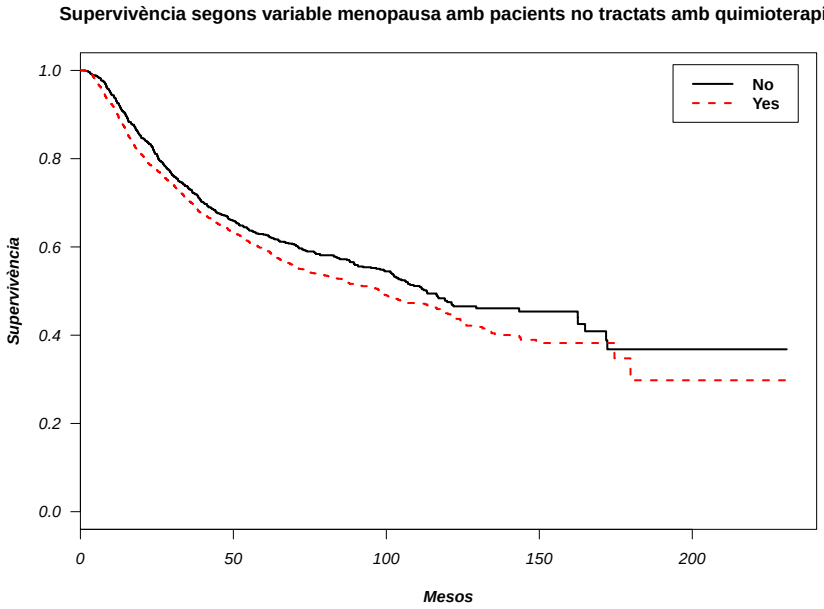

Figura 4.5: Corbes de supervivència mitjançant l'estimador Kaplan-Meier.

Segons els gràfics, la menopausa és un factor influent quan els pacients no han rebut tractament de quimioteràpia, mentres que quan els pacients han rebut tractament de quimioteràpia és un factor no influent.

Efectivament, aplicant el test log-rank al subconjunt de pacients no tractats per quimioteràpia obtenim els següents resultats:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	821	398	434	2.95	4.68
meno=Postmenopausal	1581	783	747	1.71	4.68

$$\chi^2 = 4.7 \text{ on } 1 \text{ degrees of freedom, } p = 0.0306$$

Per tant, el p-valor és inferior al nivell de significació α, en conseqüència, el factor menopausa és significatiu pels pacients no tractats amb quimioteràpia.
Per altre costat, quan apliquem el test log-rank al subconjunt de pacients tractats per quimioteràpia obtenim els següents resultats:

treat	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
meno=Premenopausal	491	284	286.1	0.0149	0.0985
meno=Postmenopausal	89	53	50.9	0.0835	0.0985

$$\chi^2 = 0.1 \text{ on } 1 \text{ degrees of freedom, } p = 0.754$$

El p-valor que obtenim és superior al nivell de significació α i, per tant, el factor menopausa no és significatiu pels pacients tractats amb quimioteràpia.

Conclusió

Hem introduït els mètodes estadístics de l'anàlisi de supervivència i el tractament per dades censurades. Primer als temes 1 i 2 presentant els conceptes bàsics estadístics que es requereixen i després amb l'explicació de les dades censurades i/o truncades, típiques en aquest tipus d'estudis. Després hem conegut el tractament no paramètric d'aquest tipus de dades tant en l'estimació de funció de supervivència i risc com en els test de comparació de diferents corbes de supervivència. Per últim, s'han exposat test d'hipòtesi per comparar diferents poblacions amb l'objectiu de dónar rigor a les comparacions entre corbes de supervivència.

El plantejament ha estat considerar mostres d'una població, a cadascuna d'elles se'ls hi assigna un tractament i es compara la supervivència de les poblacions segons el tractament. D'aquesta manera, estudiem si existeixen diferències entre tractaments o si pel contrari la seva eficiència és la mateixa. Els casos on hem aplicat aquesta metodologia han estat els assaigs clínics: remissió per a la leucèmia aguda i les dades de Rotterdam sobre el càncer de mama.

L'assaig de remissió va consistir en el seguiment de 42 nens amb leucèmia aguda de 11 hospitals diferents d'Estats Units amb una parcial o completa remissió de l'enfermetat. Per a l'estudi es va agrupar fent parelles de pacients amb el mateix estat de remissió i tractats al mateix hospital, aleatòriament a un pacient de la parella se'l tractà amb 6-mercaptopurina i l'altre amb placebo. Al nostre treball hem analitzat si havien diferències entre la població tractada amb prednisona i la població no tractada. Fent les corbes de supervivència ja vam sospitar que seguien diferents funcions de supervivència i quan vam aplicar els test de log-rank i Peto-Peto vam poder afirmar aquesta intuïció. Vam voler veure si l'estat de remissió influïa als test anteriors i vam veure que, efectivament, afectava i, a més, es feia més pronunciada la diferència entre tractaments.

A les dades de Rotterdam sobre el càncer de mama teníem la informació de 2982 dones a les que se'ls hi va detectar el primer tumor mamari, la variable d'interès fou el temps des de la primera intervenció quirúrgica fins a l'aparició del primer dels events: recurrència locorregional, tumor contralateral o mort per càncer de mama. L'interés de l'estudi va consistir en estudiar diferents factors associats a la supervivència. Quan hem treballat damunt d'aquestes dades hem escollit els factors menopausa, tamany del tumor i quimioteràpia per estudiar la seva influència a la supervivència dels pacients. En primer lloc, vam estimar les funcions de supervivència associades a les variables menopausa i tamany. Al segon cas pareixia clar que la funció de supervivència variava segons el tamany del tumor, al cas de la menopausa eren prou similars les gràfiques i "a ull"no es podia dir si les diferències eren significatives. Quan vam construir els test d'hipòtesi vam comprovar que, efectivament, la variable tamany influïa a la supervivència i, per l'altre costat, la variable menopausa no va resultar ser significativa per la supervivència. No contests amb això, vam contruir un test estratificat per veure si el factor menopausa era significatiu segons el factor tamany escollit i un altre test per comprovar si el factor menopausa era significatiu segons el factor quimioteràpia. Al primer test vam obtenir com a resultats que el factor menopausa no era significatiu mentres que al segon test va resultar que si era un factor que influïa a la supervivència. Observant les gràfiques i aplicant el test de log-rank al subconjunt de la mostra tractada i no tractada amb quimioteràpia vam concloure que el factor menopausa

era significatiu a la població dels pacients que no havíen set tractats amb quimioteràpia, mentres que no era significatiu pels pacients tractats amb quimioteràpia.

Apèndix

A l'anàlisi de supervivència fet a aquest treball per les dades reals dels assaigs clínics dels nens amb leucèmia aguda i per les dones amb càncer de mama hem realitzat els càlculs amb el programari estadístic *R*. Dediquem aquest apèndix a exposar les comandes usades gràcies a les quals hem pogut calcular els estimadors i fer els tests per comparar poblacions.

Leucèmia

Per aquest assaig les comandes i sortides del R van ser:

```
> table(remision$cens)

 0  1
12 30
> table(remision$treat)

6-MP Control
21      21
> table(remision$status)

 1  2
10 32

> srem<-with(remision,Surv(time,cens))
> srem
 [1]  1  22   3  12   8  17   2  11   8  12   2   5   4  15   8  23   5  11
   4   1   8  10   7  32+ 23  22   6
[28] 16  34+ 32+ 25+ 11+ 20+ 19+  6  17+ 35+  6  13   9+  6+ 10+

> survfit(srem~1)
Call: survfit(formula = srem ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     42      42      42      30      12       8      22
> svf<-survfit(srem~1)
> smsvf<-summary(svf)
> svf2t<-survfit(srem~treat,remision)
> summary(svf2t)
Call: survfit(formula = srem ~ treat, data = remision)

                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
```

```
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                treat=Control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000     NaN           NA           NA


> #Gràfica Supervivència segons estimador KM

> x11()
> par(font=2,font.axis=3,font.lab=4,las=1)
> plot(svf2t,col=1:2,conf.int=T,lwd=2)
> plot(svf2t,col=1:2,xlab='Setmanes',ylab="Supervivència",mark.time=F,lty=1:2,lwd=2)
> title('Supervivència segons tractament')
> legend('topright',levels(remision$treat),col=1:2,lty=1:2,lwd=2,inset=0.025)


> #Estimador Nelson-Aalen
> svfsh<-survfit(srem~treat,remision,type="fh2")
> (smsvfsh<-summary(survfit(srem~treat,remision,type="fh2")))
Call: survfit(formula = srem ~ treat, data = remision, type = "fh2")

                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.860  0.0767        0.723        1.000
    7     17       1    0.811  0.0874        0.657        1.000
   10     15       1    0.759  0.0971        0.591        0.975
   13     12       1    0.698  0.1081        0.516        0.946
   16     11       1    0.638  0.1159        0.447        0.911
   22      7       1    0.553  0.1318        0.346        0.882
   23      6       1    0.468  0.1405        0.260        0.843

                treat=Control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9070  0.0642       0.7895        1.000
    2     19       2   0.8140  0.0862       0.6615        1.000
```

```
    3     17       1   0.7675  0.0936       0.6043        0.975
    4     16       2   0.6745  0.1041       0.4985        0.913
    5     14       2   0.5815  0.1099       0.4015        0.842
    8     12       4   0.3955  0.1100       0.2293        0.682
   11      8       2   0.3026  0.1044       0.1539        0.595
   12      6       2   0.2097  0.0943       0.0868        0.506
   15      4       1   0.1633  0.0873       0.0573        0.466
   17      3       1   0.1170  0.0787       0.0313        0.437
   22      2       1   0.0710  0.0693       0.0105        0.481
   23      1       1   0.0261     Inf       0.0000        1.000


> #risc acumulat
> cym.haz<--log(svfsh$surv)
> cym.haz
 [1] 0.15025063 0.20907416 0.20907416 0.27574082 0.27574082 0.35907416 0.44998325
0.44998325 0.44998325
[10] 0.44998325 0.59284039 0.75950706 0.75950706 0.75950706 0.75950706 0.75950706
0.09761905 0.20580618
[19] 0.26462971 0.39379638 0.54214803 0.92750156 1.19535870 1.56202537 1.81202537
2.14535870 2.64535870
[28] 3.64535870


> #gràfica nelson-aalen
> x11()
> par(font=2,font.axis=3,font.lab=4,las=1)
> plot(svfsh,col=1:2,conf.int=T,lwd=2)
> plot(svfsh,col=1:2,xlab='Setmanes',ylab="Supervivència",mark.time=F,lty=1:2,lwd=2)
> title('Supervivència segons tractament')
> legend('topright',levels(remision$treat),col=1:2,lty=1:2,lwd=2,inset=0.025)


> #logrank
> survdiff(Surv(time,cens)~treat, data=remision)
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = remision)

               N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.46      16.8
treat=Control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
> #peto peto
> survdiff(srem~treat,remision,rho=1)
Call:
survdiff(formula = srem ~ treat, data = remision, rho = 1)

               N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21     5.12    12.00      3.94      14.5
treat=Control 21    14.55     7.68      6.16      14.5
```

```
 Chisq= 14.5  on 1 degrees of freedom, p= 0.000143
> #estratificado
> survdiff(srem~treat+strata(status),remision)
Call:
survdiff(formula = srem ~ treat + strata(status), data = remision)

              N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.51      17.9
treat=Control 21       21     10.7      9.96      17.9

 Chisq= 17.9  on 1 degrees of freedom, p= 2.28e-05
```

Dades Rotterdarm

Per aquest assaig les comandes que es van usar són similars al cas anterior, no es varen imprimir en pantalla les dades degut a la gran quantitat que es tenia (2982). Així doncs, les comandes i sortides del R van ser:

```
> #Descriptiva inicial
> table(rot.red$cens)

   0    1
1464 1518
> table(rot.red$grade)

   2    3
 794 2188
> table(rot.red$size)

    T1 (<=20 mm) T2 (>20, <=50mm)       T3 (>50mm)
            1387             1291              304
> table(rot.red$meno)

 Premenopausal Postmenopausal
          1312           1670
> table(rot.red$hormon)

  No  Yes
2643  339


> srem<-with(rot.red,Surv(reltime,cens))
> survfit(srem~1)
Call: survfit(formula = srem ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
 2982.0  2982.0  2982.0  1518.0    98.0    89.3   104.9
> svf<-survfit(srem~1)
> smsvf<-summary(svf)
```

```
> svf2t<-survfit(srem~grade,rot.red)


> #km segun tamaño
> km<-survfit(srem~size,rot.red)
> x11()
> par(font=2,font.axis=3,font.lab=4,las=1)
> plot(km,col=1:3,conf.int=T,lwd=2)
> plot(km,col=1:3,xlab='Mesos',ylab="Supervivència",mark.time=F,lty=1:2,lwd=2)
> title('Supervivència segons variable tamany')
> legend('topright',levels(rot.red$size),col=1:3,lty=1:2,lwd=2,inset=0.025)

> #km segun menopausia
> km<-survfit(srem~meno,rot.red)
> x11()
> par(font=2,font.axis=3,font.lab=4,las=1)
> plot(km,col=1:2,conf.int=T,lwd=2)
> plot(km,col=1:2,xlab='Mesos',ylab="Supervivència",mark.time=F,lty=1:2,lwd=2)
> title('Supervivència segons variable Menopausa')
> legend('topright',levels(rot.red$meno),col=1:2,lty=1:2,lwd=2,inset=0.025)
> #Estimador Nelson-Aalen Menopausa
> na.meno<-survfit(srem~meno,rot.red,type="fh2")
> #risc acumulat
> cym.haz<--log(na.meno$surv)

> #gràfica nelson-aalen
> x11()
> par(font=2,font.axis=3,font.lab=4,las=1)
> plot(na.meno,col=1:2,conf.int=T,lwd=2)
> plot(na.meno,col=1:2,xlab='Mesos',ylab="Supervivència",mark.time=F,lty=1:2,lwd=2)
> title('Supervivència segons variable Menopausa')
> legend('topright',levels(rot.red$meno),col=1:2,lty=1:2,lwd=2,inset=0.025)
> #Estimador Nelson-Aalen tamany
> na.size<-survfit(srem~size,rot.red,type="fh2")
> #risc acumulat
> cym.haz<--log(na.size$surv)


> #logrank
> survdiff(srem~meno,rot.red)
Call:
survdiff(formula = srem ~ meno, data = rot.red)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
meno=Premenopausal  1312      682      699     0.425      0.79
meno=Postmenopausal 1670      836      819     0.363      0.79

 Chisq= 0.8  on 1 degrees of freedom, p= 0.374
> survdiff(srem~meno,rot.red,rho=1)
Call:
survdiff(formula = srem ~ meno, data = rot.red, rho = 1)
```

```
                      N Observed Expected (O-E)^2/E (O-E)^2/V
meno=Premenopausal  1312      494      511     0.557      1.33
meno=Postmenopausal 1670      623      606     0.470      1.33

 Chisq= 1.3  on 1 degrees of freedom, p= 0.248
> survdiff(srem~size,rot.red)
Call:
survdiff(formula = srem ~ size, data = rot.red)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
size=T1 (<=20 mm)     1387      565      802      69.9     149.0
size=T2 (>20, <=50mm) 1291      745      611      29.3      49.1
size=T3 (>50mm)        304      208      105     101.1     109.1

 Chisq= 202  on 2 degrees of freedom, p= 0
> survdiff(srem~size,rot.red,rho=1)
Call:
survdiff(formula = srem ~ size, data = rot.red, rho = 1)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
size=T1 (<=20 mm)     1387      391    582.4      62.7     170.6
size=T2 (>20, <=50mm) 1291      559    454.1      24.2      52.8
size=T3 (>50mm)        304      168     81.2      92.0     124.3

 Chisq= 229  on 2 degrees of freedom, p= 0
> survdiff(srem~size+strata(chemo),rot.red)
Call:
survdiff(formula = srem ~ size + strata(chemo), data = rot.red)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
size=T1 (<=20 mm)     1387      565      798      67.9     144.6
size=T2 (>20, <=50mm) 1291      745      614      27.8      46.9
size=T3 (>50mm)        304      208      106      98.7     106.8

 Chisq= 197  on 2 degrees of freedom, p= 0
> survdiff(srem~meno+strata(chemo),rot.red)
Call:
survdiff(formula = srem ~ meno + strata(chemo), data = rot.red)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
meno=Premenopausal  1312      682      720      1.99      4.52
meno=Postmenopausal 1670      836      798      1.79      4.52

 Chisq= 4.5  on 1 degrees of freedom, p= 0.0336
> survdiff(srem~meno+strata(size),rot.red)
Call:
survdiff(formula = srem ~ meno + strata(size), data = rot.red)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
meno=Premenopausal  1312      682      685   0.00950    0.0174
```

```
meno=Postmenopausal 1670      836      833   0.00781    0.0174

 Chisq= 0  on 1 degrees of freedom, p= 0.895
```

Bibliografia

[Kle] Klein, John P.; Moeschberg, Melvin L.: *Survival Analysis. Techniques for Censored and Truncated Data*. United States: Springer, 1997.

[Gom] Gómez, Guadalupe; Olga Julià; Klaus Langohr: *Análisis de supervivencia*. Bellaterra: Publicacions de la Universitat Politècnica de Barcelona, 2011.

[Tsa] Tsai, Wei-Yann; Crowley, John: *A large saple study of generalized maximum likelihood estimators from incomplete data via self-consistency*. The Annals of Statistics, Vol. 13, No. 4 (Dec., 1985), pp. 1317-1334.

[Gai] M. Gail; K. Krickeberg; J. Samet; A. Tsiatis; W. Wong: *Statistics for Biology and Health* United States: Springer, 2005.